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...
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,\(^3\) 24 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.\(^7\)

**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.
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...
by guest on July 9, 2017 http://stroke.ahajournals.org/ Downloaded from

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). Magnetic resonance apparent diffusion coefficient volume significantly correlates with GOS ($r=0.73$), mRS ($r=0.68$), and BI ($r=0.67$; each $P<0.001$).

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

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. 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). The BI has high internal consistency (as indicated by Cronbach $\alpha$ 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. 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, the Structured Interview for the mRS, or a training program 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%.

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); 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. 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.

Blinded assessment by telephone interview is reliable with BI 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). True reliability with several hundred raters in a

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.
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. 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 blinded, source data verification, cross-trial consistency, and improved statistical power in a single package. 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%.

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).

Clinically Relevant Shifts
Each of the mRS categories except 5 to 6 represents a clinically meaningful difference in health state. 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. 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. 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). This is consistent with a large number of recent acute stroke trials, including all of the major alteplase, desmoteplase, and prourokinase trials. Within VISTA, 24 trials involving 26,898 patients had outcome assessed at 3 months and 5 trials of 7,211 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. 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. The hemicraniectomy trials extended follow-up to 12 months. 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; language or cultural backgrounds; 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 reported improvement in inter-rater 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. In a more recent randomized evaluation of the structured interview versus standard mRS scoring, reliability was not nearly as good as in earlier studies.

There have been other methods proposed to improve reliability in mRS scoring, including video recording and central scoring. Initial results of this approach still indicate further work needs to be performed to enhance reliability using this method. 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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\textsuperscript{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\textsuperscript{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.\textsuperscript{59} 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.\textsuperscript{59} 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.\textsuperscript{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).\textsuperscript{58} It has begun to be added as a secondary measure in stroke trials.\textsuperscript{59}

Rather than require a second outcome measure,\textsuperscript{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


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


Contemporary Outcome Measures in Acute Stroke Research: Choice of Primary Outcome Measure
Kennedy R. Lees, Philip M.W. Bath, Peter D. Schellinger, Daniel M. Kerr, Rachael Fulton, Werner Hacke, David Matchar, Ruchir Sehra and Danilo Toni

Stroke. 2012;43:1163-1170; originally published online March 15, 2012; doi: 10.1161/STROKEAHA.111.641423
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/43/4/1163

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2012/03/15/STROKEAHA.111.641423.DC1

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/
SUPPLEMENTAL MATERIAL

to

Contemporary outcome measures in acute stroke research: choice of primary outcome measure.

Kennedy R Lees MD FESO, Philip MW 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 Organisation Outcomes Working Group

*ESO Outcomes Working Group (*committee member)
Hernan Altman, Philip MW 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*
**Supplemental Table 1.** Frequency of scales as outcome measure in 473 trials listed on clinicaltrials.gov website that involve a treatment intervention for stroke patients and provide sufficient protocol description, and that had not completed by end December 2006.

