The Canadian Journal of Neurological Sciences (2025), 1–10
doi:10.1017/cjn.2025.10111

Canadian Journal of Neurological Sciences
Journal Canadien des Sciences Neurologiques

Review Article

The Effect of Training and Certification for the NIHSS and the mRS on Rater Performance: A Systematic Review

Davis MacLean[1], Diana Kim[2], Richard H. Swartz[3], Shelagh B. Coutts[2,4,5,6] and Aravind Ganesh[2,5,6]

[1]Department of Medicine, University of Calgary Cumming School of Medicine, Calgary, Canada, [2]Calgary Stroke Program, Department of Clinical Neurosciences, University of Calgary Cumming School of Medicine, Calgary, Canada, [3]Department of Medicine (Neurology), Sunnybrook Health Sciences Centre, University of Toronto, Toronto, ON, Canada, [4]Department of Radiology, Cumming School of Medicine, University of Calgary, Foothills Medical Centre, Calgary, Canada, [5]Department of Community Health Sciences and the O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Foothills Medical Centre, Calgary, Canada and [6]Hotchkiss Brain Institute, University of Calgary, Foothills Medical Centre, Calgary, Canada

ABSTRACT: ***Background:*** Up-to-date certification of the National Institutes of Health Stroke Scale (NIHSS) and modified Rankin Scale (mRS) is often required for clinical trials, representing a significant burden on clinical investigators globally. ***Aims:*** This systematic review sought to determine if NIHSS or mRS training, re-training, certification or recertification led to improvements in the reliability or accuracy of ratings as well as other relevant user metrics (e.g., user confidence). ***Results:*** Among 4227 studies, 100 passed screening and were assessed for eligibility with full-text review; 23 met inclusion criteria. Among these 23 studies, 22 examined NIHSS training and/or certification, and only a single study included examined the effect of training on mRS performance. Ten of 23 included studies were conference abstracts. The study designs, interventions and outcome measurement of the included studies were heterogeneous. In the case of the NIHSS, two studies found increased accuracy after NIHSS training, and a third study showed statistically significant though clinically trivial decreases in error rate with training. The remaining 19 studies showed no benefit of NIHSS training as it relates to reliability or accuracy outcomes. The single included mRS study did not show the benefit of training. ***Conclusion:*** Although data are sparse with heterogeneous training protocols and outcomes, there is no compelling evidence to suggest benefit of healthcare professionals completing NIHSS or mRS training, certification or recertification. At the very least, recertification/re-training requirements should be reconsidered pending the provision of robust evidence.

RÉSUMÉ : Effets de la formation et de la certification sur la performance des évaluateurs pour l'échelle d'évaluation de l'AVC des *National Institutes of Health* et pour l'échelle modifiée de Rankin: une revue systématique. ***Contexte :*** Une certification à jour pour l'échelle d'évaluation de l'AVC des *National Institutes of Health* (NIH) et de l'échelle modifiée de Rankin (EMR) est souvent requise pour des essais cliniques, ce qui représente un fardeau important pour les chercheurs du monde entier. ***Objectifs :*** Cette revue systématique a cherché à déterminer si la formation, le recyclage, la certification ou la re-certification pour l'échelle d'évaluation de l'AVC des NIH ou pour l'EMR ont permis d'améliorer la fiabilité ou la précision des évaluations ainsi que d'autres paramètres pertinents pour l'utilisateur, sa confiance par exemple. ***Résultats :*** Sur 4227 études, 100 ont été retenues et ont fait l'objet d'un examen complet. De ce nombre, 23 répondaient à nos critères d'inclusion. Parmi ces 23 études, 22 portaient sur la formation et/ou la certification pour l'échelle d'évaluation de l'AVC des NIH et une seule étude incluse portait sur l'effet de la formation pour l'EMR en lien avec la performance. À noter que 10 études sur 23 étaient des résumés de conférence. Les modèles d'étude, les interventions et la mesure des résultats des études incluses se sont avérés hétérogènes. Dans le cas des NIH, deux études ont constaté une augmentation de la précision après une formation et une troisième a montré une diminution statistiquement significative, bien que cliniquement insignifiante, du taux d'erreur des évaluateurs. Cela dit, les 19 autres études n'ont montré aucun avantage de la formation des NIH en termes de fiabilité ou de précision. Enfin, l'unique étude à propos de l'EMR n'a pas montré de bénéfice en lien avec une formation. ***Conclusion :*** Bien que les données soient rares et que les protocoles de formation et les résultats soient hétérogènes, il n'existe pas de preuves convaincantes des avantages pour les professionnels de la santé de suivre une formation, une certification ou une re-certification portant sur l'échelle d'évaluation de l'AVC des NIH ou sur l'ERM. À ce sujet, les exigences en matière de re-certification/formation devraient à tout le moins être reconsidérées dans l'attente de preuves solides.

Keywords: Certification; modified Rankin Scale; mRS; National Institutes of Health Stroke Scale; NIHSS; training

(Received 1 January 2025; final revisions submitted 2 April 2025; date of acceptance 1 May 2025)

Corresponding author: Davis MacLean; Email: davis.maclean@ucalgary.ca

Cite this article: MacLean D, Kim D, Swartz RH, Coutts SB, and Ganesh A. The Effect of Training and Certification for the NIHSS and the mRS on Rater Performance: A Systematic Review. *The Canadian Journal of Neurological Sciences*, https://doi.org/10.1017/cjn.2025.10111

