Contemporary Outcome Measures in Acute Stroke Research
Choice of Primary Outcome Measure

Kennedy R. Lees, MD, FESO; Philip M.W. Bath, MD, FESO; Peter D. Schellinger, MD, FESO; Daniel M. Kerr, BSc; Rachael Fulton, MSc; Werner Hacke, MD, FESO; David Matchar, MD; Ruchir Sehra, MD; Danilo Toni, MD, FESO; for the European Stroke Organization Outcomes Working Group

Background and Purpose—The diversity of available outcome measures for acute stroke trials is challenging and implies that the scales may be imperfect. To assist researchers planning trials and to aid interpretation, this article reviews and makes recommendations on the available choices of scales. The aim is to identify an approach that will be universally accepted and that should be included in most acute trials, without seeking to restrict options for special circumstances.

Methods—The article considers outcome measures that have been widely used or are currently advised. It examines desirable properties for outcome measures such as validity, relevance, responsiveness, statistical properties, availability of training, cultural and language issues, resistance to comorbidity, as well as potential weaknesses. Tracking and agreement among outcomes are covered.

Results—Typical ranges of scores for the common scales are described, along with their statistical properties, which in turn influence optimal analytic techniques. The timing of recovery on scores and usual practice in trial design are considered.

Conclusions—The preferred outcome measure for acute trials is the modified Rankin Scale, assessed at 3 months after stroke onset or later. The interview should be conducted by a certified rater and should involve both the patient and any relevant caregiver. Incremental benefits at any level of the modified Rankin Scale may be acceptable. The modified Rankin Scale is imperfect but should be retained in its present form for comparability with existing treatment comparisons. No second measure should be required, but correlations with supporting scales may be used to confirm consistency in direction of effects on other measures. **(*Stroke*. 2012;43:1163-1170.)**

Key Words: acute stroke ■ outcomes ■ interpretation ■ randomized controlled trials

See related articles, p 935 and 1171.

From a recent review of functional outcome measures in published stroke trials, at least 47 options were identified.[1] This wide range presents a challenge to investigators and regulators who are unfamiliar with the field and implies that the available outcome scales may be imperfect. To assist researchers planning trials and to aid interpretation, this article reviews and makes recommendations on the available choices of scales. The aim is to identify an approach that will be universally accepted and that should be included in most acute trials, without seeking to restrict options for special circumstances.

Thus, the article first considers outcome measures that have been widely used or are currently advised. It examines desirable properties for outcome measures such as validity, relevance, responsiveness, statistical properties, availability of training, cultural and language issues, resistance to comorbidity, as well as potential weaknesses. Tracking and agreement among outcomes are covered.

Typical ranges of scores for the common scales are described, because these have a bearing on their use in certain case mixes. It also affects their statistical properties, which in turn influence optimal analytic techniques, a topic reserved for a separate article.[1a] Finally, it is relevant to examine the timing of recovery on scores and usual practice in trial design.

Primary sources of data include recent reviews, analyses conducted specifically for this review based on data from the Virtual International Stroke Trials Archive (VISTA), the deliberations of the National Institute of Neurological Diseases Common Data Elements (NINDS CDE) project, information about contemporary stroke trials registered with clinicaltrials.gov and ISRCTN, recommendations issued by the European Medicines Agency and the Food and Drug Administration of the United States, and examples from the literature.[1–6]

Current Practice

The European Medicines Agency Points to Consider document published in 2001 refers to Barthel Index (BI), modified

Received April 21, 2011; accepted November 7, 2011.
James C. Grotta, MD, was the Guest Editor for this paper.
The online-only Data Supplement is available with this article at http://stroke.ahajournals.org/lookup/suppl/doi:10.1161/STROKEAHA.111.641423/-/DC1.
Correspondence to Kennedy R. Lees, FRCP, FESO, University of Glasgow, Western Infirmary, 44 Church Street, Glasgow, UK G11 6NT. E-mail k.r.lees@clinmed.gla.ac.uk

Rankin Scale (mRS), and Glasgow Outcome scale, and to 4 neurological severity scales: Scandinavian, National Institutes of Health, Canadian, and Unified.[5] It states that BI had been the most widely used functional outcome scale in stroke trials.

Quinn et al[1] undertook a systematic review of functional outcome measures that had been used in stroke trials published over the period 2001 to 2006, identifying 126 trials with a median of 100 patients in each, 47 outcome measures featured, with mRS most prevalent (64.3%) and BI second (40.5%). The National Institutes of Health Stroke Scale (NIHSS) was in third place at 27.8% but was selected as primary outcome more often than BI. One hundred trials used a functional measure as primary outcome, most often mRS. Heterogeneity in choice of measures and their analysis was substantial. Fifteen outcome measures were used across 70 trials of investigational medicinal products. However, only mRS, BI, National Institutes of Health Stroke Scale (NIHSS), Scandinavian Stroke Scale, and Glasgow Outcome Scale (GOS) were each used in >5% of trials. Of these, only the first 3 featured as primary outcome measure in >5% of trials.

Within VISTA,[3] 21 acute trials of ischemic stroke and 6 trials that included hemorrhagic stroke recorded mRS, NIHSS, and BI. All trials completed within the past decade included all 3 measures. Eleven of the ischemia trials included the Scandinavian Stroke Scale, whereas no intracerebral hemorrhage trial did so. Only 5 of 12 trials completed in the past decade included Scandinavian Stroke Scale. The Scandinavian Stroke Scale and NIHSS may be interconverted, however.[7]

After extensive review and consultation, the NINDS Common Data Elements group selected mRS, NIHSS, BI, and EuroQol as most relevant for acute stroke use, with conditional support for the Functional Independence Measure and GOS.[4] For activities of daily living or functional status, 2 measures were recommended as core or potential primary measures: BI and mRS.

