A Critical Appraisal of Stroke Evaluation and Rating Scales

Patrick D. Lyden, MD, and Gabriel T. Lau, BS

Background: To judge the efficacy of new, putative stroke therapies, we need a method to measure neurological deficit accurately in groups of patients before and after treatment. No single measurement technique has yet proven to be universally acceptable, but one approach is the use of rating instruments that summarize the neurological deficit found on clinical examination. Currently, stroke assessment scales may be based on the examination of physical deficits, an inventory of activities of daily living, or a global evaluation of functional outcome.

Summary of Review: Scientific methods for authenticating stroke scales are available in the psychometric and statistical literature. We review currently available stroke scales for their validity and reliability and propose investigations needed to refine further the standardized measurement of neurological deficit following stroke.

Conclusions: We suggest that clinical stroke trials include a physical deficit scale and a global rating during the acute phase and that an activities of daily living scale be added at later points in recovery. (Stroke 1991;22:1345–1352)

Stroke has proven to be a difficult disease to study because the clinical manifestations are highly variable, there are many causes, survivors tend to recover, and the extent of the recovery is variable.1-6 Perhaps most importantly, recovery is difficult to measure quantitatively because the traditional neurological examination is uniquely suited to the accurate description of a single patient, but it is ill-suited to group description over time as required for large-scale clinical investigations.1,7-9 Despite these limitations in measuring outcome, the advent of new, potentially effective treatments for stroke has brought an increase in the number of clinical stroke trials now in planning stages. Planners of clinical trials will need to understand some of the issues underlying the design and use of outcome scales. Recently, excellent reviews of clinimetrics and proper clinical stroke trial design have been published.1,7,8 To date, however, there has not been a review of frequently used stroke assessments and their validity and reliability. We briefly review here some of the important issues in the use of stroke scales. We then provide a summary of currently available scales. Finally, we discuss future directions in the effort to design a more useful stroke scale.

Designing and Using Scales

Reviews of clinical assessments have identified practical attributes useful in the development, evaluation, and application of stroke scales.1,8 It is generally agreed that quantitative assessment should be simple and relatively easy to use for the clinician and patient. Limiting the number of variables in the scale contributes to simplicity and utility of an assessment, but may result in a loss of completeness and sensitivity.1,8 Also, scales must be valid and reliable, statistical concepts that have been reviewed elsewhere.1,3 Here we will briefly discuss these concepts in relation to assessing stroke scales.

Validity

A measurement is valid if it accurately describes the underlying phenomenon or disease.7,10,11 For example, the validity of carotid ultrasound can be shown by correlating the estimates of arterial stenosis with measurements from cerebral angiography. However, for psychophysiological variables such as behavior or performance of a test maneuver, there is rarely an independent “gold standard,” and in this setting validity is a more difficult concept to understand and demonstrate.8,10 To examine the validity of a test of behavior, psychometricians frequently discuss three types of validity: criterion, construct, and content.10 Criterion validity is the demonstration of the accuracy of an assessment compared with a particular
standard, the criterion, using correlation coefficients, concordance, or percentage agreement. This concept is particularly useful when an obvious gold standard exists for use as a criterion. The ultrasound/angiogram example mentioned above illustrates criterion assessment using a concurrent measurement. Quantitative measurements of stroke size on brain imaging may also appear to be a gold standard measure of stroke because they are numerical, highly reproducible, and relatively easy to obtain. However, an abnormality seen on computed tomography, magnetic resonance imaging, xenon cerebral blood flow, or positron emission tomography may not necessarily provide a more accurate assessment of stroke severity because some patients with rather small lesions can be severely impaired and vice versa.

Similarly, predictive outcomes such as length of hospitalization or mortality are appealing and useful criteria because they are easily measured and represent important milestones during recovery. However, such outcomes represent only one aspect of stroke severity, and other comorbid conditions unrelated to stroke contribute to morbidity and length of stay.