Introduction

The National Institutes of Health Stroke Scale (NIHSS) and the modified Rankin Scale (mRS) are widely used to evaluate stroke severity and functional outcomes, respectively, in clinical practice and research settings.[1–4] The NIHSS is the current standard for quantifying stroke-related impairments to guide treatment decisions and to determine subsequent neurological improvement and deterioration, while the 90-day mRS is the most commonly used primary outcome measure in randomized controlled trials in stroke. As such, accurate performance of these scores has become an essential competency for stroke clinicians and trial personnel.[2,5,6]

In the 1980s, the NIHSS was developed as a systematic method for evaluating stroke impairments[7] and rose to prominence, particularly after its use in the landmark National Institute of Neurological Disorders and Stroke alteplase trial.[2,8–10] The use of the NIHSS was propelled forward by validation studies and the development of a validated videotape-based training program, making this scale easily applied to many contexts and locations (e.g., multi-site trials). In the original NINDS rt-PA Stroke Trial study, raters were required to recertify 6 months after initial certification and then yearly thereafter.[9,11] Yearly recertification has remained the current standard.[2,12] The mRS was published in 1988 after modifications were made to the original 1957 Rankin Scale in order to increase its comprehensiveness and applicability to modern stroke practice[3,13–15] Compared to the NIHSS certification and training requirements, the mRS training and certification landscape has been more heterogeneous, though mRS certification is typically required for participation in stroke trials and occasionally in clinical practice as well.[16] Online video scenario-based certification is widely used, and yearly recertification is typically recommended, particularly for those participating in clinical trials.

To date, however, there has been little critical examination of the benefits of such training requirements. Such examination is crucial, particularly given the substantial time commitment required for mandatory training, certification and annual recertification. This systematic review sought to determine if NIHSS or mRS training, re-training, certification or recertification led to improvements in rater performance and user-reported metrics.

Methods

This review is reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The protocol for this review was registered on November 8, 2023, on PROSPERO (registration number: CRD42023476934).

Search strategy

Searches (without date or language restrictions) of MEDLINE, EMBASE and Cumulated Index in Nursing and Allied Health Literature (CINAHL) databases were performed from inception until March 24, 2024. The complete search strategy is listed in the online Supplemental Material and included the following terms (and associated synonyms/mapped subject subheadings): "National Institutes of Health Stroke Scale," "Rankin scale," "Training," "Certification," "Accreditation," "Education," "Teaching" and "Learning." Reference lists of the included studies were hand searched for other relevant items that were not retrieved in the original search.

Table 1. Inclusion and exclusion criteria

Inclusion criteria:
1. Includes training and/or certifications as an intervention in the study (either prospectively or retrospectively)
2. Reports an outcome (see Table 2) and compares this to another group (either control group or other intervention group or historical group [i.e., same participant group, pre- and post-intervention])
3. At least a subset of participants is clearly identified as clinicians and/or allied health professionals (including students in these fields)
Exclusion criteria:
1. No comparator group included
2. Non-English language
3. Use nonstandard versions of the scales (e.g., simplified National Institutes of Health Stroke Scale)

Table 2. Predefined outcomes for inclusion criteria

• Reliability
o Inter-rater reliability
o Intra-rater reliability
• Accuracy
o Score accuracy relative to benchmark/gold standard
o Error rates and types
• End user comfort/confidence in using tool
• Certification success/pass rate (for studies examining training)
• Or any other outcomes related to effectively ascertaining the National Institutes of Health Stroke Scale or modified Rankin Scale that may arise during the process of screening

Study selection

Title and abstract screening, followed by full-text review, were independently completed by two authors (DM and DK) with conflicts resolved by a third author (AG). Inclusion and exclusion criteria are reported in Table 1. Studies were included if they (1) included training or certification as an intervention, (2) reported outcomes (pre-defined outcomes of interest included in Table 2) in reference to a comparator group (i.e., control group, other intervention group or historical group or pre- vs. post-intervention comparison) and (3) included stroke clinicians or allied health professionals (including students in these fields). Outcomes of interest included measures related to reliability, accuracy, user confidence and certification/training pass rate. Any other outcomes related to effectively conducting the NIHSS or mRS scoring systems that arose during the process of screening were specified in the protocol to be included; no such additional outcomes were found.

Studies were excluded if there was no comparator group (e.g., simply a descriptive report of NIHSS pass rate or of inter-rater reliability within a single trained group) or were not published in the English language. Published peer-reviewed conference abstracts and proceedings were included. Screening was performed using Covidence software.[17]

Data extraction

Given the broad inclusion criteria of this review, it was anticipated that there would be significant heterogeneity between included studies in terms of interventions, study groups and outcome measures. As such, a broad and narrative style of data extraction was pursued. We collected the following data: study design; training, certification and recertification details; participants' health professional role; level of training or experience; outcomes; and key