A further systematic examination of trials involving interventions for stroke in progress from 2007 to 2010 has been undertaken, based on registrations on clinicaltrials.gov and not restricted to acute stroke (unpublished data). Across 473 trials, at least 191 forms of outcome measure are described and at least 63 unique measures are listed as primary outcome (online-only Supplemental Table I, http://stroke.ahajournals.org). Again, mRS was most prevalent and most often used as primary outcome. The NIHSS was second most prevalent. Barthel remains in third place and was the primary measure in only 8 trials. The Fugl-Meyer scale, which only measures motor function, was in fourth place; however, it is typically only used in rehabilitation trials.[1] Note that European Medicines Agency stated that quality of life was not the primary purpose of stroke treatment and that any quality of life scale used should have been validated for use in stroke, also noting that work toward such validation was desirable.[5] Cognitive, mood, and quality of life scores are uncommonly reported. Cognition, mood, and language function are covered in a third article prepared by the ESO Outcomes Working Group.[7a]

Description of Most Common Outcome Measures

The mRS measures the degree of disability or dependence in daily activities.[8–10] It requires an interview or assessment with the patient or caregivers and can be completed in ≈5 minutes. It is scored on a hierarchical ordinal scale from 0 to 6, with 0 indicating no symptoms and 6 indicating death.

The NIHSS is a 15-item scale to record neurological examination findings in acute stroke. It records deficits affecting level of consciousness, language, neglect, visual field loss, extraocular movement, motor strength, ataxia, dysarthria, and sensory loss.[11] Scores range from 0 (no deficit) to 42 (maximal deficit), but because of the scoring rules when testing is limited in uncooperative patients, the grading of severely affected patients with scores >20 to 25 is likely to be unreliable. NIHSS mainly is used in documenting baseline stroke severity and initial changes in condition. It takes <10 minutes to complete but requires a trained assessor in the presence of the patient.

The BI measures performance in 10 basic activities of daily living and mobility.[12] It is usually scored from 0 to 100, with higher scores indicating increased likelihood of being able to live at home with a degree of independence.[13] It takes ≈10 minutes to assess.

The GOS is a hierarchical ordinal scale for describing disability and handicap in patients with brain injury, scored from 1 (death) to 5 (good recovery).[14] In its extended version (extended GOS, scored 1–8), the last 3 ratings have upper and lower categories.[15] It can be completed by interview in 5 minutes.

The EuroQol-5D is a generic instrument to measure health outcome.[16,17] It is in 2 parts: EuroQol-5D and EuroQol visual analog scale. The EuroQol-5D records a single digit response to 5 questions on mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, respectively. As a result, there are 243 unique coded health states utilizing the 5 categories. These can be compared across various diseases to determine quality of life. The EuroQol visual analog scale is a self-rating of health-related quality of life, recorded from 0 (worst state) to 100 (perfect health).

Associations Among Scales and Typical Outcome Distributions

There are strong correlations among the 3 most widely used outcome measures. In part II of the NINDS alteplase trial, Lu et al[18] found lower correlations than those from VISTA, but these were based on dichotomization, not ordinal use of mRS. Among 9275 patients from VISTA, 3-month mRS had a Spearman rank correlation with BI of −0.94 and with NIHSS score of 0.91. NIHSS correlated with BI ($r_s = -0.85$). From the distribution of the outcomes, it is evident that BI suffers from floor and ceiling effects in this typical stroke trial population, and NIHSS also may have a ceiling effect (Figure 1, online-only Supplemental Figure I). Thus, there are strong correlations among outcome measures, especially if the full range of the scales is examined, whereas dichotomized outcomes may reflect contrasting levels of recovery because of selection of dissimilar cut-points.

The NINDS trial took advantage of apparent variation among scale results when analyzing the early alteplase trials, gaining statistical power from the use of correlated outcomes.[18,19] The advantage of this approach is lower when correlations are close.