Construct validity is demonstrated by examining the relations among a newly created test and other tests to show that the new test measures the same "construct." This conceptual approach is most useful when a definite criterion for comparison does not exist, as in the measurement of intelligence or anxiety. Again, correlation coefficients or other regression methods are used to demonstrate construct validity. We hesitate to accept demonstrations of construct validity as evidence that any given stroke outcome measure is "accurate" because it is not yet certain that neurological deficit represents a single construct. Implicit in the design of all stroke scales is the assumption that observable phenomena such as weakness, sensory loss, aphasia, and gaze palsy all supply some information about a single underlying construct. Alternatively, neurological deficit may consist of separate and unrelated constructs such as impairment in motor or higher cortical function. If this is the case, no single stroke scale will be useful for all subjects, and separate, symptom-specific scales will be necessary.

Content validation relies on expert opinions and reviews of the literature. Statistical methods are generally not helpful, and an assertion of content validity may simply state that the items in the scale are reasonable variables that adequately measure stroke. For example, the content validity of the Canadian Neurological Scale was reported through clinical experience and a literature review. The National Institutes of Health (NIH) scale items were selected from previously studied scales and the experience of the investigators, although the report did not specifically identify the process as content "validation." We are concerned that selection of items in this manner may lack the mathematical sophistication needed to avoid relatively low contributing variables and redundant components. That is, a panel of experts convened to select items for a new stroke scale may unwittingly include items that either do not add discriminant value or redundantly measure the same thing as another item(s). Nevertheless, in the development of stroke scales, clinical judgment and experience are probably sufficient to ensure that the content of the scale relates in a valid way to stroke.

Reliability

A test is reliable if the measurement error (variance) is minimal. Reliability includes both reproducibility among observers and consistency among scale items. Reliability is important because the error is increased if there is poor reproducibility over consecutive tests or among separate observers. Interobserver reliability measures the agreement among different people performing the assessment, whereas intraobserver reliability measures the repeatability of the assessment from the same examiner, during repeat assessments. Observer reliability is best assessed with the $\kappa$ statistic, a measure of agreement developed for the study of nonparametric ratings by observers defined as

$$\kappa = \frac{(P_o - P_e)}{(1 - P_e)}$$

where $P_o$=observed proportion of agreement and $P_e$=proportion of agreement expected by chance. Thus, $\kappa$ is preferred over general measures of agreement because it accounts for agreement expected on the basis of chance. The degree of assessment is conventionally interpreted as follows:

- $<0$ poor
- $0$--.20 slight
- $.21$--.40 fair
- $.41$--.60 moderate
- $.61$--.80 substantial
- $.81$--1.00 almost perfect

Weighted $\kappa$ statistics are frequently used for scale items that contain more than two possible responses, in order to more accurately reflect the amount of disagreement among observers. The weights given for disagreements depend on the severity of disagreement between any two choices on the scale item and are chosen by the investigator to suit particular needs.

Internal consistency measures the variation within the assessment, that is, the degree to which the items may measure one construct. It is a function of the number of items in the scale and the mean correlation between them. Measuring internal consistency is similar to repeatedly dividing a test into halves and finding the correlation between both halves. It can be calculated using the KR20 formula developed by Kuder-Richardson for items with only two possible responses, or Cronbach's $\alpha$ coefficient, a more gen-
eral tool for measuring items with more than two possible responses. \textsuperscript{10,19} Cronbach's $\alpha$ is a coefficient of correlation, with values ranging from 0 to 1. Low values may indicate that more than one construct is present in the rating scale; however, compared with other correlations, $\alpha$ values are often high. Alpha coefficients are considered "good" if $>0.8$ and excellent if $>0.9$. \textsuperscript{1} High values indicate good consistency, but high values may also indicate redundancy among the variables measured by virtue of including excess items measuring the same thing. Such redundancy results in an inefficient scale that requires extra time to administer. Redundancy will also obscure other problems, such as poor reproducibility, because correlation coefficients and $\kappa$ will be falsely elevated. \textsuperscript{20}