https://doi.org/10.1017/cjn.2025.10111

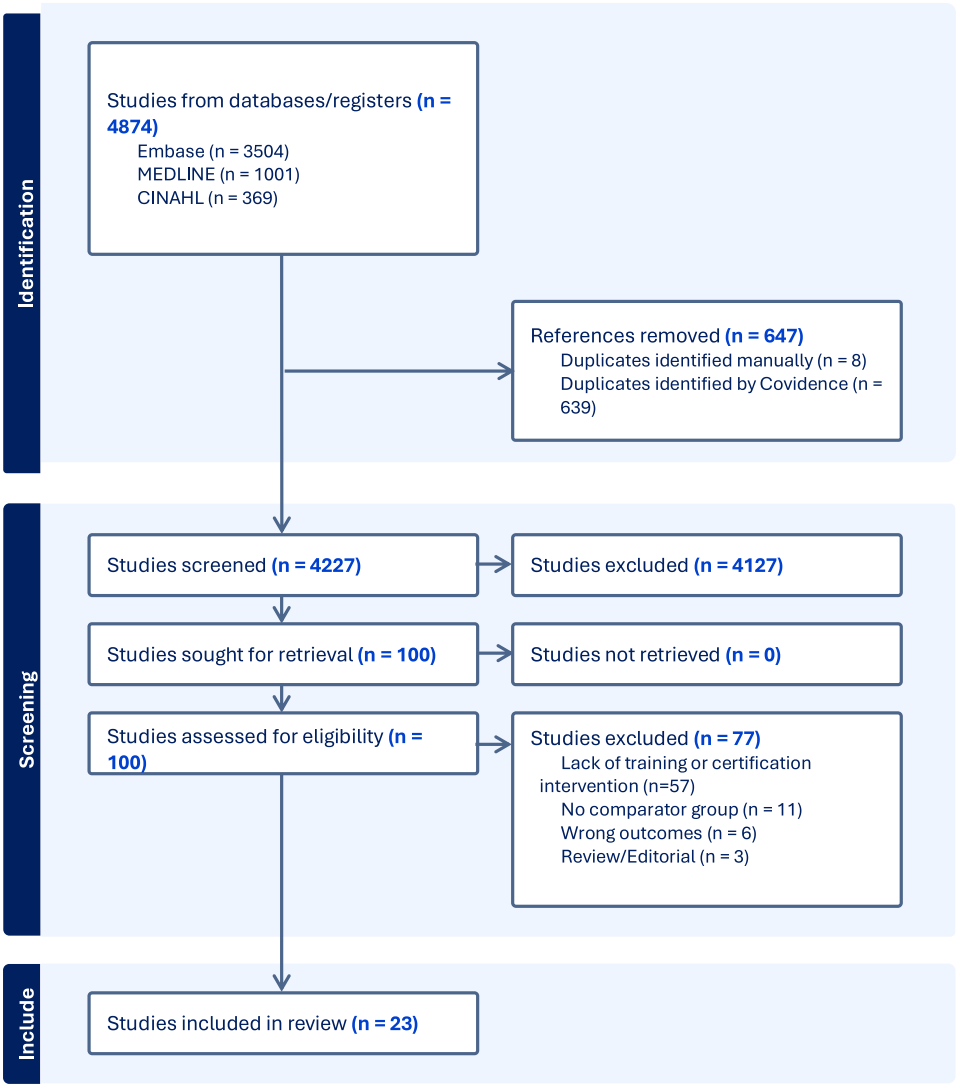

Figure 1. PRISMA diagram of the included studies.

findings. Data extraction was completed in duplicate by two authors (DK and DM), with any conflicts resolved by consensus.

Data synthesis

Given the diversity of interventions, study groups and outcomes, a narrative style data synthesis was used. In the review registration protocol, methods for a meta-analysis were outlined, but given the heterogeneity observed across studies, a meta-analysis was not considered to be appropriate.

Assessment of study quality and risk of bias

Risk of bias assessment of the included studies was performed using the ROBINS-I (Risk Of Bias In Non-randomized Studies - of Interventions)[18] tool for observational studies and using the CROB (Cochrane Risk of Bias)[19] tool for any included randomized control trial.

Results

After removal of duplicates, 4227 studies were screened, 100 of which were assessed for full-text eligibility, with 23 meeting criteria for inclusion in this review. Of the studies excluded, the majority (56/77) were excluded due to a lack of a training, certification or recertification intervention. Other reasons for exclusion included the absence of an appropriate outcome or a lack of a comparator group. The PRISMA flow diagram is shown in Figure 1.

Study characteristics

Tables 3–6 summarize the characteristics of the included studies. In total, 23 studies were included, of which 10 (43%)[20–29] were conference abstracts. Publication dates ranged from 1997 to 2023, with 16 of 23 (70%)[12,20–28,30–35] being published after 2014. The majority were observational studies (18/23, 78%),[8,12,20–29,33,36–39] with 5 (21%)[30–32,34,40] being randomized controlled trials (Table 3). Overall, there was a wide range of the number of participants in included studies, spanning from 4 (39) to 1,313,733.[12]

Seven of 23 (30%)[23,25,27,28,34,35,39] studies included exclusively physicians as participants, of which 5 of 8 (63%)[23,25,27,28,34] included only physicians in training (residents or medical students). An additional 6 of 23 (26%)[8,12,33,36–38] studies included physicians as well as other health professionals. Seven of 23 (30%)[20–22,24,29,30,40]

