Is the NIHSS Certification Process Too Lenient?

Nancy K. Hills^a S. Andrew Josephson^a Patrick D. Lyden^c
S. Claiborne Johnston^{a, b}

Stroke Sciences Group, Departments of ^aNeurology and ^bEpidemiology and Biostatistics, University of California, San Francisco, Calif., and ^cVA Medical Center, San Diego, and University of California San Diego Stroke Center, San Diego, Calif., USA

Key Words

Stroke • NIH Stroke Scale, reliability • Scoring the NIHSS certification test

Abstract

Background and Purpose: The National Institutes of Health Stroke Scale (NIHSS) is a widely used measure of neurological function in clinical trials and patient assessment; inter-rater scoring variability could impact communications and trial power. The manner in which the rater certification test is scored yields multiple correct answers that have changed over time. We examined the range of possible total NIHSS scores from answers given in certification tests by over 7,000 individual raters who were certified. **Methods:** We analyzed the results of all raters who completed one of two standard multiple-patient videotaped certification examinations between 1998 and 2004. The range for the correct score, calculated using NIHSS 'correct answers', was determined for each patient. The distribution of scores derived from those who passed the certification test then was examined. **Results:** A total of 6,268 raters scored 5 patients on Test 1; 1,240 scored 6 patients on Test 2. Using a National Stroke Association (NSA) answer key, we found that correct total scores ranged from 2 correct scores to as many as 12 different correct total scores. Among raters who achieved a passing score and were therefore qualified to administer the NIHSS, score distribu-

tions were even wider, with 1 certification patient receiving 18 different correct total scores. **Conclusions:** Allowing multiple acceptable answers for questions on the NIHSS certification test introduces scoring variability. It seems reasonable to assume that the wider the range of acceptable answers in the certification test, the greater the variability in the performance of the test in trials and clinical practice by certified examiners. Greater consistency may be achieved by deriving a set of 'best' answers through expert consensus on all questions where this is possible, then teaching raters how to derive these answers using a required interactive training module.

Copyright © 2009 S. Karger AG, Basel

Introduction

The National Institutes of Health Stroke Scale (NIHSS) has proven to be a valuable tool for assessing neurological impairment and predicting stroke outcome [1, 2]. The NIHSS is the culmination of a process that began with the need to develop a systematic means by which to clinically measure the severity of cerebral infarction in light of the introduction of new therapies for acute stroke. Previously limited clinical measurements and unvalidated rating scales were formalized in a scale developed at the University of Cincinnati Stroke Center as a system for

examining patients with acute cerebral infarction [1]. The scale was developed further by researchers at the National Institute of Neurologic Disorders and Stroke (NINDS) to quantify patient status and outcomes in stroke clinical trials. Early versions of the scale, tested among stroke researchers on both live [1, 3] and videotaped patients [4], all showed poor reliabilities for items intended to score facial movement, limb ataxia, and dysarthria, with moderate or excellent agreement on the remaining items. In 1994, a videotaped training and certifying instrument was devised in order to systematically train investigators and coordinators of clinical trials to rate patients as consistently as possible on the NIHSS [6]. This is particularly important for multi-site clinical trials in which multiple investigators in different centers are measuring outcomes at variable times; inconsistency in the use of the scale might affect power and the ability to detect a true effect [3–5]. This development of a required, uniform certification process, and the use of actual patients for scoring, has been an important step in helping clinical trials in stroke become more rigorous.

This training/certification process was primarily aimed at skilled, trained stroke neurologists who were presumed to have prior experience in using the NIHSS. Over time, however, the NIHSS certification process has been expanded; as greater numbers of institutions seek Joint Commission certified stroke center status, an increasing number of those seeking certification to administer the scale have little or no training in neurology, and include study personnel who have no formal medical training [6, 7]. This group diverges appreciably from those for whom good test-retest reliability among various different raters was reported by Brott et al. [1]. It has become increasingly important, therefore, to determine whether the current training portion of the certification process adequately teaches all examiners, particularly those without prior training in stroke neurology or without medical training, to score patients comparably. In addition to extending beyond its originally intended administrators, the NIHSS has been utilized beyond its original and validated purpose into areas not currently supported by high-quality clinical evidence such as specifying clinical trial inclusion criteria [8, 9], defining clinically important change from pre- to post-treatment [1, 7], functioning as a primary [7, 10] or secondary [8, 11–16] endpoint in trials, determining differences between treatment groups in some studies [1, 17], assessing stroke severity, and planning patient care by determining appropriate treatment for stroke patients [18, 19]. While the scale should not be used for purposes for which it has not

been adequately validated, widespread documentation of such use places an even greater burden on the certification process.