Approaches to Stroke Assessment

It is crucial for those using stroke assessment scales to understand the nature of the resultant data in order to properly summarize the results of the investigation and to apply appropriate statistical tests. In general, data about a disease may be classified into one of four categories. \textsuperscript{10,21} Observations may be nominal, based on category without respect to rank; ordinal, based on a ranking into one of several ordered states; interval level, based on the difference between two numerical scores; or ratio-level compared to an absolute origin. For example, classifying stroke subtype yields nominal level data whereas measuring motor performance as absent, poor, or normal would yield ordinal, or rank-level, data. A physiological variable based on a defined zero point, such as Fahrenheit temperature, gives interval-level data, and a variable with a true zero, such as height or weight, yields ratio-level data. Powerful parametric tests, such as Student's $t$ or analysis of variance, should not be applied to stroke scale responses that are at best ordinal-level rankings. \textsuperscript{17,18} In general, we believe that no currently used scale yields data that are interval-level data. Although data are recorded numerically, with values ranging from 0 to 1, interval-level data have the property of an absolute origin and the ability to quantify differences between items. Interval-level data are considered to be "good" if $>0.9$. \textsuperscript{1} High values indicate good consistency, but high values may also indicate redundancy among the variables measured by virtue of including excess items measuring the same thing. Such redundancy results in an inefficient scale that requires extra time to administer. Redundancy will also obscure other problems, such as poor reproducibility, because correlation coefficients and $\kappa$ will be falsely elevated. \textsuperscript{20}

Global Outcome Scales

The investigator renders a global ranking of each patient by assigning the patient to one of a limited number of broad classifications. It is important that the discrete categories be arranged hierarchically, so that patients can be compared in a "worse than/better than" fashion. \textsuperscript{9} Global scales are simple and easy to use, but some global scales have the disadvantage of poor reproducibility due to a lack of standard demarcation between categories. \textsuperscript{1} It is also widely believed that global rankings lack sensitivity, that is, the power to detect small but significant changes in deficit. However, sensitivity depends on reproducibility: a scale with more choices, and correspondingly more variability among examiners, will actually have a smaller probability of detecting a meaningful difference between study groups, compared with a short scale with a few, well-demarcated choices. \textsuperscript{1,17,18} This issue can be clarified by considering the extreme case of a binary rating. Judgments such as alive/dead or normal/abnormal tend to show the highest concordance among observers, whereas multi-choice judgments, such as type of aphasia, have the lowest. \textsuperscript{22,23}

The Rankin scale is a 5-point global assessment that categorizes patients based on their ability to perform previous activities and their requirements for assistance. Groups are ranked from I (no significant disability) to V (severe disability, requires constant attention). \textsuperscript{24} A modified form uses a 0–5-point scale, adding a category for patients with no symptoms. \textsuperscript{25} Although the modified scale has been used to assess outcome in several trials, no formal validation study has been reported. \textsuperscript{23,26} The modified scale has moderate interobserver reliability. \textsuperscript{25}

The Glasgow Outcome Scale was developed as a companion to the Glasgow Coma Scale but has received far less attention. \textsuperscript{27,28} The development of the scale has not been published, nor are reliability or validity data available. The scale consists of five well-defined outcomes: death, persistent vegetative state, severe disability, moderate disability, and good recovery. The scale has been used in studies of outcome following head injury and nontraumatic coma but not for stroke. \textsuperscript{27,29}

Physical Deficit Scales

The clinician uses these scales to describe stroke-related deficit based on the neurological examination. Although data are recorded numerically, with items weighted in accordance with severity, it is critical to note again that such rankings are ordinal and not interval-level data. \textsuperscript{8,18} Thus, the total score, a sum of the individual rankings, may be a misleading number. \textsuperscript{9}