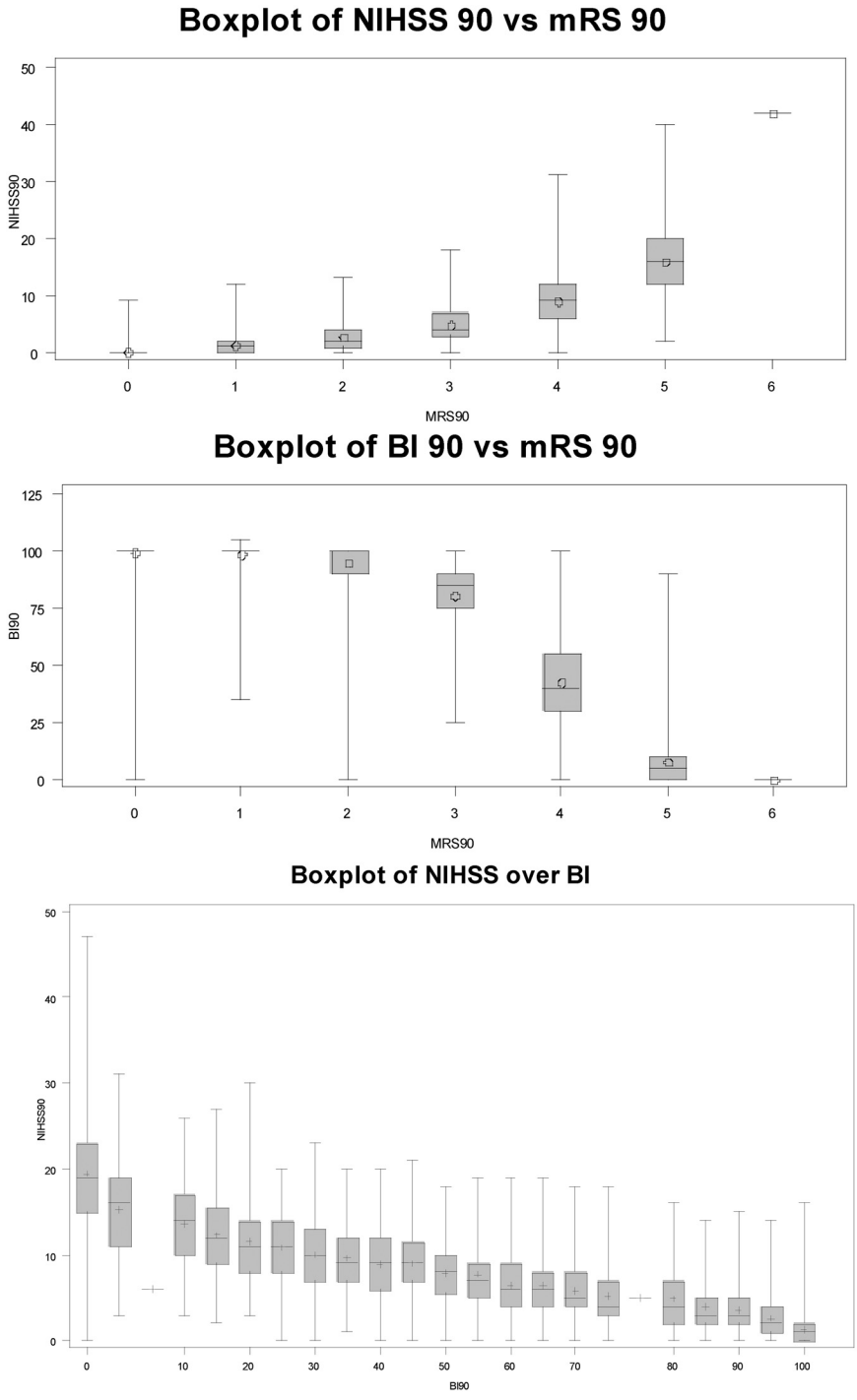

Figure 1. Associations among National Institutes of Health Stroke Scale (NIHSS), Barthel Index (BI), and modified Rankin Scale (mRS). Data from >9000 patients in VISTA.

From VISTA, the mean±SD distribution of mRS scores at 3 months (0–6, respectively) were 10.3%±4.8%, 16.4%±3.5%, 12.6%±2.8%, 15.3%±2.2%, 20.8%±3.5%, 7.5%±2.0%, and 17.0%±5.6% for a population of 14 708 patients across 15 trials with median baseline NIHSS score of 13. These should be contrasted with outcomes for patients with severe stroke (Figure 2).

Value of Second Measure

The European Medicines Agency Points to Consider document suggests that the mRS is chosen as a primary end point.[5] It states that if ordinal analysis rather than dichotomized analysis is used, then a second scale, such as a neurological scale, also should be analyzed, and it declares that dichotomization of the neurological scales would be discouraged (online-only Supplemental Table II). In contrast, the Working Group favored ordinal analysis and did not support the European Medicines Agency guidance that a statistically weaker and clinically less relevant measure using NIHSS should be given equal place with mRS when the most clinically relevant and statistically reliable assessment of mRS is used. The difficulty in predicting the most suitable

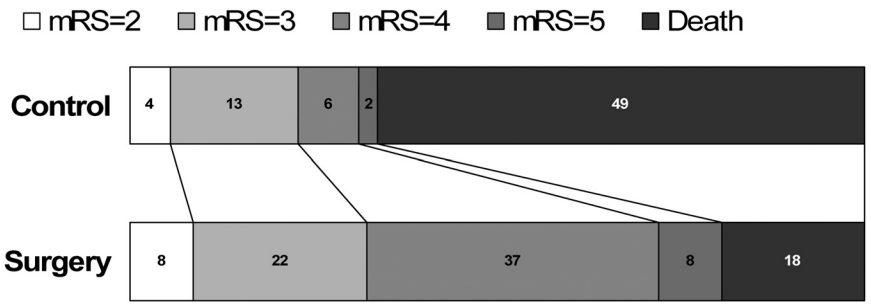

Figure 2. Outcomes assessed by modified Rankin scale after 1 year among patients with severe ischemic stroke, treated with or without decompressive surgery. Based on published data.[20,21]

dichotomy for mRS, dependence on case mix, and failure to incorporate information on potential harm at other levels of the scale each renders this counterintuitive; ordinal analysis is statistically more powerful, is more robust to variation in case mix among trials, and better expresses the importance to clinicians and patients of functional gains at any level of mRS. If a second measure should be used, it would be reasonable to require that it simply tracks with mRS in a predicted manner and shows changes in a similar direction, but it would not be reasonable to require a similar level of statistical significance for a less powerful and less relevant measure.