https://doi.org/10.1017/cjn.2025.10111

Table 3. Included studies related primarily to initial NIHSS training

NIHSS training	Author and publication date	Study design	Article type	Training descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	Harring et al., 2023[31]	Randomized control trial	Manuscript	All students enrolled received validated NIHSS training program for paramedics (45 min lecture and practice demonstration by certified physician, followed by 45 min simulation session). Participants then participated in voluntary simulation practice of NIHSS (GameStroke game for intervention group; in-person simulation for controls). After 2 months, participants then participated in a NIHSS clinical proficiency test.	50 paramedic students	Game-based versus in-person NIHSS simulation.	Accuracy (mean difference from true NIHSS score)	No statistically significant difference in the mean difference from the true NIHSS score was 0.64 in the game group and 0.69 in the control group.
	Koka et al., 2020[32]	Randomized control trial	Manuscript	Comparing an author-designed interactive e-learning module with the original NIHSS video training.	39 paramedics	E-learning module group versus control group	Study quiz at the end of training with 50 questions aimed at assessing overall NIHSS knowledge (primary outcome)	E-learning participants performed better than controls in the post-study quiz (36/50 vs. 33/50, $p = 0.04$).
	Suppan et al., 2020[34]	Randomized control trial	Manuscript	Participants completed either a novel e-learning NIHSS training module or standard didactic video training.	75 medical students	Comparison of e-learning module group versus traditional didactic videotape learning group.	Performance on a post-training knowledge quiz (score/50)	Participants in the e-learning group performed better than those in the traditional video training group (38 correct answers, versus 35 correct answers, $P < .001$)
	Dancer et al., 2017[30]	Randomized control trial	Manuscript	The trained NIHSS group watched a standard 55-minute NINDS NIHSS training DVD. Untrained groups received no training in stroke assessment. Note this study also examined a modified NIHSS scoring system (NIHSS-PE) and the effect of training.	122 nursing students	Trained versus untrained NIHSS group.	Accuracy (deviation between each participant's total score and experts' total score)	There was a numerical increase in deviation from expert scores in participants untrained in NIHSS compared to those who were trained (4.0 vs. 2.9 per NIHSS) though statistical significance was not reported. Pooled results of trained NIHSS-PE (a modified scale) and NIHSS versus untrained users were analyzed, and the authors report trained users had scores significantly closer to the expert scores. Score deviation from expert was 2.7 ± 2.3 in the trained group versus 3.5 ± 2.5 in the untrained group ($p = 0.011$).
	Chiu et al., 2009[40]	Randomized control trial	Manuscript	This study compared two NIHSS training programs: Interactive Computer-Assisted Instruction (ICAI) versus Instructor-led Videotape Learning Program (IVLP).	84 nurses	Pre-training scores versus post-training scores. IICAI versus IVLP	Score verification unit (measure of percentage of correct scores). Learner satisfaction (16-item questionnaire)	Both groups' scores on the assessment of correctness (SVU) significantly increased ($F = 35.50$, $p = 0.00$) after training. There was no significant difference between the changes in the two groups ($F = 0.02$, $p = 0.89$).

Table 3. Included studies related primarily to initial NIHSS training (*Continued*)

NIHSS training	Author and publication date	Study design	Article type	Training descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	Schmulling et al., 1998[39]	Observational study	Manuscript	Two raters were previously trained in NIHSS use (using standard training video); the other two raters received no instruction other than the NIHSS form itself, which provides few details on performing the stroke scale.	4 neurologists	Trained raters versus untrained raters	Interobserver reliability	The interobserver reliability (k) of trained raters was 0.61 (SD = 0.17) and 0.33 (SD = 0.22) among untrained raters.
	Goldstein et al., 1989[8]	Observational study	Manuscript	Participants took part in a training session on the use of the NIHSS, which included written instructions as well as standardized videotaped patient examinations.	59 (30 physicians, 29 study coordinators)	Previous exposure to NIHSS training tapes	Levels of agreement (assessed with intraclass correlation coefficients)	Levels of agreement were not affected by previous exposure to NIHSS training and certification tapes.
	Parker et al., 2023[25]	Observational study l	Abstract	Full day of simulation training with six common stroke scenarios.	15 (resident physicians)	Pre- vs post-simulation scores	Confidence (unit/measure not stated)	15 of15 reported increased confidence in their ability to perform the NIHSS.
	Graves et al., 2021[22]	Observational study	Abstract	In-person NIHSS training with live demonstration of exam technique.	896 nurses	Prior online standardized NIHSS versus in-person NIHSS training program	Accuracy (not further defined)	Authors state that evaluations show improved learner confidence, skills and knowledge (though no data or description of the evaluations/metrics used are included).
	Shoemaker et al., 2019[26]	Observational study	Abstract	An in-person class was added in this study to supplement standard NIHSS didactic videotape training with in-person discussion and demonstrations of the NIHSS scale.	24 (nurses and paramedics)	Pre- versus post-class metrics	Comfort in performing the NIHSS (5-point Likert scale)	Overall confidence in performing the NIHSS improved from 2.1/5 before the class to 4.2 after the class (statistical significance not reported).
	Grace, 2013[21]	Observational study	Abstract	Educational program that included a review of relevant neuroanatomy, a review of NIHSS use and scoring on a simulated patient, followed by a debriefing discussion.	174 nurses	Pre- versus post-educational session	User-perceived NIHSS competency	Mean self-perceived NIHSS competency prior to the session was 3.28 compared to 3.90 post-session ($t(172) = 13.99$, $p<0.01$).
	Margiotta et al., 2018[23]	Observational study	Abstract	In-person 1-hour didactic training session on how to perform the NIHSS, followed by two simulated stroke cases.	9 incoming neurology residents	Pre- and post-training/simulation surveys	Resident confidence in performing the NIHSS	Prior to training, only 44% of residents indicated that they felt comfortable performing the NIHSS compared to after training where 100% of residents indicated that they felt confident performing the NIHSS.
	Wadhwa, 2017[27]	Observational study	Abstract	In-person participation in five simulated acute stroke cases.	36 junior neurology residents	Simulation group versus historical controls who did not participate in simulation training	Resident confidence in administering NIHSS (unit of measure not reported)	Significant improvement in NIHSS utilization was observed in the simulation cohort. Further details/data not described.

(*Continued*)

Table 3. Included studies related primarily to initial NIHSS training (*Continued*)

NIHSS training	Author and publication date	Study design	Article type	Training descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	Gill et al., 2016[20]	Observational study	Abstract	High fidelity patient simulation of in-hospital ischemic stroke.	8 intensive care unit nurses	Pre- versus post-simulation	Confidence in ability to perform the NIHSS	This abstract states that nurses were confident in their ability to perform the NIHSS though no specific results or data were reported.
	Wendell et al., 2018[28]	Observational study	Abstract	In-person participation in two simulated stroke patients.	10 junior neurology residents	Pre- and post-simulation survey	Comfort level of performing the NIHSS (5-point Likert scale)	Resident reported increased comfort in performing the NIHSS after participating in the simulation (3.35/5 vs. 4.25/5 $p = 0.03$).
	McDavid et al., 2015[24]	Observational study	Abstract	Standard online NIHSS training with the option of in-person additional training. This was followed by a competency evaluation, and those who failed the competency evaluation were required to complete remedial in-person training prior to re-taking the competency test.	114 nurses	Standard online training versus standard online training + in-person training	Competency test pass rate	8 of 9 (89%) of nurses who completed additional in-person training passed the competency evaluation compared to 69 of 105 (66%) who only completed standard online training.