The Mathew stroke scale is a neurological evaluation reported in 1972 and originally used in a study on glycerol therapy. \textsuperscript{30} The method by which the scale was developed was not described. It measures mentation, cranial nerve function, motor power, global
disability status, reflexes, and sensation on a 100-point scale, with lower scores reflecting a more severe deficit.\(^\text{30}\) Concurrent construct validity has recently been reported in a 101-patient study that shows high correlation with other scales.\(^\text{31}\) A study of 344 patients found that low scores predicted early demise but not stroke subtype.\(^\text{32}\) Using a slightly modified version of the Mathew scale, considerable interobserver variability was reported in a 12-patient study by Gelmers et al.\(^\text{33}\) Agreement among investigators ranged from “slight” (\(k=0.159\)) for hemianopsia to almost perfect in areas including reflexes and loss of consciousness. The Mathew scale was found to have a low internal consistency (\(\alpha=0.54\)) compared with other measures,\(^\text{31}\) perhaps suggesting the presence of more than one construct underlying the scale. Recently, a modified Mathew scale has been used in studies of nimodipine or hemodilution for acute stroke.\(^\text{34,35}\) In these investigations, a linear transformation was performed on late outcome scores correcting for initial or baseline score. Group measurements could then be made by comparing changes from baseline over a short period of time. This approach may hold great promise in future studies, as it significantly increases the power of the test.\(^\text{34,35}\)

The NIH stroke scale is a 15-item assessment based on three previously used scales.\(^\text{12}\) The scale items use a 3- or 4-point scale, with higher scores indicating greater deficit. The subscale items encompass level of consciousness, vision, extracerebral movements, facial palsy, limb strength, ataxia, sensation, and speech and language. The NIH scale showed good concurrent correlation with the volume of cerebral infarction measured by computed tomography at 7 days (\(r=0.74\)) and a good predictive relation between initial NIH score and 7-day computed tomography scan in a study of 65 patients.\(^\text{12,13}\) Scores also correlated with 3-month outcome, as measured by a previously unvalidated global index of placement, performance, and location.\(^\text{12}\) Moderate to substantial interrater and intrarater agreement (mean \(k=0.69\), 24 patients) was originally reported among neurologists.\(^\text{12}\) A separate 20-patient study found moderate to substantial interrater agreement using a modified NIH scale in nine of 13 items.\(^\text{30}\) A significant advantage of this scale is that it was designed to be used by nonneurologists. Reliability in the hands of such users is generally good.\(^\text{12}\)

The Toronto Stroke Scale was used in a study of steroid therapy in cerebral infarction, but development of the scale was not addressed.\(^\text{37}\) It is a 317-point, 11-category assessment measuring consciousness, paresis, sensory impairment, hemianopsia, aphasia, higher cortical function, mental confusion, forced gaze, incoordination, dysarthria, and dysphagia. Each item is measured from 0 to 3.5 or 4, reflecting amount of deficit. Although no studies of validation or reliability were initially reported, a high correlation of the Toronto Stroke Scale compared with three other scales was subsequently found.\(^\text{31}\) The same study also revealed that the Toronto Stroke Scale had only moderate internal consistency (\(\alpha=0.72\)). Observer reliability has not been formally reported.

The Canadian Neurological Scale was initially reported by Côté in 1986 and was developed based on a review of the literature and clinical experience.\(^\text{38}\) It was designed to be useful in the hands of neurologists and nonneurologists such as house officers and stroke unit nurses.\(^\text{38}\) The eight-item scale measures level of consciousness, orientation, speech, motor function, and facial weakness for a maximum score of 10 points in the normal patient. A separate section measuring face, arm, and motor response is used for patients with comprehension defects. Validation and reliability assessment has been formally addressed and extensively measured by Côté et al.\(^\text{3}\) Content validity was established by collaboration of selected neurologists in choosing items and their weights, as well as from a review of the literature.\(^\text{3,38}\) Criterion validity of the Canadian Neurological Scale was reported concurrently through comparison to neurological examinations and an unvalidated four-item global neurological assessment. Criterion validity was also suggested by an association with morbidity and mortality events measured by the Katz ADL index. Patients with low total scores had higher mortality at 6 months, greater incidence of repeat vascular events, and lower independence.\(^\text{16}\) The same group formally addressed construct validation by claiming differences between the Glasgow Coma Scale and Canadian Neurological Scale as discriminate evidence, and the responsiveness of Canadian Neurological Scale scores to neurological status changes as a convergent demonstration of validity.\(^\text{3,16}\) Moderate to perfect interobserver agreement was demonstrated within items, with high correlation (\(r=0.924\)) between the scores of two raters.\(^\text{16}\) Internal consistency ranged from \(\alpha=0.896\) to 0.792 in two reports.\(^\text{16,38}\)