Desirable Properties

Desirable properties of scales measuring outcome are validity, reliability, responsiveness, and convenient statistical properties. Criterion validity, agreement with a gold standard, is difficult to assess in absence of an accepted standard. Correlation with infarct volume, convergent validity, is moderate for NIHSS and GOS, and slightly lower for BI (Supplemental Table III).[22] Magnetic resonance apparent diffusion coefficient volume significantly correlates with GOS (r=0.73), mRS (r=0.68), and BI (r=0.67; each P<0.001).[23]

The mRS is closely related to hospital bed occupancy and health care costs.[24] The construct validity confirmation that the concept that is measured, for example, a group of motor functions versus various higher cortical functions, is acceptable for NIHSS, BI, and mRS.[25,26] Expert opinion and literature review appear satisfied that these 3 scales also have relevant content, ie, have "content validity."[2]

Reliability, relative freedom from random error, depends both on the homogeneity of the construct that it measures and the reproducibility within and among observers who apply the scale. The NIHSS has a high intraobserver reliability, with intraclass correlation coefficient of 0.93 between ratings 3 months apart.[27] Interobserver reliability is also high, with an overall interclass correlation coefficient of 0.95 (raters underwent formal training and certification of the scale with a standard videotaped program).[25] The BI has high internal consistency (as indicated by Cronbach α of 0.98), implying redundancy of items. Intraobserver and interobserver reliabilities are also quite high, with Pearson r scores ranging from 0.89 to 0.99.[28] In contrast, mRS may have lower internal reliability, because it conflates motor and cognitive elements with environmental and historical elements; resumption of usual activities can depend on previous social interest, recovery of motor function, psychological motivation, and even legal permissions.

Training is available for NIHSS, BI, and mRS, with certification procedures for NIHSS and mRS. A formal scoring system may be used for the mRS, such as the Rankin focused assessment,[29] the Structured Interview for the mRS,[30,31] or a training program[32] to determine the score that best describes the current state of the subject. The GOS reliability is improved by structuring the interview, giving interobserver agreement of 92%.[15]

For trial use, responsiveness, ie the capacity to detect intrasubject changes over time or between treatments, is crucial. The NIHSS is useful for serial monitoring of patients after stroke to detect neurological worsening. A change of 4 points or more is interpreted as clinically meaningful in the multicenter registry of intravenous thrombolysis (SITS-ISTR);[33] some trials use a threshold of 2 points. Although not validated for this purpose, a quantifiable change in NIHSS can be readily recognized and may prompt further diagnostic studies or treatment. In a prospective study comparing 5 outcome measures in 1530 patients 100 days after ischemic stroke, the mRS was more responsive to changes in functional status and better-differentiated changes in mild-to-moderate disability than BI, which suffered a ceiling effect in milder stroke.[34] In contrast, the mRS has more limited sensitivity over short time intervals, especially during hospitalization before patients have attempted their usual roles and activities and because of a substantial clinical threshold between each point in the scale.

Only the mRS and GOS are hierarchical scales; the NIHSS and BI have scores that may be attained in various combinations of subitems that are not necessarily equivalent in their relevance to outcome. The distribution of final outcomes on BI is U-shaped, which renders it insensitive to subtle change between populations and forces dichotomization (which weakens statistical power). The NIHSS has a bimodal skewed distribution of final scores. The GOS and mRS show a more even distribution and have a near-optimal number of categories to offer at least 95% of the discriminatory power compared to a continuous scale. Statistical analysis is discussed in a second article.[1a]

Blinded assessment by telephone interview is reliable with BI[35] and has been used with mRS, although it is not specifically validated. Published values for external reliability of the mRS (ie, the kappa for interobserver agreement) under ideal circumstances with a few very experienced raters range from 0.25 to 0.72, with a mean of 0.46 (95% CI, 0.41–0.51).[36] True reliability with several hundred raters in a

multicenter trial will be at the poorer end of the range. The Spearman Brown prediction formula shows that increasing the number of ratings per patient from 1 to 4 will improve reliability substantially, eg, from 0.25 to 0.57, or from 0.46 to 0.77.[37,38] Recently, central adjudication of video-recorded mRS interviews has been found feasible and valid, with the advantage that multiple raters may provide a score offering blinding, source data verification, cross trial consistency, and improved statistical power in a single package.[39] Involving 4 raters in central review will deliver improved power equivalent to an increase in sample size of at least 5% to 10%, although the sample size gains based on the most conservative estimates of existing reliability may exceed 30%.[37–40]

International use requires transferability across language and cultures. English, Mandarin, Spanish, Thai, and Italian validated translations of NIHSS are available; mRS is also available in English and 11 other languages, with comparability tested across cultures and language (eg, Mandarin and English).[41,42]

Clinically Relevant Shifts

Each of the mRS categories except 5 to 6 represents a clinically meaningful difference in health state.[24,43] The extent of a shift that may be needed for drug or device registration and marketing approval purposes is beyond the scope of this article, because it will depend on cost, safety, resource availability, and local policy. It is evident, however, that any improvement in health state that can be measured on mRS will be evident to patients and caregivers in day-to-day life will be associated with substantial reductions in duration of hospitalization and will translate into health care savings within the first 3 months after stroke.[24,44] Improvements at different levels of the scale are not equal, but because each of the potential improvements is important and has clinical and societal benefit, it should not be mandatory to distinguish these changes from each other when considering overall effect, provided that the benefit is monotonic, ie, that no health state boundary is worsened.[45] In the event of worsening at one level, a balanced decision would be required; however, just as dichotomization could show benefit despite such an adverse effect at another level of the scale, ordinal analysis will normally produce a significantly positive result only if the overall trend is positive despite such adverse consequences. Thus, the ordinal approach protects against false claims of benefit that could be outweighed by harm at other levels.