However, neurologic examination items, which previously have been found difficult for raters to score, continue to be problematic. Some of these, when viewed in a video image, are difficult to characterize despite multiple-angle presentations. To overcome the inherent limitations of video technology, the scoring algorithm for the certification test allows for multiple acceptable answers to many of the 15 individual items on which each patient is scored. Because the overall NIHSS score is a summation of the scores on these 15 individual stroke characteristics, items with multiple allowable scores necessarily impact the reliability of the total score. This grading practice, whereby a single ‘best’ answer has never been agreed upon, was more feasible when certification was aimed primarily at trained stroke neurologists. However, given the much wider use of the scale, this practice may well be detrimental to the teaching aspect that is key to the certification process. We aimed to determine whether the method for setting criteria for passing the certification tests contributes to an unacceptable level of scoring variability which could then be carried over into use of the scale in actual clinical practice. Such a finding could potentially lead to improvements to the training segment of the certification process that might increase scoring consistency of the NIHSS.

Methods

Analyses in this paper are based on the original VHS certification tapes, developed in 1994. While the current certification videos have changed, thousands of raters were certified using the original version, thus impacting research data for a number of years. These original tapes consisted of a 45-min taped training program and two taped certification tests [20]. After watching the training tape, raters were asked to score 5 videotaped patients on Certification Tape 1. Six months later, each investigator was to review the training tape and score the 6 patients on Certification Tape 2. After completion of the NINDS rt-PA for Acute Stroke Trial, the National Stroke Association (NSA) obtained copies of the training and certification videotapes for scoring and determination of eligibility to administer the NIHSS outside of the original clinical trial. While a training module was provided, examiners were not required to complete it before attempting certification, and raters did not receive any immediate feedback on their performance, either during or after the test.

Scoring the NIHSS Certification Test

To score the test, an outlier method of grading was developed for use in the NINDS rt-PA for Acute Stroke Trial. Initially, answers chosen by at least 12% of the original 162 NINDS t-PA in-

investigators were deemed to be acceptable, i.e., 'correct'. Responses chosen by less than 12% were considered to be outliers [20]. Rater tests including fewer than five outliers on Test 1 or six outliers on Test 2 were considered passing; these raters were judged to be competent to administer the NIHSS during the trial. Given the expectation that real-world users would not be able to match answers given by experts who were present at the time the video was recorded, this score sheet was updated periodically (using the same 12% outlier rule and grading method) to reflect answers given by real world raters [P.D.L., pers. commun.].

NIHSS Data

In order to determine how much variability this grading system potentially introduced into the NIHSS score, we examined data from all certification tests submitted to the NSA between December 1, 1998 and August 24, 2004, using an NSA answer key that was based on the original NINDS scoring algorithm. The two available certification videotapes included a total of 11 unique patients, with 15 items (counting sub-items) for each patient. A total NIHSS score could be calculated only for those patients for whom all 15 questions were answered; we included only those certification tests for which a complete score could be calculated for every patient included in the test. For raters who took the *same* test more than once, we included only their first test, whether or not they received a passing score on this test.

Using NIHSS 'acceptable answers', we first calculated 'correct' scores for each of the 11 test patients. This consisted of summing the answers for the 15 items pertaining to each patient, taking into account those items for which multiple answers were considered to be acceptable. We thus determined for each videotaped patient the NIHSS scores that would be deemed correct. We then scored rater certification tests to determine which raters had passed the test, and examined the distribution of individual NIHSS patient scores in this group of raters (i.e., those judged to be qualified to administer the stroke scale).

Results

Our study included a total of 9,171 certification examinations completed over a 6-year period. Of these, 7,681 raters took Test 1 and 1,480 raters took Test 2. After excluding incomplete and repeat tests, 7,405 tests were eligible for these analyses: 6,268 unique raters each scoring 5 patients on Test 1 and 1,123 scoring 6 patients on Test 2. An unknown number of these users would have received training in using the scale, but this training was not required.