The Hemispheric Stroke Scale is a 100-point neurological assessment designed for a trial of hemodilution for the treatment of acute ischemic stroke.\(^\text{39}\) It was designed using empirical methods, choosing items based on their practicality, variability, and reliability. It measures level of consciousness, language, cortical function, motor function, and sensory capability on varying scales, with higher scores reflecting more deficit. The scale notably includes the Glasgow Coma Scale as a measure of level of consciousness. Concurrent validity was initially assessed in a 16-patient study that showed good agreement between the Hemispheric Stroke Scale and an unvalidated six-item global assessment (Pearson \(r=0.89\)) and showed reasonable agreement (\(r=0.95\)) with the Barthel Index in two studies and with the Toronto Stroke Scale (\(r=0.91\)) and the Mathew scale (\(r=-0.92\)) in a 101-point study.\(^\text{31}\)\(^\text{39}\) Interobserver reliability between two raters was originally reported as good (\(r=0.95\)), and the assessment has been shown to have high internal consistency (\(\alpha=0.88\)).\(^\text{31,39}\) However, no measure of intrarater reliability has been
reported, and the scale suffers from the redundant inclusion of several closely related items.

Several motor assessment scales have been designed for single studies, and some of these studies have yielded useful insights. The concept of sequential motor recovery was shown by graphing the stages of motor return for the arm and leg as defined on the Fugl-Meyer Assessment Scale. A modification of this scale assesses gross mobility and gait. During these validation studies, factor analysis was used to show construct validity. The report does not include actual factor loadings, but from inspection of the limited data available, it is clear that only one factor could legitimately be extracted from the data, despite the fact that the scale is very long and relatively complex. This suggests a high degree of redundancy and oversampling in this scale.

In 1984, Allen devised an outcome assessment by recording several features of 128 stroke patients and using multivariate analysis to find features that predicted functional dependence. Old age, complete limb paralysis, depression of consciousness at onset, and combination hemiplegia/hemianopia with higher cerebral dysfunction were found to predict functional dependence whereas uncomplicated hemiplegia was a predictor of independence. Regression weights were derived for these variables and a predictive equation developed. Using the equation, 89% of 156 stroke patients at 6 months were correctly allocated to one of two outcome groups. The study suggests a unique and potentially useful approach in designing future scales through the use of multivariate regression. If a truly appropriate outcome variable can be developed, the coefficients derived from the multivariate regression equation would be very useful in assigning weights to individual scale items.

Activities of Daily Living Scales

ADL scales measure performance in occupational functions useful for independent living. Scales may include basic functions, such as continence, to more complex tasks, such as preparing food. The ADL scales may be administered through patient interview, observation, or self-assessment. Items may be graded on a scale or on a pass/fail basis. Although ADL measurements are useful in studies of rehabilitation, their use may be somewhat impractical in the setting of acute stroke.

The Barthel Index is a widely used ADL scale that comprises 10 weighted items measuring feeding, bathing, grooming, dressing, bowel control, bladder control, toileting, chair/bed transfer, ambulation, and stair climbing. A maximum score of 100 indicates competence in all 10 items. It is among the most thoroughly studied of scales, and numerous groups have reported concurrent validation. The Barthel Index has also been shown to predict length of hospital stay and discharge placement, and higher scores are associated with independent community living, more person–person contacts, and more activity in community affairs. The Barthel Index has substantial interrater reliability (κ=0.80). Activities of daily living reported by telephone interview have been shown to correlate highly with ADL measured from direct examination. Excellent internal consistency has also been reported (α=0.96).