<table>
<thead>
<tr>
<th>Outcome measure.</th>
<th>Number of trials using outcome measure.</th>
<th>Number of trials using outcome measure as primary endpoint.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRS</td>
<td>130</td>
<td>59</td>
</tr>
<tr>
<td>NIHSS</td>
<td>106</td>
<td>15</td>
</tr>
<tr>
<td>Barthel Index</td>
<td>85</td>
<td>8</td>
</tr>
<tr>
<td>Fugl Meyer</td>
<td>61</td>
<td>35</td>
</tr>
<tr>
<td>Modified Ashwood Scale</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>6 min walk</td>
<td>31</td>
<td>8</td>
</tr>
<tr>
<td>Stroke Impact Scale (SIS)</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Gait Kinematics</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>Wolf Motor Function Test</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>Action Research Arm Test</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>Walking speed</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>Mini-mental State Examination (MMSE)</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Functional Independence Measure (FIM)</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Motor Activity Log</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>10m Walk test</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Box and Blocks Test</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>VO2 Max</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Functional Ambulation Categories</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Glasgow Outcome Scale</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Activities of daily living</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Hamilton Rating Scale for Depression</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Step Activity Monitor</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Colour Trails 1 and 2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Timed Up and Go</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Jhelsen-Taylor Test (JTT)</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Visual Analogue pain scale</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Motor Assessment Scale</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>SF-36</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Berg Balance Test</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Berg Balance Scale</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Motricity Index</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Boston Naming Test</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Health Related Quality of Life</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Rivermead Mobility Index</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Grip Strength</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Functional Reach</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Clinical Global Impression (CGI) score of functional disability</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>9-hole peg test</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>MRC grading scale</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Beck Depression Inventory</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Line Cancellation Test</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Arm Motor Ability Test (AMAT)</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Geriatric Depression Scale</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Centre for Epidemiological Studies- Depression scale (CES-D)</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Hospital anxiety and depression (HAD) scale</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Picture Naming</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Chedoke-McMaster Stroke Assessment</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Walking endurance</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Speech Accuracy</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Brief Pain Inventory (BPI)</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Western Aphasia Battery</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Functional Communicative Ability</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Instrumental Activities of Daily living</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Montreal cognitive Assessment (MOCA)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Balance Confidence Scale</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Combined NIHSS, mRS, and BI</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CLOX-2 (Copied Clock drawing Test)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Modified Emory Functional Ambulation Profiles (MEFAP)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Alzheimer’s Diseases assessment scale- Cognition (ADAS Cog)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Reintegration to normal living</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Stroke Specific</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Neuropsychiatric Inventory Questionnaire (NPI-Q)</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Four Square Stepping Test</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Canadian Occupational Performance Scale</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Blessed-Roth Memory test</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Health related goal attainment</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
### Dynamic Gait Index
- 2

### Participation Assessment Scale
- 2

### New Word Learned
- 2

### European Stroke Scale
- 2

### Scandinavian Stroke Scale
- 2

### Global Assessment Scale (GAS)
- 2

### Racho los amigos functional test
- 2

### Movement Imagery Questionnaire
- 2

### 3 minute walk
- 2

### Stair climbing
- 2

### Physical activity for the elderly scale
- 2

### NINDS-AIREN
- 2

### Fullerton Advanced Balance (FAB) Scale
- 2

### Step Test
- 2

### Falls Efficacy Scale
- 2

### Physician Global Assessment
- 2

### Wisconsin card sort
- 2

### Stroop Test
- 2

### Repeatable battery for the assessment of neuropsychological status (RBANS)
- 2

### WHOQoL-BREF
- 2

### Disability Assessment Scale
- 2

### NINCDS- CSN Vascular cognitive impairment battery
- 2

### SF-12
- 2

### Montgomery Asberg Depression Rating Scale (MARDS)
- 2

### Gait Symmetry Index
- 2

### Functional outcomes (unspecified)
- 40

### Cognition (unspecified)
- 38

### Quality of Life (unspecified)
- 20

### Neurological Status (unspecified)
- 17

### Clinical Outcomes (unspecified)
- 16

### Depression (unspecified)
- 15

### Strength (unspecified)
- 14

### Balance (unspecified)
- 11

### Fatigue (unspecified)
- 6

### Spatial Neglect (unspecified)
- 3

### Pain (unspecified)
- 2

### Other scales of unspecified type
- 6

### Other Cognitive Scales (each used once only)
- 46

### Other Physical Scales (each used once only)
- 43

### Other pain scales (each used once only)
- 5

### Other Quality of Life Scales (each used once only)
- 5

---

These trials include all forms of stroke interventional research in patients listed on clinicaltrials.gov, and are not restricted to acute stroke. At least 191 forms of outcome measure are described and at least 63 unique measures are listed as primary outcome.
Supplemental Table 2. EMA Points to Consider: suggestions on primary endpoint.5

The primary efficacy variable should be specified a priori, depending on the effect expected from the study drug.

One option may be to measure the proportion of surviving patients who regain functional independence after stroke (survival free of disability or with only minor disability), as estimated by a functional outcome scale or a more global scale of disability or handicap.

The cut-off point for what is considered a favourable outcome (which may include minor disability) has to be defined and justified in the study protocol. This allows a dichotomic analysis, which is easy to perform and to interpret. In this case results from a neurological outcome scale should be supportive as a secondary efficacy variable.

Alternatively it may be shown that an agent effectively moves patients from the severe outcome to the moderate disability group and from moderate disability to the recovered group i.e. that the drug effect applies across all grades of severity of stroke, moving patients to a higher grade of independence in their activities of daily living. Again, for this kind of analysis, clinically relevant shifts need to be defined and justified in the study protocol. In this case a categorical analysis provides more information on the drug effect than a dichotomic analysis.