As a result of the outlier method of scoring, we found in our representative set of correct answers that 20 of 75 questions (27%) on Test 1 and 26 of 90 questions (29%) on Test 2 had multiple 'correct' answers. When combined into total NIHSS scores for each patient, a range of 'correct' total scores was possible for each, and no patient was characterized by a single best total score (table 1). In

Table 1. Range of 'correct' total scores possible

	Questions with multiple correct answers, %	Acceptable NIHSS scores	Spread (in points)
Test 1			
Patient 1	27	3–7	5
Patient 2	33	3–9	7
Patient 3	13	1–3	3
Patient 4	27	17–21	5
Patient 5	33	10–16	6
Test 2			
Patient 1	33	16–21	6
Patient 2	33	4–10	7
Patient 3	7	2–3	2
Patient 4	27	5–9	5
Patient 5	13	1–3	3
Patient 6	60	24–35	12

the best case, 1 patient had 2 acceptable total scores; in the most extreme case, a patient could be correctly assigned any of 12 different total scores. For this latter patient, as a result, any score between 24 and 35 would be considered correct by the established grading standards. The average number of 'correct' total scores per patient was 5.5.

By accepted grading standards, a rater was considered to have passed the certification test if no more than five outlying responses for Test 1 or six outliers for Test 2 were chosen. (While this reflects one outlier per patient on average, in actuality, outliers are totaled over all patients, and not on a per patient basis.) Of the 6,268 unique raters who scored Test 1, 4,396 (70%) received a passing score; 1,059 (94%) of the 1,123 raters who took Test 2 passed the certification test. Among these raters who achieved a passing score and therefore were certified to administer the NIHSS, the distribution of total scores for the individual patients varied even more widely. The minimum number of different assigned scores was 4 for 1 patient, while the maximum number was 18. These variations were independent of the severity of the symptoms (fig. 1).

Due to the limitations of video representations of patient examinations, the range of acceptable answers differed between individual examination components. Test items that are well known to be difficult to score consistently – aphasia, dysarthria, and facial paresis – ended up with multiple allowable answers as a result of the scoring algorithm used. For example, 9 of the 11 patients (82%) had multiple correct answers for the facial paresis item consistent with the fact that videography

Fig. 1. For each patient scored in the two certification tests, the distribution of scores indicates all the NIHSS scores that were considered to be correct. For example, for patient 6 on Test 2, a rater who assigned the patient a score of 24, and a rater who assigned that patient a score of 36 (as well as any score in between), would be considered to have calculated the 'correct' score. While separation *between* patients is acceptable, the variability of 'correct' scores *within* patients is not. The minimum number of different assigned scores was 4 for 1 patient (Test 2, Patient 5), while the maximum number was 18 (Test 2, Patient 6). Only scores derived by raters that achieved a passing score and therefore were certified to administer the NIHSS were included.

cannot portray facial weakness as accurately as an in-person examination. Of 11 patients, 8 (73%) had multiple correct items for the language item, a difficulty that has never been satisfactorily explained, and could be due to the method of training. As a result, 4 patients could be correctly classified as having either no or mild aphasia, 3 could be scored as having either mild or moderate aphasia, and 1 could be correctly identified as having either moderate or severe aphasia. In 3 patients, three answers were acceptable: 2 patients could be correctly classified as having either no, minor, or partial paralysis on the facial palsy item, and another patient could be categorized as having either no ataxia, ataxia in one, or in more than one limb. In another instance, a patient was 'correctly' diagnosed on the extinction item with either no abnormality or profound hemi-inattention, with the intermediary category counting as an outlier. For the items defining dysarthria and extinction, 5 of 11 patients (45%) had multiple correct answers. Items for which consistency would appear to be more easily reached, such as motor function, demonstrated inconsistency, with the left arm and right leg having multiple answers allowed in 27 and 36% of the patients, respectively. Level of consciousness was not deemed to be precisely determinable in all cases, with multiple answers acceptable for 2 of the 11 patients (18%).