The Kenny Self-Care Evaluation measures six items, including bed, transfer, locomotion, dressing, personal hygiene, and feeding activities. Total scores range from 0 (totally dependent) to 24 (independent in all six categories). In a study of 149 stroke survivors, a high degree of agreement between the Kenny Self-Care Evaluation and the Barthel Index and Katz Index of ADL was found. No formal studies of reliability have been noted.

The Katz Index of ADL was developed to evaluate chronically ill patients, including those with stroke, hip fractures, and multiple sclerosis. The Katz scale summarizes performance in six areas: bathing, dressing, toileting, transfer, continence, and feeding. Patients are defined as either dependent or independent in each area and are ranked into groups based on dependency. Katz ADL originally contained eight rankings (A–G, other), but has since been simplified to a 0–6 scale (6=dependent in all six areas). Katz ADL and the Barthel Index are highly correlated (κ=0.77), and Katz ADL predicts independence as well as the Barthel Index and Kenny Self-Care Evaluation. The Katz index has a strong relation to nursing time required by patients. Observer reliability for the Katz index is not available; however, excellent internal consistency has been reported.

The Activity Index was developed, using components of other scales, for measuring the functional capacity of acute stroke patients. The index, which includes four mental capacity items, six measures of motor function, and five measures of ADL, showed good correlation with the Katz Index of ADL and the Barthel Index. Admission Activity Index scores also showed good concurrent association with the Rankin scale (r=−0.94). The same group reported predictive validity by using linear regression to show scores 3 weeks after stroke and at discharge to be related to admission scores. Similar to the development of the Fugl-Meyer Assessment, factor analysis originally showed that the motor activity components alone would be sufficient to constitute the index, yet other items remain in the scale. No observer reliability was originally reported, but two studies showed the Activity Index to have high internal consistency (α=0.94–0.97), perhaps reflecting excessive redundancy.

Discussion

The development of new stroke therapy is hindered by the lack of an ideal measure of recovery that can be used to compare treated and control groups of patients. Although several stroke scales exist, several problems (notably, the lack of explicit testing of validity and reliability) restrict the widespread use of any one of them.

The concept of validation of stroke assessments has been explicitly addressed in the stroke literature.
is also a useful scientific method for identifying the redundant items in a scale is factor analysis. It should be viewed with an appreciation that redundancy falsely elevates reliability estimates.

Finding suggests that a very brief, yet accurate, inventory could be designed around a motor rating. Alternatively, if there is no single neurological deficit construct that can be measured, a new scale must be defined that rates patients on several dimensions, such as motor, language, vision, and so on. Such a scale would be difficult to use, and clinical trials would probably need to enroll greater numbers of subjects.

In contrast to validity, reliability seems easy to describe and measure in the context of stroke-scale design. Most of the assessments reviewed here have high internal consistency and good correlation with other scales, thus demonstrating good coherence both within the scale and with other scales. However, high correlation among scale items also indicates redundancy and a lack of efficiency. Redundancy inflates reliability measures such as Cronbach’s α and κ statistics due to repetitive ratings of the same construct. Such scales are not efficient because they require the clinician to collect more data than the minimum required to assess a patient’s condition. Therefore, in selecting a stroke scale for a clinical trial, reliability measures such as κ and α should be viewed with an appreciation that redundancy falsely elevates reliability estimates.

One method to reduce the number of superfluous and redundant items in a scale is factor analysis. It is also a useful scientific method for identifying the constructs underlying a scale. Factor analysis was used to determine the number of independent dimensions underlying the modified Fugl-Meyer Motor Assessment and the Hamrin Activity Index. The latter authors concluded that there were three factors measured by their scale and named these factors arm function, leg function, and standing leg movements. However, the limited data presented in the paper justify extracting only one factor from the data. This example illustrates an important use and limitation of factor analysis. Simplistically put, the analysis seeks to explain the variables used in the scale by linking similar variables into a smaller number of factors. A 30-question scale might provide ratings on three factors, for example. The advantage is that it is easier to describe the results in terms of three factors than to summarize the responses to all 30 questions. We suggest that a principal components factor analysis, when applied to a suitable data set, will be extremely useful in selecting items for inclusion in a stroke rating scale. We should note, however, that the correct application of factor analysis requires a skilled statistician working closely with knowledgeable clinicians. The method is limited by the fact that it is not always clear how many factors may be justifiably extracted from the data set. It is possible to over- or under-extract factors, depending on how the analysis is designed. Another limitation is the need for a very large data set, on the order of 10 subjects for every variable in the scale. Finally, factor analysis cannot be used to determine the weights to be applied to the items composing the scale. Depending on the ultimate use of the scale, such weights may not be necessary, but if weights are to be used, a useful approach to their construction is multivariate regression.