Timing of End Points

The Food and Drug Administration offers guidance for device trials on timing of recordings, suggesting that 30 and 90 days should be considered, presumably implying that 90 days is the primary end point (online-only Supplemental Table IV).[6] This is consistent with a large number of recent acute stroke trials, including all of the major alteplase, desmoteplase, and prourokinase trials. Within VISTA,[3] 24 trials involving 26 898 patients had outcome assessed at 3 months and 5 trials of 7211 patients extended follow-up to 6 months or longer; and 3 trials of 184 patients completed assessment at 1 month or earlier. On timing of assessments, the NINDS CDE group concluded that acute stroke studies intended to demonstrate durable clinical benefit should assess outcome using a clinically meaningful measure of stroke disability at 90 days.[4] Evaluation of clinical outcomes beyond 90 days was encouraged. The CLEAR-III trial of intraventricular recombinant tissue plasminogen activator for intraventricular hemorrhage includes outcome measured at 6 months, as did the STICH trial of surgical intervention.[46] The hemicraniectomy trials extended follow-up to 12 months.[20,21] Thus, trials that concentrate on patients with the most severe stroke syndromes and involving surgical interventions that could be associated with early morbidity have allowed longer for potential recovery to be realized and for any short-term adverse effects to resolve. Extending outcome beyond 3 months may allow more extraneous events unrelated to the stroke to accrue (eg, myocardial infarction, cancer, trauma) that could attenuate any treatment effect.

Undertaking an outcome assessment at 3 months may be recommended as standard for all trials intending to demonstrate sustained benefit of acute treatment in stroke, except that a later outcome measure is acceptable in circumstances in which stroke severity is substantial, in which the early risk-to-benefit ratio may be relatively unfavorable but is expected to reverse with sustained recovery, or in which early benefits are suspected to favor people with higher competing risks.

Applying mRS in Practice

The mRS has the advantage of being easy and quick to administer and reflects patients' outcomes in practical real-world settings. Although use of the mRS is widespread, it is often administered in different ways, making careful consideration of both the scale in general and its use in practice important. Issues affecting the value of the mRS as the main primary end point have included interviews being conducted in person versus via the telephone; variable or no standard training, particularly in differentiating the critical increments in the middle of the scale (ie, score of 1 versus 2 versus 3) that often become critical when dichotomous end points are used;[2] language or cultural backgrounds;[47] and different methods of determining success such as different dichotomous cut points or ordinal analyses.

A major focus of research on the application of the mRS has been on methods to reduce variability and to improve reliability. Wilson et al[31,48] reported improvement in interrater reliability with the use of a structured interview to standardize the mRS scoring. When such an interview is administered via the telephone, the reliability is substantially reduced and becomes difficult to recommend.[2,49] In a more recent randomized evaluation of the structured interview versus standard mRS scoring, reliability was not nearly as good as in earlier studies.[50]

There have been other methods proposed to improve reliability in mRS scoring, including video recording and central scoring.[51] Initial results of this approach still indicate further work needs to be performed to enhance reliability using this method.[52] A novel method (Rankin focused assessment) still being fully evaluated even suggests utilizing other elements of medical history and NIHSS scoring to help

reduce variability.[29] There is one method to improve mRS scoring that seems to have consensus agreement. This is using a digital-based training program with certification.[32] Training using commercial vendors, such as trainingcampus.com, has become standard for nearly all sponsored multicenter studies. Other ways to reduce potential bias and variability have included the use of blinded scoring of mRS by evaluators who have not been involved in the intervention being studied and using single evaluators for all patients at a particular site. Whether these methods are effective remains to be seen.

Regarding cultural variability, more research is needed to determine sources of variability between countries. A study of scoring after formalized digital training (with "real-life" patients speaking English) revealed there were substantial variations between countries that could not be fully explained by difference in native language.[47]

Finally, determining success for a study outcome with mRS continues to be inconsistent across studies. Some investigators have proposed ideal binary cut-off criteria for a dichotomous outcome.[53] Others have suggested that ordinal outcomes may be most useful in many cases by providing information about improvements that do not reach the dichotomous threshold.[2,54] Some suggest that the analysis method used may be different based on the type of intervention being tested, such as thrombolysis versus neuroprotection.[54,55]

Commentary

There is an evident consensus among both trialists and regulators that functional outcome measures are appropriate for trials intending to demonstrate sustained benefit of acute treatment in stroke. There is evident agreement that the NIHSS has become the neurological scale of choice but that it is not ideal as the primary end point of a trial. It lacks meaning for patients and does not closely reflect social or health care needs, quality of life, or health economics; its statistical properties are poor. The BI is the most widely recommended activities of daily living scale, but it has fallen out of favor as primary end point for acute stroke trials because of its ceiling effect, poor responsiveness, undesirable statistical properties, and reliance on motor function to the exclusion of quality of life and cognitive function. The mRS is widely favored by trialists and regulatory authorities. Although far from perfect, it has favorable clinimetric properties and there is widespread familiarity with it. There is extensive experience with mRS in trials of medical and surgical interventions for acute stroke across a range of severity and in all countries. A 90-day recording of mRS has been available for almost every acute trial conducted in the past decade, and thus comparisons among treatment effects can be undertaken. The mRS scores show an association with quality of life and with economic measures. Each category on mRS reflects a different length of hospital stay and associated short-term health care cost.[24] This is not the case for NIHSS or for BI. Certain categories on BI are also associated with changing bed occupancy and cost, but the relationship is not graded across the entire scale as it is with mRS. Apart from the range 90 to 100, BI scores offer little useful information in this regard. The study by Dawson[24] examined resource use only in terms of bed occupancy over the course of the first 90 days. It remains possible that BI scores would be more informative in predicting longer-term use of other types of support services.