Discussion

The video training and certification of raters on the NIHSS was developed by the NINDS to train clinical trial investigators and coordinators (primarily trained, skilled stroke neurologists) to measure data on patients reliably and consistently across sites and over time [7, 20]. Unfortunately, with wider use of self-certification of stroke neurologists, as well as many without training in the field of stroke neurology or even without any formal medical education, and often without required training, there has been considerable degradation in the quality of these measurements. Both the American Stroke Association [21] and the NSA [22] offer a free, online version of the test through their websites. However, the accompanying training module is not mandated prior to taking the certification examination. Even when utilized, the training module provides no immediate feedback to allow raters to recognize and learn from their mistakes before attempting the certification test, despite the fact that studies have shown feedback to be one of the most inexpensive and influential methods of learning available [23–25]. While not all institutions choose to rely solely on the self-certification process, the fact that the majority of raters will be certified in this manner makes it imperative that the certification process be as effective as possible

at teaching raters to score patients consistently on the NIHSS.

From its inception, problems with reliability in scoring of the NIHSS – whether live or via video – have been noted [3, 26]. When the certification tapes were first designed and tested, agreement beyond chance for the Test 1 (scored by 162 raters) was poor for one-third of the 15 questions comprising the scale, moderate for another third, and excellent only for the remaining third [20]. Although agreement was somewhat better on Test 2, only 64 original raters had actually scored these patients [20].

In 2001, a modified version of the NIHSS, the mNIHSS, was devised [27]. This test dropped problematic (facial weakness, ataxia, and dysarthria) and redundant (level of consciousness) items. The number of items showing poor agreement decreased to 3 of 22 (14%) items on the modified scale, but only 12 of 22 items (55%) exhibited excellent rater agreement (up from 40% for the 30 combined items of the original test). Despite general consensus about the problematic nature of some of the test questions and the apparent improvement of the scale by their removal, the modified scale has not been widely adopted. Neither has there been discussion about items not currently included in the scale. The NIHSS, which might benefit from the omission of some items, may also be strengthened by the addition of components to test attention, memory, and visual-spatial function, inclusions which could lead to more balanced non-dominant hemispheric scores.

In 2005, an updated version of the certification test for the full NIHSS was developed, encompassing a more balanced representation of patient disabilities and utilizing DVD technology, an improvement over the prior VHS technology [6]. The outlier grading system continued to be utilized, however, and inadequate rater agreement continued to be observed. Of 15 items, 13 proved to have either poor (2) or moderate (11) levels of rater agreement [6]. Clearly, a new approach to improve reliability is warranted. A possible alternative to changing the stroke scale itself may lie in endeavoring to improve the ability of the certification process to teach raters to derive answers that are as *consistent* as possible.

It is important to remember that a patient's final NIHSS score is a summation of 15 individually scored questions. The scoring system developed for NIHSS certification in the original clinical trial allows for multiple 'correct' answers to many of these questions (about 30% of questions across all videotaped patients per certification test version) to adjust for the limitations of video rep-

resentations of patient examinations. Furthermore, the 'correct' answers have been periodically reevaluated based on rater responses to the questions, such that the same videotaped patient could have different total correct scores depending on when the certification test was taken. As greater numbers of users enter the certification system with little or no training, the number of 'outliers' necessarily increases. This adaptive scoring approach results in greater variability in the *total* 'correct' scores, and ultimately reaches a point where a rater can be certified despite an unacceptably large number of mistakes on individual test items. In the original two videotapes, fully 73% of the 11 unique patients in the two certification tests have between 5 and 12 'correct' *total* scores, and no patient has a single correct total score. In the most extreme case, the videotaped patient can be correctly scored with an overall NIHSS score ranging from 24 to 35. Allowing substantial variability in individual patient scores increases the range of a passing score on the certification test, thus setting the pass/fail bar lower than should be acceptable. The certification process may therefore fail to teach raters to correctly categorize future stroke patients, thus leading to raters with less reliability, who therefore affect the results of clinical trials in which they participate. Outside of the clinical trial setting, this variability must have worsened over the ensuing years as well.

A precise scale ideally should have a single correct answer for each of the items tested. Consensus as to what constitutes a correct answer should, if possible, also take into account potential language and cultural differences in scoring, given the increasing number of international clinical trials. In a digital training resource for modified Rankin Scale (mRS) assessment, e.g., fully translated training packages, with native speakers overdubbing the patient interviews, have been made available for a variety of languages [28]. An analysis of the variability between countries using these packages revealed that differences in scoring were not entirely a function of language, but appeared to have some association with sociocultural factors related to perceptions of disability and handicap [29]. Others, however, in studying the effect of language on scoring of the NIHSS itself, have found that all venues – website, group, individual – yield similar results, and that items that film poorly in English remain poor in other languages, making the scale robust to such important variables [P.D.L., pers. commun.]. The absence of geographic information made such an effect impossible to assess in our current data.