NIHSS = National Institutes of Health Stroke Scale; NINDS = National Institute of Neurological Disorders and Stroke.

Table 4. Included studies related primarily to initial National Institutes of Health Stroke Scale (NIHSS) certification

NIHSS certification	Author and publication date	Study design	Article type	Certification descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	Lyden et al., 2009[38]	Observational study	Manuscript	Validation study of novel NIHSS training tapes (as described in Lyden et al., 2005[37])	8214 (nurses and physicians)	Previously certified users versus noncertified users. Novice users versus experienced users	Interclass correlation (ICC)	There was no difference in ICC between previously certified (0.82 [0.70–0.9]) and noncertified users (0.94 [0.80–0.97]).
	Lyden et al., 2005[37]	Observational study	Manuscript	Evaluation of a new set of NIHSS training and videotapes developed by the study authors.	112 raters (nurses and physicians)	Previously certified raters versus noncertified users	Agreement (ICC)	There was no difference in levels of agreement between previously certified users and noncertified users (data not reported). There was no difference in ICC between previously certified (0.92 [0.79–1]) and noncertified users (0.95 [0.87–1]).

Table 5. Included studies related primarily to repeat National Institutes of Health Stroke Scale (NIHSS) training or certification

NIHSS re-training or recertification	Author and publication date	Study design	Article type	Training/certification descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	McLoughlin et al., 2022[33]	Observational study	Manuscript	Examined NIHSS scoring when completed via telemedicine.	15 (nurses and physicians)	Comparing recertified participants against non-recertified participants	Inter-rater reliability	For five NIHSS items, there was similar reliability between certification sub-groups, better in the recertification group for five NIHSS items and worse in the recertification group for five items.
	Anderson et al., 2020[12]	Observational study	Manuscript	Training was completed via online training documents and videos. Data from three NIHSS certification vendors were included, one of which required training before certification or recertification, the second did not and the third adopted a training requirement part-way through the included study period.	1,313,733 physicians and nurses (distributions of each not published)	New versus repeat certification users. Pre-certification training versus no pre-certification training.	Changes in accuracy (measured by comparing each user total score to the correct total score). Changes in technical errors (9 pre-specified errors defined by study authors) over time.	There was no difference in accuracy or technical errors between first time certification users or repeat users overall. One vendor group (n = 255,147) that required repeat training before recertification showed a statistically significant but trivial decrease in technical error rate over repeat certification (0.014/year; $P < 0.05$) Vendor group two that did not require repeat training showed a statistically significant but clinically negligible increase in technical error rate over time (0.13 error/year; $P < 0.001$)
	Josephson et al., 2006[36]	Observational study	Manuscript	Retrospective data from all certification exams submitted to the National Stroke Association from December 1998 to August 2004.	7405 raters (physicians, nurses and other healthcare providers)	Repeated tests (same version of test, 1065/7405 raters, took the same version of the test 2–4 times).	Pass rate and level of agreement	Retaking the test did not improve agreement with the most common response (p = 0.78) nor improve the pass rate ($p = 0.85$).
	You et al., 2010[29]	Observational study	Abstract	Repetitive NIHSS training program (details not provided)	12 stroke unit nurses	Inter-rater reliability after first NIHSS training sessions versus after completion of repetitive sessions.	Inter-rater reliability/ concordance (exact unit/measure not specified)	Authors state reliability/concordance improved over the course of repetitive training though details/data not reported)

Table 6. Included studies related primarily to modified Rankin Scale training or certification

Modified Rankin Scale	Author and publication date	Study design	Article type	Training descriptions	Participants and professionals	Comparator groups	Outcome measure	Key findings
	Pozarowszczyk et al., 2023[35]	Observational study	Manuscript	No formal training included though compared pairs of physicians group by certification status (either both pair members certified or a single member certified)	102 stroke patients evaluated by both a stroke unit physician and a rehabilitation ward physician (number of included physicians not included)	Grouped by physician pair (certified pair vs. pairs containing one certified user and one noncertified user)	Agreement (measured as a percentage [overall agreement] and kappa)	The was no significant difference in agreement between pairs of certified raters and pairs of raters including a single certified rater

studies included only nurses (or nursing students), 2 of 23 (9%)[31,32] included paramedics (or paramedic students) and 1 of 23 (4%)[26] included both paramedics and nurses.

Risk of bias assessment

Of the included observational studies, 10 of 18 (56%) were deemed to have a critical risk of bias. The 10 included abstracts in this review were the same 10 deemed to have a critical risk of bias. Three of 18 (17%)[12,37,38] observational studies were deemed to have moderate risk of bias, and 5 of 18 (28%)[8,33,35,36,39] were rated as serious risk of bias. Of the five included randomized control, three of five (60%) were rated as having high risk of bias,[31,34,40] and two of five (40%) were rated as having some concern for bias.[30,32] A summary graphic of the risk of bias assessment is included in the supplemental online Supplemental Material (Supplemental Figures 1 and 2)

Study findings

Twenty-two of 23 included studies were related to the NIHSS, with just one included mRS study.[35] Twelve of 23 studies examined the effect of training compared to no training (i.e., no formal instruction or exposure to training tapes and simply access to the standard NIHSS or mRS scoring form that contains limited instructions). Of these training studies, three[30,35,39] examined performance among different groups of participants (trained or untrained), and the other nine used historical controls (i.e., the same participant group before and after a training intervention). Two of the three studies examining different cohorts of participants who were trained or untrained reported numerical differences in outcomes between trained and untrained users, but statistical tests were not reported in either.