In a relatively small number of patients (n=435), Spieler et al[56] found that by month 12 after discharge, the costs of stroke care amounted to 17 799 euros (16 440–19 158) per patient; the initial hospitalization accounted for 42% of this cost, rehabilitation accounted for 29%, and ambulatory care accounted for 8%. These costs were mostly concentrated within the first 3- to 6-month period. After 46 months without recurrence, the cost of ambulatory care outweighed the cost of the first 6 months. Handicap levels explained 43% of the variance of costs ($P<0.0001$) and, according to the Rankin scale divided into 3 classes (0–2, 3, and 4–5), cumulative costs over time differed considerably.

Improvements on mRS can be demonstrated in response to treatment and show reasonable associations with other outcomes. The mRS has room for improvement that should be considered in operational use of the score. For example, efforts to reduce rater-based effects should be minimized to optimize statistical power.[39] Because of the nature of the score of 6 being fixed and offering no further information, treating 5 and 6 together should be considered. In any case, high mRS scores have a utility for most patients that approximates that of death.[57] It is desirable to include mRS at 90 days in all future trials intending to demonstrate sustained benefit of acute treatment in stroke. For trial success, it should be sufficient to demonstrate that the investigational treatment has produced an improvement in the mRS at 90 days or later, compared to control. The GOS correlates with mRS and has similar properties but has not been as widely studied in acute stroke, as opposed to neurotrauma trials. Good recovery is not clearly defined and does not distinguish symptomatic from asymptomatic patients. GOS offers no advantage over mRS and cannot be used for comparison of treatment effects of other interventions. One approach that may add value to mRS is that of home time.[44] This simple measure records the number of nights that patients spend within their original homes or in a relative's private home in the first 90 days after stroke onset, contrasted with nights in any sort of institutional environment. It correlates with mRS and economic measures, is robust and objective, and, provided that it is stratified or adjusted for country, it appears responsive and useful as an outcome measure (online-only Supplemental Figure II).[58] It has begun to be added as a secondary measure in stroke trials.[59]

Rather than require a second outcome measure,[5] with the associated risks of confusing the true properties of the significance testing (alpha and beta), the relative importance of each measure, and the meaning of any combined end point ("improvement in disability level associated with reduced chance of measurable neurological deficit"), it is recommended that supporting scales are used to confirm that the direction of effects is similar when measured in other ways, and that correlations among scales remain similar to those reported in the literature, or that discrepancies can be explained. However, these are not required to demonstrate statistical significance in their own right.

Conclusions

The preferred outcome measure for acute trials is the mRS assessed at 3 months after stroke onset or later. The interview should be conducted by a certified rater and should involve both the patient and any relevant caregiver. Incremental benefits at any level of the mRS may be acceptable. The mRS is imperfect but should be retained in its present form for comparability with existing treatment comparisons. No second measure should be required, but correlations with supporting scales may be used to confirm consistency in direction of effects on other measures.

Acknowledgments

The working group was sponsored by the European Stroke Organisation (ESO), with administrative assistance supplied by Marisa Kretschi on behalf of the ESO. ESO Outcomes Working Group (*committee member): Hernan Altman, Philip M.W. Bath* (session 2 lead), Martin Bland, Natan Bornstein,* John Boscardin, Stephen M. Davis, Avinoam Dayan,* Geoffrey Donnan, Wolfgang Eisert,* Gary A. Ford, Werner Hacke, George Howard, Markku Kaste,* Michael Krams, Kennedy R. Lees* (working group chair; session 1 lead), Didier Leys, Patrik Lyden, David Matchar, Carlos Molina, John Norrie, Bo Norrving, Frank Rathgeb, Joshua Resnick, Steve Richieri, Jeffrey Saver, Peter D. Schellinger* (session 3 lead), Ruchir Sehra,* Yoram Solberg, Danilo Toni,* Thomas Truelsen, Nils Wahlgren, Andrew Weiss.*

Disclosures

The authors and members of the working group are employed by a range of academic and commercial organizations involved in stroke research, including pharmaceuticals and devices. The manuscript was drafted by K.R.L., with assistance of P.D.S., P.M.W.B. and coauthors, and was approved by all members of the working group. No commercial organization was involved in the content or decision to publish.