However, it must be recognized that expecting raters to achieve absolute exact agreement on a 42-point scale

composed of 15 different questions per patient is unrealistic. Exactly what should constitute a passing score on a certification test is a complex question, given that each rater scores 15 questions on each of 5 or 6 patients. A training package for the mRS, a much less complex scale with only one question per patient, faces similar issues of agreement, and classifies a rater's ranking of a patient as either 'correct' if agreed upon by two experts and more than 50% of trainees, and as 'acceptable' if one expert and a substantial minority of assessors in the pilot study chose it. A passing score on the certification test is then based on the number of correct and acceptable responses given across all patients. This is roughly similar to the approach taken to each of the 15 individual questions on the NIHSS; however, allowing a number of 'acceptable' individual answers has been shown to lead to a very wide range in 'acceptable' total scores. A more effective approach would be to derive a correct, or best, answer for each individual item through expert consensus, keeping to an absolute minimum the number of responses where multiple answers cannot be avoided due to problems of interpretation caused by videography (if these items were not eliminated altogether). A determination could then be made as to what range of total patient scores would be acceptable.

Such a program would necessarily be supported by a computer testing procedure that includes an expanded, highly interactive and *required* training process. One possibility would allow raters to score a number of sample patients and receive feedback about their answer choices *after every question* before going on to take the actual certification test. They thus would have an opportunity (1)

to gain a better understanding of what experts consider to be the correct scoring of a given patient, (2) to see what mistakes they have made and why, and (3) to learn from their mistakes throughout the training process. This new system could then be evaluated in comparison to the current system, in order to determine if greater rater consistency was indeed achieved.

While such changes to the NIHSS certification process may require a greater time commitment on the part of those who wish to administer the test, the gains in reliability and consistency could well outweigh the disadvantages and costs of development. In the meantime, caution should be exercised in the use of the NIHSS for purposes for which it has not been validated. Until a new system can be evaluated and shown to have less variability in the raters who are certified, a patient's NIHSS score should not be used as a standard of exclusion or primary endpoint in clinical trials, nor should it be heavily weighted in the making of treatment decisions or communication between practitioners about patient status. Given the increasing pressure to extend the use of the NIHSS to include these functions, and the often-stated desire that it truly provide a common language with which to compare patients across different sites, raters, and points in time [22], such an effort would seem highly justified.

Acknowledgment

This study was supported by the NIH/NINDS (NS02254 and P50NS44148).

References

- 1 Brott T, Adams HP Jr, Olinger CP, Marler JR, Barsan WG, Biller J, Spilker J, Holleran R, Eberle R, Hertzberg V, et al: Measurements of acute cerebral infarction: a clinical examination scale. *Stroke* 1989;20:864-870.
- 2 Muir KW, Weir CJ, Murray GD, Povey C, Lees KR: Comparison of neurological scales and scoring systems for acute stroke prognosis. *Stroke* 1996;27:1817-1820.
- 3 Goldstein LB, Bertels C, Davis JN: Interrater reliability of the NIH Stroke Scale. *Arch Neurol* 1989;46:660-662.
- 4 Albanese MA, Clarke WR, Adams HP Jr, Woolson RF: Ensuring reliability of outcome measures in multicenter clinical trials of treatments for acute ischemic stroke. The program developed for the trial of ORG 10172 in acute stroke treatment (TOAST). *Stroke* 1994;25:1746-1751.
- 5 Josephson SA, Hills NK, Johnston SC: NIH Stroke Scale reliability in ratings from a large sample of clinicians. *Cerebrovasc Dis* 2006;22:389-395.
- 6 Lyden P, Raman R, Liu L, Grotta J, Broderick J, Olson S, Shaw S, Spilker J, Meyer B, Emr M, Warren M, Marler J: NIHSS training and certification using a new digital video disk is reliable. *Stroke* 2005;36:2446-2449.
- 7 NINDS rt-PA Stroke Study Group: Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. *N Engl J Med* 1995;333:1581-1587.
- 8 Furlan A, Higashida R, Wechsler L, Gent M, Rowley H, Kase C, Pessin M, Ahuja A, Callahan F, Clark WM, Silver F, Rivera F: Intra-arterial pro-urokinase for acute ischemic stroke. The PROACT II study: a randomized controlled trial. *Prolyse in acute cerebral thromboembolism. JAMA* 1999;282:2003-2011.
- 9 IMS II Trial Investigators: The Interventional Management of Stroke (IMS) II Study. *Stroke* 2007;38:2127-2135.