The five RCTs included four studies of different training approaches (game-based vs. in-person,[31] e-learning vs. original video training[32,34] and computer-assisted instruction vs. instructor-led video learning).[40] Only one study randomized participants to training versus no training[30] and examined NIHSS score performance. This 2017 study[30] focused on nursing students. The authors reported a numerical deviation from expert scores that was greater in the untrained group (4.0 vs. 2.9 per NIHSS score) though confidence intervals and statistical analyses were not reported. Of note, Dancer et al. (2017)[30] reported a statistically significant increase in deviation from expert scores in untrained participants versus trained when a pooled analysis of both NIHSS and a modified NIHSS scale (NIHSS-PE [Plain English]) was completed and the authors report trained users had scores significantly closer to the expert scores (score deviation from expert was 2.7 ± 2.3 in the trained group vs. 3.5 ± 2.5 in the untrained group [$p = 0.011$]). An observational study compared trained versus untrained NIHSS raters[39] but reported a numerical difference in agreement among trained raters compared to untrained raters, in only four participants.[39]

The only mRS study compared groups of trained versus untrained raters[35] and showed no statistical difference between pairs of trained raters or a trained rater paired with an untrained rater.[35]

Nine[21–23,25–29,40] studies examined historical cohorts by comparing pre-training versus post-training scores, eight[21–23,25–29] of which were conference abstracts that generally commented on participant confidence measures pre- and post-training. Five studies [21,23,25,26,28] (all abstracts) reported numerical or statistically significant increases in participant (usually resident physicians) confidence in performing the NIHSS. The only full-length manuscript analyzing pre- and post-test scores was by Chiu et al. (2009)[40] and was designed to compare two NIHSS training methods among a group of nurses (computer assisted vs. instructor led), although the authors do report a significant increase in score verification unit (a surrogate of accuracy) after training in both groups.

No studies examined the effect of initial certification though seven[8,29,33,36–38] studies examined the effect of re-training or recertification. The largest of these was published by Anderson et al. (2020),[12] which included results of 1,313,733 unique NIHSS certification tests. In this study, no difference was observed in accuracy or error rate between first-time certification users compared to users completing repeat certification. The study did show a small but statistically significant (0.014/year, $P < 0.05$ [confidence interval not reported]) decrease in error rate from one year to the next for groups that required repeat online training prior to each repeat certification exam. On the other hand, there was a similar small (0.013/year, $P < 0.001$ [confidence interval not reported]) but statistically significant increase in error rate, compared to prior performance, among a group that did not require repeat training prior to recertification. Note that this study compared results within groups (i.e., customers of different NIHSS training vendors) and with historical controls within these groups rather than statistically comparing between trained versus untrained groups. The remaining six[8,29,33,36–38] included studies found no significant change in reliability or agreement measures with repeat training and/or repeat certification. For example, a study by Lyden et al. (2009),[38] which included 2416 previously certified NIHSS raters and 1414 uncertified raters undergoing online certification/recertification with required pre-training, showed no statistical difference in reliability between previously certified and first-time certification users.

Five studies[24,26,31,32,40] examined differences between types of training, which included current standard training, novel computer-assisted methods and instructor-led in-person training. Two of five[32,40] found statistically significant benefit in the novel computer module/e-learning groups compared to in-person or traditional online methods. Specifically, Koka et al. (2020)[32] report that e-learning participants performed better than controls on a post-study quiz (36/50 vs. 33/50 correct, $p = 0.04$), and Chiu et al. (2009)[40] report an increase in the percentage of correct scores ($p = <0.01$) after their novel e-learning training. One of five[31] studies showed no difference between a computer module group and instructor-led group, and two of five[24,26] reported improved NIHSS skills with the addition of in-person training. Specifically, a conference abstract by McDavid et al. (2015)[24] reported an increased pass rate on competency evaluation (89% in face-to-face training group vs. 68% in the online group), although the sample size was small and statistical analysis was not reported. A conference abstract by Shoemaker et al. (2019)[26] reported an increase in user confidence after in-person training from 2.1 to 4.2 on a 5-point Likert scale, although again, statistical analysis was not reported.

Discussion

The results of this review highlight the limited and heterogeneous evidence for current NIHSS and mRS training and certification practices. Of the 12 included studies examining NIHSS training, only 2 studies showed an increase in accuracy of NIHSS scoring after training, and a single study showed a very small decrease in

https://doi.org/10.1017/cjn.2025.10111

year-to-year error rate after re-training. The remaining studies, which included large observation studies, did not show improvement in accuracy or reliability of NIHSS scores with training. No included studies demonstrated a benefit of NIHSS certification or recertification. A number of studies reported subjective improvements in user confidence after training although these were limited to conference abstracts judged to be at critical risk of bias. Only one mRS study met our criteria, and this showed no significant difference in agreement between pairs of certified raters and pairs of certified versus noncertified raters.

Several of the larger studies[12,36–38] included in this review pooled results of multiple healthcare providers (largely physicians and nurses) though it is important to consider that different healthcare provider groups (physicians, nurses, pre-hospital providers) likely have different experience with the NIHSS in daily practice and may benefit from different training and certification standards. Given that the only studies showing improvement in NIHSS accuracy were those including only nurses or nursing students, this may suggest some benefit to NIHSS training among these groups. Studies that included physicians only were largely conference abstracts and generally commented on confidence in performing the NIHSS scale. For physicians in training, who constituted the majority of the physicians included in these studies, there seems to be a signal that training may increase user confidence in scale performance although interpretation of these results is limited by the high risk of bias among these reports.