References

1. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. *Int J Stroke*. 2009;4:200–205.
1a. Bath PMW, Lees KR, Schellinger PD, Altman H, Bland M, Hogg C, et al. Statistical analysis of the primary outcome in a acute stroke trials. *Stroke*. 2012;43:1171–1178.
2. Kasner SE. Clinical interpretation and use of stroke scales. *Lancet Neurol*. 2006;5:603–612.
3. Ali M, Bath PM, Curram J, Davis SM, Diener HC, Donnan GA, et al. The Virtual International Stroke Trials Archive. *Stroke*. 2007;38:1905–1910.
4. Stroke CDE Working Group. Stroke CDE Standards. Available at: http://www.commondataelements.ninds.nih.gov/Stroke.aspx 2010. Accessed February 28, 2011.
5. The European Agency for Evaluation of Medicinal Products – Evaluation of Medicines for Human Use. Points to consider on clinical investigation of medicinal products for the treatment of acute stroke. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003342.pdf London 20 September 2001 CPMP/EWP/560/98. Accessed April 21, 2011.
6. US Food and Drug Administration. Guidance for Industry and FDA Staff - Pre-Clinical and Clinical Studies for Neurothrombectomy Devices. Available at: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071403.htm June 18, 2007 Accessed April 21, 2011.
7. Gray LJ, Ali M, Lyden PD, Bath PM. Virtual International Stroke Trials Archive Collaboration. Interconversion of the National Institutes of Health stroke scale and Scandinavian stroke scale in acute stroke. *J Stroke Cerebrovasc Dis*. 2009;18:466–468.
7a. Schellinger PD, Bath PMW, Lees KR, Bornstein NM, Uriel E, Eisert W, Leys D for the European Stroke Organization Outcomes Working Group. Assessment of additional endpoints relevant to the benefit of patients after stroke: what, when where, in whom. *International J Stroke*. 2012; in press.
8. Rankin J. Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scott Med J*. 1957;2:200–215.
9. Farrell B, Godwin J, Richards S, Warlow C. The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: final results. *J Neurol Neurosurg Psychiatry*. 1991;54:1044–1054.
10. 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.
11. Lyden P, Brott T, Tilley B, Welch KM, Mascha EJ, Levine S, et al. Improved reliability of the NIH stroke scale using video training. NINDS TPA Stroke Study Group. *Stroke*. 1994;25:2220–2226.
12. Mahoney F, Barthel D. Functional evaluation: the Barthel Index. *Md Med J*. 1965;14:61–65.
13. Granger CV, Dewis LS, Peters NC, Sherwood CC, Barrett JE. Stroke rehabilitation: analysis of repeated Barthel index measures. *Arch Phys Med Rehabil*. 1979;60:14–17.
14. Teasdale GM, Pettigrew LE, Wilson JT, Murray G, Jennett B. Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of the Glasgow Outcome Scale. *J Neurotrauma*. 1998;15:587–597.
15. Wilson JT, Pettigrew LE, Teasdale GM. Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. *J Neurotrauma*. 1998;15:573–585.
16. The EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16:199–208.
17. Xie J, Wu EQ, Zheng ZJ, Croft JB, Greenlund KJ, Mensah GA, et al. Impact of stroke on health-related quality of life in the noninstitutionalized population in the United States. *Stroke*. 2006;37:2567–2572.
18. Lu M, Tilley BC, NINDS t-PA Stroke Trial Study Group. Use of odds ratio or relative risk to measure a treatment effect in clinical trials with multiple correlated binary outcomes: data from the NINDS t-PA stroke trial. *Statist Med*. 2001;20:1891–1901.
19. Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T, et al. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. *Stroke*. 1996;27:2136–2142.
20. Vahedi K, Hofmeijer J, Juettler E, Vicaut E, George B, Algra A, et al. Early decompressive surgery in malignant infarction of the middle cerebral artery: a pooled analysis of three randomised controlled trials. *Lancet Neurol*. 2007;6:215–222.
21. Hofmeijer J, Kappelle LJ, Algra A, Amelink GJ, van Gijn J, van der Worp HB, et al. Surgical decompression for space-occupying cerebral infarction (the Hemicraniectomy After Middle Cerebral Artery infarction with Life-threatening Edema Trial [HAMLET]): a multicentre, open, randomised trial. *Lancet Neurol*. 2009;8:326–333.
22. Saver JL, Johnston KC, Homer D, Wityk R, Koroshetz W, Truskowski LL, et al. Infarct volume as a surrogate or auxiliary outcome measure in ischemic stroke clinical trials. The RANTTAS Investigators. Stroke. 1999;30:293–298.
23. Engelter ST, Provenzale JM, Petrella JR, DeLong DM, Alberts MJ. Infarct volume on apparent diffusion coefficient maps correlates with length of stay and outcome after middle cerebral artery stroke. *Cerebrovascular Diseases*. 2003;15:188–191.
24. Dawson J, Lees JS, Chang TP, Walters MR, Ali M, Davis SM, et al. Association between disability measures and healthcare costs after initial treatment for acute stroke. *Stroke*. 2007;38:1893–1898.
25. Lyden PD, Lu M, Levine SR, Brott TG, Broderick J, NINDS rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. *Stroke*. 2001;32:131–1317.
26. Granger CV, Hamilton BB, Gresham GE. The stroke rehabilitation outcome study–Part I: General description. *Arch Phys Med Rehabil*. 1988;69:506–509.
27. Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. *Arch Neurol*. 1989;46:660–662.
28. Shah S, Vanclay F, Cooper B. Improving the sensitivity of the Barthel Index for stroke rehabilitation. *J Clin Epidemiol*. 1989;42:703–709.
29. Saver JL, Filip B, Hamilton S, Yanes A, Craig S, Cho M, et al. Improving the reliability of stroke disability grading in clinical trials and clinical practice: the Rankin Focused Assessment (RFA). *Stroke*. 2010;41:992–995.
30. Bruno A, Shah N, Lin C, Close B, Hess DC, Davis K, et al. Improving modified Rankin Scale assessment with a simplified questionnaire. *Stroke*. 2010;41:1048–1050.

31. Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, Muir KW, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. *Stroke*. 2002; 33:2243–2246.
32. 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.
33. Wahlgren N, Ahmed N, Dávalos A, Ford GA, Grond M, Hacke W, et al. Thrombolysis with alteplase for acute ischaemic stroke in the Safe Implementation of Thrombolysis in Stroke-Monitoring Study (SITS-MOST): an observational study. *Lancet*. 2007;369:275–282.
34. Weimar C, Kurth T, Kraywinkel K, Wagner M, Busse O, Haberl RL, et al. Assessment of functioning and disability after ischemic stroke. *Stroke*. 2002;33:2053–2059.
35. Shinar D, Gross CR, Bronstein KS, Licata-Gehr EE, Eden DT, Cabrara AR, et al. Reliability of the activities of daily living scale and its use in the telephone interview. *Arch Phys Med Rehabil*. 1987;68:723–728.
36. Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale: a systematic review. *Stroke*. 2009;40:3393–3395.
37. Spearman, Charles C. Correlation calculated from faulty data. *Br J Psychol*. 1910;3:271–295.
38. Brown W. Some experimental results in the correlation of mental abilities. *Br J Psychol*. 1910;3:296–322.
39. McArthur K, Lees KR. Central Adjudication of Modified Rankin Scale Disability Assessments in Acute Stroke Trials. Report to Chief Scientist Office on grant CZB/4/595. January 14, 2011.
40. Quinn TJ, Dawson J, Johnson PCD, McArthur K, Walters MR, Weir CJ, et al. Beneficial effect of improving modified Rankin Scale reliability on sample size for stroke trials. Proceedings of European Stroke Conference in Barcelona. *Cerebrovasc Dis.* 2010;29(Suppl 2):18.
41. McArthur K, Xing H, Quinn TJ, Dawson J, Walters MR, Sun Wei, et al. Reliability and validity of a translated modified Rankin Scale assessment–a pilot study in Mandarin and English. *Stroke*. 2011;42:e247.
42. Pezzella FR, Picconi O, De Luca A, Lyden PD, Fiorelli M. Development of the Italian version of the National Institutes of Health Stroke Scale: It-NIHSS. *Stroke*. 2009;40:2557–2559.
43. Hong KS, Saver JL. Quantifying the value of stroke disability outcomes: WHO global burden of disease project disability weights for each level of the modified Rankin Scale. *Stroke*. 2009;40:3828–3833.
44. Quinn TJ, Dawson J, Lees JS, Chang TP, Walters MR, Lees KR, et al. Time spent at home poststroke "home-time" a meaningful and robust outcome measure for stroke trials. *Stroke*. 2008;39:231–233.
45. Samsa GP, Matchar DB. Have randomized controlled trials of neuroprotective drugs been underpowered? An illustration of three statistical principles. *Stroke*. 2001;32:669–674.
46. Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, et al. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet*. 2005;365:387–397.
47. 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.
48. Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke*. 2005;36:777–781.
49. Newcommon NJ, Green TL, Haley E, Cooke T, Hill MD, Wilson MD, et al. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin scale. *Stroke*. 2003; 34:377–378.
50. Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified Rankin scale. *Stroke*. 2009;40:762–766.
51. Quinn TJ, Dawson J, Walters MR, Lees KR. Reliability of the modified Rankin Scale. *Stroke*. 2007;38:e144.
52. Quinn TJ, Dawson J, Walters MR, Lees KR. Predicting variability in modified Rankin Scale assessment. *Cerebrovascular Dis.* 2009;27(Suppl 6):I-121.
53. Uyttenboogaart M, Stewart RE, Vroomen PC, De Keyser J, Luijckx GJ. Optimizing cutoff scores for the Barthel Index and the modified Rankin scale for defining outcome in acute stroke trials. *Stroke*. 2005;36: 1984–1987.
54. Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. *Neurology*. 2009;72:1310.
55. OAST Collaboration. Calculation of Numbers-Needed-to-Treat (NNT) in parallel group trials assessing ordinal outcomes: Case examples from acute stroke and stroke prevention. *Int J Stroke*. 2011;6:472–479.
56. Spieler JF, Lanoe JL, Amarenco P. Costs of stroke care according to handicap levels and stroke subtypes. *Cerebrovasc Dis*. 2004;17:134–142.
57. Samsa GP, Matchar DB, Goldstein L, Bonito A, Duncan PW, Lipscomb J, et al. Utilities for major stroke: Results from a survey preferences among persons at increased risk for stroke. *Am Heart J*. 1998;136: 703–713.
58. Mishra NK, Shuaib A, Lyden P, Diener HC, Grotta J, Davis S, et al. Home time is extended in ischemic stroke patients who receive thrombolytic therapy: a validation study of home time as an outcome measure. *Stroke*. 2011;42:1046–1050.
59. Rosenberg G, Bornstein N, Diener HC, Gorelick PB, Shuaib A, Lees K, et al. The Membrane-Activated Chelator Stroke Intervention (MACSI) Trial of DP-b99 in acute ischemic stroke: a randomized, double-blind, placebo-controlled, multi-national pivotal phase III study. *Int J Stroke*. 2011;6:362–367.