- 10 Clark WM, Wissman S, Albers GW, Jhamandas JH, Madden KP, Hamilton S: Recombinant tissue-type plasminogen activator (alteplase) for ischemic stroke 3–5 hours after symptom onset. The ATLANTIS study: a randomized controlled trial. Alteplase thrombolysis for acute noninterventional therapy in ischemic stroke. *JAMA* 1999;282: 2019–2026.
- 11 Clark WM, Williams BJ, Selzer KA, Zweifler RM, Sabounjian LA, Gammans RE: A randomized efficacy trial of citicoline in patients with acute ischemic stroke. *Stroke* 1999;30:2592–2597.
- 12 Davis SM, Lees KR, Albers GW, Diener HC, Markabi S, Karlsson G, Norris J: Selfotel in acute ischemic stroke: possible neurotoxic effects of an NMDA antagonist. *Stroke* 2000; 31:347–354.
- 13 Del Zoppo GJ, Higashida RT, Furlan AJ, Pesin MS, Rowley HA, Gent M: PROACT: a phase II randomized trial of recombinant pro-urokinase by direct arterial delivery in acute middle cerebral artery stroke. PROACT investigators. *Stroke* 1998;29:4–11.
- 14 Grotta J: Lubeluzole treatment of acute ischemic stroke. The US and Canadian Lubeluzole Ischemic Stroke Study Group. *Stroke* 1997;28:2338–2346.
- 15 Hacke W, Kaste M, Fieschi C, Toni D, Lesaffre E, von Kummer R, Boysen G, Bluhmki E, Hoxter G, Mahagne MH, et al: Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Study (ECASS). *JAMA* 1995;274: 1017–1025.
- 16 Hacke W, Kaste M, Fieschi C, von Kummer R, Davalos A, Meier D, Larrue V, Bluhmki E, Davis S, Donnan G, Schneider D, Diez-Tejedor E, Trouillas P: Randomised double-blind, placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). Second European-Australasian Acute Stroke Study investigators. *Lancet* 1998;352:1245–1251.
- 17 Johnston KC, Connors AF Jr, Wagner DP, Knaus WA, Wang X, Haley EC Jr: A predictive risk model for outcomes of ischemic stroke. *Stroke* 2000;31:448–455.
- 18 Schlegel D, Kolb SJ, Luciano JM, Tovar JM, Cucchiara BL, Liebeskind DS, Kasner SE: Utility of the NIH Stroke Scale as a predictor of hospital disposition. *Stroke* 2003;34:134–137.
- 19 Schlegel DJ, Tanne D, Demchuk AM, Levine SR, Kasner SE: Prediction of hospital disposition after thrombolysis for acute ischemic stroke using the National Institutes of Health Stroke Scale. *Arch Neurol* 2004;61:1061–1064.
- 20 Lyden P, Brott T, Tilley B, Welch KM, Mascha EJ, Levine S, Haley EC, Grotta J, Marler J: Improved reliability of the NIH Stroke Scale using video training. NINDS tPA Stroke Study Group. *Stroke* 1994;25:2220–2226.
- 21 American Stroke Association: NIH Stroke Scale training online, 2007.
- 22 National Stroke Association: NIH Stroke Scale, 2007.
- 23 Ende J: Feedback in clinical medical education. *JAMA* 1983;250:777–781.
- 24 Knowles M, Holton E, Swanson R: *The Adult Learner*. Houston/TX, Gulf Publishing, 1998.
- 25 Wood BP: Feedback: a key feature of medical training. *Radiology* 2000;215:17–19.
- 26 Kasner SE: Clinical interpretation and use of stroke scales. *Lancet Neurol* 2006;5:603–612.
- 27 Lyden PD, Lu M, Levine SR, Brott TG, Broderick J: A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. *Stroke* 2001;32:1310–1317.
- 28 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.
- 29 Quinn TJ, Dawson J, Walters MR, Lees KR: Variability in modified Rankin scoring across a large cohort of international observers. *Stroke* 2008;39:2975–2979.