Taken together, the results of this review highlight important deficiencies of the evidence behind current NIHSS training and certification practices. At the very least, it seems reasonable to revisit current annual recertification requirements for the NIHSS and mRS for clinicians practicing in stroke. For example, in the study by Anderson et al. (2020),[12] which had the largest sample size included in this review, the authors suggested that NIHSS mastery for physicians and nurses is stable over time, that repeat training and certification lead to no clinically significant differences and that the required interval for recertification should be lengthened.[12] Based on the current review, there is little evidence to support recertification at all. A first certification may be reasonable to increase user confidence. Such an approach has been adopted by the Clinical Dementia Rating Scale,[41] which requires initial certification though no mandatory recertification; additionally, other scales that are recognized as clinical standards (e.g., the Glasgow Coma Scale) do not require mandatory training or certification.

Limitations of this review include the heterogeneity and generally high risk of bias of the included studies. Yet, it is precisely because of a lack of high-quality evidence that certification standards must be questioned. We opted to be comprehensive in the types of studies we included in order to provide as complete a picture as we could of the available evidence in this space.

Finally, it is worth noting that medicine is rampant with resource-intensive practices with little evidence for their use.[42] It is important for health and research professionals to critically examine current practices and standards in order to seek evidence that justifies the current practice or, in the absence of this, seriously reconsider the practice in question. Revising the current training and certification practices has the potential to improve clinical trial efficacy and reduce investigator burden. In this case, however, while there is a lack of evidence for current NIHSS and mRS training regimens, this does not necessarily mean that these practices are ineffective; however, it does underscore the need for higher-quality data to continue justifying the current practices as well as to seek possible evidence-based alternative practices. Questioning the current requirements seems reasonable, and effort should be made to achieve professional consensus on more efficient and rational strategies that maintain the validity of these scales. Pending higher quality evidence, it is important for professional stroke organizations and trial steering committees to be transparent about their proposed approaches to NIHSS and mRS certification and their rationale in published statements, in order to promote consistency across sites in national and, ideally, international trials. Such concerted approaches would also help provide reassurance and a united front to regulatory bodies and clinical trial sponsors as opposed to a haphazard approach of individual sites refusing to pursue recertification.

Conclusions

The results of this review highlight the sparsity and heterogeneity of studies examining whether NIHSS or mRS training, re-training, certification or recertification improves the reliability and accuracy of ratings or other user metrics. In the case of the NIHSS, there is some evidence to suggest a lack of benefit of the current training and certification regimen in terms of accuracy and reliability of the ratings. For the mRS, more work is clearly needed to quantify the effects of training and/or certification in general. Overall, there is an absence of evidence to support current NIHSS and mRS certification practices; at the very least, recertification requirements should be reconsidered pending the provision of robust evidence.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/cjn.2025.10111

Data availability statement. Data is available upon request.

Acknowledgments. None.

Author contributions. D.M. was responsible for primary manuscript writing, methodology development, study screening and data extraction.

D.K. was responsible for manuscript writing, study screening and data extraction.

R.H.S. provided a review of the initial manuscript and subsequent writing and edits.

S.B.C. provided a review of the initial manuscript and subsequent writing and edits.

A.G. provided oversight of the entire project as well as manuscript development.

Funding statement. No funding to declare.

Competing interests. D.M. reports no competing interests.

D.K. reports no competing interests.

R.H.S. reports salary support by the Department of Medicine (Sunnybrook HSC, University of Toronto); grants by Bastable-Potts Chair in Stroke Research, CIHR, NIH and Ontario Brain Institute; participation in an advisory board for Hoffman LaRoche Inc.; and stock options with FollowMD Inc.

S.B.C. reports grants from the Canadian Institute of Health Research during the conduct of the DOUBT study and grants from the Heart and Stroke Foundation of Canada, Genome Canada and Boehringer Ingelheim outside the submitted work

A.G. reports membership in editorial boards of *Neurology, Neurology: Clinical Practice, and Stroke*; research support from the Canadian Institutes of Health Research, Alberta Innovates, Campus Alberta Neurosciences, Government of Canada – INOVAIT Program, Government of Canada – New Frontiers in Research Fund, Microvention, Alzheimer Society of Canada, Alzheimer Society of Alberta and Northwest Territories, Heart and Stroke Foundation of Canada, Panmure House, Brain Canada, MSI Foundation and

the France-Canada Research Fund; payment or honoraria for lectures, presentations or educational events from Alexion, Biogen and Servier Canada; and stock or stock options in SnapDx Inc. and Collavidence Inc. (Let's Get Proof).

Ethical statement. Systemic review – ethics and informed consent not required.