Although much basic research on stroke scales remains to be done, many new putative stroke treatments must be tested now. Pending further development of stroke deficit measurements, we suggest the following approach. The clinical investigator must first clarify the needs and purpose of the study. A study of rehabilitative therapy will likely require an ADL or a global outcome assessment. On the other hand, a trial of acute medical treatment for stroke will require a more complex approach. In the acute phase, ADL scales may be difficult to use with precision. Therefore, at early time points a deficit score and global assessment will be useful. Of currently available deficit scales, the NIH and Canadian Neurological Scale instruments appear to be the most rigorously studied at this time. Reliability and limited validity data exist, and variability measures needed for sample size estimations are available. The Rankin is the most widely studied global assessment, and there are limited data available for planning its use. At later time points, 3 or 6 months after acute treatment, a deficit and global outcome measure might be augmented with an ADL scale. The ADL adds information about the patient’s functional abilities and describes the patient in terms of...
his command of his environment.6,51 The Barthel Index is one of the most studied measures of stroke outcome, and detailed validity and reliability data are available.7,53-55,64 Because it has been shown that several ADL scales are highly predictive of each other, the Barthel is as acceptable as any other.20,47,53

Future research must focus on at least two areas. First, how many items are needed to adequately measure stroke deficit in groups of patients? It must be accepted that a complete description of the neurologist's status of each individual is neither possible nor necessary.3 Rather, the clinical investigator needs to know only that a treated group of patients is different from, or the same as, a control group. Thus, selection of test items should be driven by concerns of reliability, validity, and efficiency. Only highly reproducible items that are clearly related to neurologist deficit should be included. Items that lengthen the scale but add little discriminant value should be avoided as being inefficient.65 Once items are selected, comparison to several external criteria of severity, using multivariate regression, will help clarify the accuracy of the new scale.

Second, how should progress be measured? In previous stroke treatment trials, the usual approach has been to compare gross outcome scores among treat-5

ment groups.9 However, this approach lacks statistical power, as the data should be analyzed using nonparametric tests.1,7,11 Some investigators have found interesting changes in outcome score over time.4,6,2,64,65 It seems rational to suspect that an analysis over a short interval early in recovery may accurately measure stroke-related cerebral damage and might even predict eventual outcome. Another promising idea is to correct outcome scores for differences at baseline. Further studies must clarify the impact of such data transformations on power and biasing, but it seems logical to expect that it will be more useful to study relative changes than raw scores. We are certain that substantial progress in this area awaits closer collaboration among statisticians and those clinical investigators charged with designing and managing current acute stroke treatment studies.

Acknowledgments

The authors are grateful to Justin Zivin, MD, PhD, and Barbara Tilley, PhD, for helpful suggestions; to John Rothrock, MD, for reviewing the manuscript; and to Ann Stanaway for editorial assistance.

References

2. Vancley F: Functional outcome measures in stroke rehabilita-
14. Kruppman DS, Rubens AB: The validity of computed tomo-

graphic scan findings for the localization of cerebral functions: The relationship between computed tomography and hemi-

paralysis. Arch Neurol 1986;43:328-331

21. Munsat TL: Quantification of Neurologic Deficit. Boston, But-
terworth, 1989
25. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJA, van Gijn J: Interobserver agreement for the assessment of handi-
A critical appraisal of stroke evaluation and rating scales.
P D Lyden and G T Lau

Stroke. 1991;22:1345-1352
doi: 10.1161/01.STR.22.11.1345

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/22/11/1345