In that case a second primary efficacy variable should demonstrate an improvement in neurological deficit, as measured by one of the available neurological stroke scales in order to validate a specific effect of the study drug.

In addition to the demonstration of efficacy, a separate analysis of mortality is required as a safety parameter. A new agent is only acceptable for approval if there is no suspicion of a detrimental effect on mortality whatever the benefits on morbidity.
**Supplemental Table 3. Correlations of CT Infarct Volume and 3-Month Functional Outcome Measures (Correlation Coefficients)**
Reproduced with permission from Saver et al\textsuperscript{22}

<table>
<thead>
<tr>
<th>Outcome Measure</th>
<th>All Infarcts (95% CI)</th>
<th>Visible Infarcts Only (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI (n=150)</td>
<td>0.43 (0.28–0.58)</td>
<td>0.46 (0.29–0.64)</td>
</tr>
<tr>
<td>GOS (n=170)</td>
<td>0.53 (0.40–0.65)</td>
<td>0.59 (0.47–0.71)</td>
</tr>
<tr>
<td>NIHSS (n=131)</td>
<td>0.54 (0.41–0.68)</td>
<td>0.56 (0.40–0.71)</td>
</tr>
<tr>
<td>Mortality (n=191)</td>
<td>0.31 (0.18–0.45)</td>
<td>0.32 (0.16–0.49)</td>
</tr>
</tbody>
</table>
Supplemental Table 4. FDA Guidance on endpoints.* The FDA has not issued formal guidance on stroke endpoints except for thrombectomy devices. For thrombectomy devices, the FDA gives the following non-binding advice.6

Neurologic Evaluation
We recommend you obtain the NIHSS and mRS immediately following the procedure, as well as at 24 hours, 7-10 days (or at discharge from hospital), 30 days, and 90 days following the procedure. We recommend a certified examiner or neurologist who is masked to the treatment group perform assessments. We also recommend you obtain the NIHSS score, mRS, Barthel Index and Glasgow Outcome Scale scores, or any other scores used in the primary outcome measure at 30 and 90-days following the procedure. During the study, we recommend you document the severity of any subject experiencing a neurologic deterioration with the NIHSS.

Clinical Endpoints
Your clinical effectiveness endpoint should be outcome assessments at 30 days and 90 days by any appropriate, validated neurologic impairment scale, disability measure, or handicap scale. Examples of appropriate measures include the mRS, NIHSS score, Barthel Index, and Glasgow Outcome Scale. The selection of appropriate clinical endpoints and statistical approaches depends on the device and study design.

Measures of Success
Definition of study success depends on the primary efficacy measures you select. For studies using clinical outcomes using mRS, we recommend you define success as a significantly increased number of subjects having a good (score of 0-2) outcome compared to untreated controls, or equivalent outcome compared to treatment with other efficacious devices or therapies. We recommend you measure safety success in comparison to the control (equivalence or superiority). The primary safety and efficacy endpoints should include an analysis of intent-to-treat subjects, treated subjects, and observed subjects.

*Note that the ESO Working Group did not endorse this advice, particularly on timing of mRS assessments. The conclusions of the working group are contained within the main manuscript.
Supplemental Figure 1. Length of stay during first 90 days after acute stroke according to final Barthel score. Duration of stay was associated with Barthel index (P<0.0001 by ANOVA), but on Bonferroni testing of adjacent categories, significant differences were found only within the range 90 to 95-100. However, if Barthel scores were grouped (0 to 55, 60-90, and 95 to 100), then each group differed significantly from its neighbour (data not shown). Reproduced with permission from Dawson et al 2007.24
**Supplemental Figure 2.** Forest plot of mean 90-day Home time ± 95% CIs versus mRS. N=1717 (of which mRS 0=197; mRS 1= 268; mRS 2=205; mRS 3=214; mRS 4=366; mRS 5=143; mRS 6 (death)=324, P<0.0001 comparing adjacent categories except mRS 4 to 5 (P=0.37) and mRS 5 to 6 (P=0.0003). Reproduced with permission from Quinn et al 2008.44
Supplemental References


http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071403.htm  
June 18, 2007 Accessed 21 April 2011