References

1. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin Scale: implications for stroke clinical trials. *Stroke*. 2007;38:1091–1096.
2. Lyden P. Using the National Institutes of Health Stroke Scale. *Stroke*. 2017;48:513–519.
3. Kasner SE. Clinical interpretation and use of stroke scales. *Lancet Neurol*. 2006;5:603–612.
4. Broderick JP, Adeoye O, Elm J. Evolution of the modified Rankin Scale and its use in future stroke trials. *Stroke*. 2017;48:2007–2012.
5. Taylor-Rowan M, Wilson A, Dawson J, Quinn TJ. Functional assessment for acute stroke trials: properties, analysis, and application. *Front Neurol*. 2018;9:191.
6. Maguire J, Attia J. Which version of the modified Rankin Scale should we use for stroke trials? *Neurology*. 2018;91:947–948.
7. Brott T, Adams HP, Olinger CP, et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke*. 1989;20:864–870.
8. Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. *Arch Neurol*. 1989;46:660–662.
9. 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;25:2220–2226.
10. Asplund K. Clinimetrics in stroke research. *Stroke*. 1987;18:528–530.
11. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med*. 1995;333:1581–1587.
12. Anderson A, Klein J, White B, et al. Training and certifying users of the National Institutes of Health Stroke Scale. *Stroke*. 2020;51:990–993.
13. New PW, Buchbinder R. Critical appraisal and review of the Rankin scale and its derivatives. *Neuroepidemiology*. 2006;26:4–15.
14. Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke*. 1999;30:1538–1541.
15. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604–607.
16. Quinn TJ, Lees KR, Hardemark HG, Dawson J, Walters MR. Initial experience of a digital training resource for modified Rankin Scale assessment in clinical trials. *Stroke*. 2007;38:2257–2261.
17. Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org.
18. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
19. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928–d5928.
20. Gill R, Rasmussen T, Garg R, Ray J, McCoyd M, Ruland S. Abstract TP312: simulation based education for neurology nurses to improve in-hospital stroke emergency performance. *Stroke*. 2016;47:ATP312–ATP312.
21. Grace M. Abstract TMP85: enhancing stroke nursing education through simulation. *Stroke*. 2013;44:ATMP85–ATMP85.
22. Graves AM, Jones J, Bragg A. Abstract P848: a better way to NIHSS. *Stroke*. 2021;52:AP848–AP848.
23. Margiotta M, Wilhour D, D'Ambrosio R, Pineda C, Rincon F. Abstract WP295: improving resident confidence and efficiency during stroke alerts through simulation training. *Stroke*. 2018;49:AWP295–AWP295.
24. McDavid JC, Bellamy LM, Thompson CJ. Abstract NS12: is online NIHSS certification enough training. *Stroke*. 2015;46:ANS12–ANS12.
25. Parker H, Asher G, Arnold B, Hindley E, Hardern K. 35 is that the medical registrar?: bridging the gap from medical trainee to medical registrar with a simulation programme. *BMJ Lead [Internet]*. 2023;7(Suppl 1):A21–A22. https://bmjleader.bmj.com/content/7/Suppl_1/A21.
26. Shoemaker A. Abstract TP485: increasing comfort with the National Institute of Health Stroke Scale Performance. *Stroke*. 2019;50(Suppl_1), ATP485–ATP485.
27. Wadhwa A, Katyal N, Singh NN. Abstract WP316: stroke simulation improves resident confidence in acute stroke/TIA management. *Stroke*. 2017;48:AWP316–AWP316.
28. Wendell L, Reznik M, Lindquist D, et al. Code stroke simulation training benefits junior neurology residents (P3.016). *Neurology*. 2018;90:P3.016.
29. You SK, Song YA, Park HS, et al. Implementation of repetitive nihss training sessions for stroke unit nurses to improve the predictive probability to the assessment of neurologic deterioration in patients with acute stroke. *Cerebrovascular Diseases*. 2010;29(Suppl. 2):330.
30. Dancer S, Brown AJ, Yanase LR. National Institutes of Health Stroke Scale in plain English is reliable for novice nurse users with minimal training. *J Emerg Nurs*. 2017;43:221–227.
31. Harring AKV, Røislien J, Larsen K, et al. Gamification of the National Institutes of Health Stroke Scale (NIHSS) for simulation training—a feasibility study. *Adv Simul*. 2023;8:4.
32. Koka A, Suppan L, Cottet P, Carrera E, Stuby L, Suppan M. Teaching the National Institutes of Health Stroke Scale to paramedics (E-learning vs video): randomized controlled trial. *J Med Internet Res*. 2020;22:e18358.
33. McLoughlin A, Olive P, Lightbody CE. Reliability of the National Institutes of Health Stroke Scale. *Br J Neurosci Nurs*. 2022;18:S3–S10.
34. Suppan M, Stuby L, Carrera E, et al. Asynchronous distance learning of the National Institutes of Health Stroke Scale during the COVID-19 pandemic (E-learning vs video): randomized controlled trial. *J Med Internet Res*. 2021;23:e23594.
35. Pożarowszczyk N, Kurkowska-Jastrzębska I, Sarzyńska-Długosz I, Nowak M, Karliński M. Reliability of the modified Rankin scale in clinical practice of stroke units and rehabilitation wards. *Front Neurol*. 2023;14:1064642.
36. Josephson SA, Hills NK, Johnston SC. NIH stroke scale reliability in ratings from a large sample of clinicians. *Cerebrovasc Dis Basel Switz*. 2006;22:389–395.
37. Lyden P, Raman R, Liu L, et al. NIHSS training and certification using a new digital video disk is reliable. *Stroke*. 2005;36:2446–2449.
38. Lyden P, Raman R, Liu L, Emr M, Warren M, Marler J. National Institutes of Health Stroke Scale certification is reliable across multiple venues. *Stroke*. 2009;40:2507–2511.
39. Schmülling S, Grond M, Rudolf J, Kiencke P. Training as a prerequisite for reliable use of NIH Stroke Scale. *Stroke*. 1998;29:1258–1259.
40. Chiu SC, Cheng KY, Sun TK, et al. The effectiveness of interactive computer assisted instruction compared to videotaped instruction for teaching nurses to assess neurological function of stroke patients: a randomized controlled trial. *Int J Nurs Stud*. 2009;46:1548–1556.
41. Morris, J.C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*. 1993;43(11),2412–2414.
42. Feldman LS. Choosing Wisely(®): things we do for no reason. *J Hosp Med*. 2015;10:696–696.