Interconversion of Stroke Scales
Implications for Therapeutic Trials

Keith W. Muir, MRCP; Donald G. Grosset, MD; Kennedy R. Lees, MD

Background and Purpose Stroke scales are intended to measure stroke severity for the purpose of clinical trials. Scores have been used to determine trial entry, to compare patient groups within or between trials, or as a secondary end point. The use of scores as an end point in meta-analysis has not been validated, but such analyses have nevertheless been performed when equivocal results have been obtained using the main outcome measure. The different scale designs suggest that conversion of scores may not be possible. We sought to determine whether scores on different scales could be interconverted.

Methods A single observer scored 433 consecutive admissions to an acute stroke unit on the Canadian Neurological Scale, the middle cerebral artery Neurological Score (or Orgogozo scale), and the National Institutes of Health stroke scale. Data were separated into training and test sets, and linear regression was used to model conversion between scales. Prediction errors were calculated. Strokes were subdivided according to the Oxfordshire Community Stroke Project classification, and coefficients of determination were calculated for different subtypes.

Results Conversion between Canadian and middle cerebral artery Neurological scales was satisfactory ($R^2=94.7\%$), and prediction errors were acceptable (absolute prediction error, $5.0\pm5$). Conversion from the National Institutes of Health scale was worse ($R^2=87.5\%$ to Canadian and 89.0\% to Neurological Score), and prediction errors were significantly greater (Neurological Score error, $8.7\pm7$; Canadian Neurological Scale error, $8.5\pm7.3$; $P<.005$ for both). Coefficients of determination for interconversion were significantly worse for dysphasic patients with total anterior circulation strokes than for other stroke types ($P<.01$). Reweighting the motor component of the National Institutes of Health scale improved coefficients of determination and reduced prediction errors, but prediction error for conversion to the Canadian scale remained significantly greater than other conversions ($P=.001$).

Conclusions The Canadian Neurological Scale and the middle cerebral artery Neurological Score may reliably be converted. The National Institutes of Health scale cannot be used to predict these scores reliably, even with reweighting of the motor score. Interconversion is poorest for patients with dysphasia and total anterior circulation strokes. These results suggest that there will be more general difficulty in interconverting scales that use different test items and weighting. Meta-analysis using sequential changes in averaged scores from various stroke scales is not valid. (Stroke. 1994;25:1366-1370.)

Keywords • cerebrovascular disorders • clinical trials • stroke assessment

Clinical trials in acute stroke require a means of measuring functional outcome, since viable therapies must not only decrease mortality but also lessen the number of survivors who are disabled. Lacking objective means of such assessment, investigators have developed scales that seek to quantify aspects of dysfunction within the World Health Organization's definitions of impairment, disability, and handicap.1 The concepts and methodology of scaling have been thoroughly reviewed by other authors.2 Stroke scales are usually ordinal scales that attempt to quantify impairment and are used in clinical trials either to provide a shorthand method of describing the baseline characteristics of the patient group or to define trial entry criteria.

Most trials use the Barthel activities-of-daily-living index,2 a scale that measures disability and has been extensively validated, as the primary end point. However, where equivocal or conflicting results for the primary end point have been obtained, as with nimodipine and hemodilution, stroke scale scores have been used as surrogate end points and subjected to meta-analysis.4,5 Two approaches have been made to such analysis. The meta-analysis of nimodipine trials of Orgogozo et al6 combined scales after converting each score to a percentage of potential recovery on that scale in an attempt to normalize results. After calculating a combined percentage of potential improvement, the change in this quantity over time was analyzed. Asplund's meta-analysis of hemodilution trials simply calculated a mean score from several different scales and analyzed the change in this mean. Use of scales as primary outcome measures was not intended in their design, particularly since the majority of scales concentrate purely on physical impairments that may be less important for longer-term functional recovery. Also, mathematical conversion of different scales does not take into account the different designs of scales, which in practice mean that not all scales are truly ordinal. The precise impairment that different scales measure therefore differs, and so mathematical conversion may be unreliable.

There are well-argued objections to such use of scales as trial end points.6 These indicate that simple outcome measures (eg, dead or alive, disabled or able to conduct normal activities) are more reliable between observers and more relevant to patient and physician. Furthermore, many trials convert results back into broad patient
groups, which defeats the purpose of using complex (and less reliable) scoring systems. The relevance of demonstrating change in impairment score rather than change in functional status or disability is uncertain.

Despite these objections, the number of scales has proliferated in line with the number of trials performed, and a multiplicity now exists, each with its own proponents. There seems to be little prospect of agreement on a single scale as the standard for stroke trials in the near future. A main driving force behind the use of scales in trials is the difficulty in recruiting adequate numbers of patients. Many trials, particularly those involving new pharmaceutical agents, are primarily safety and tolerability studies and appropriately include only small numbers of patients. Surrogate end points are included to broaden the clinical outcome data and permit a range of exploratory analyses that may guide larger efficacy trials.

An alternative to stroke scales is expounded by the International Stroke Trial, which plans to recruit approximately 20 000 patients. Stroke scales are not used; however, information on neurological features is recorded at entry, and this is sufficient to classify patients according to the Oxfordshire Community Stroke Project (OCSP) classification. This classification has been validated both against radiological data and for interrater reliability. It divides patients into four clinical categories among which there are differences in medium and long-term outcome: total anterior circulation strokes (hemianopia, hemiplegia, and higher cortical dysfunction), partial anterior circulation strokes (any two of these three or isolated, new higher cortical dysfunction), posterior circulation strokes (isolated hemianopia or brain stem signs), and lacunar strokes (pure motor or sensory strokes, sensorimotor stroke, clumsy hand dystarthis, or ataxic hemiparesis). The major difference in outcome is between the total anterior circulation strokes, equivalent to major middle cerebral territory infarction, and all others. Although this classification is valid for broad groups of patients, it is a poor index of functional impairment or outcome in individuals. Worldwide, the classification is not in widespread use.

To determine whether the prediction of different groups was valid, we examined the predictive errors in converting between different stroke scales. The three scales studied were the Canadian Neurological Scale (CNS), the middle cerebral artery Neurological Score (NSc), and the National Institutes of Health (NIH) stroke scale (in its 31-point version). These scales were selected principally because of their use in major ongoing therapeutic trials, but also because this selection allows comparison of two scales that are similar in design (CNS and NSc) with a scale that is different in both in weighting and original conception (NIH). Additionally, the OCSP classification was used as a means of clinically subdividing patients. We did not seek to assess the reliability of scales in predicting outcome.

Methods

A single observer (K.W.M.) scored 433 consecutive admissions to an acute stroke unit simultaneously on each of the three scales (NIH, CNS, and NSc) as soon as feasible after stroke onset. Only the total scores were used in analysis. Complete data were recorded for 410 cases. Other clinical information was recorded to enable classification of stroke types according to the OCSP classification. Final OCSP classification was determined by review of clinical data by all the investigators.

The scores were plotted against each other to examine the possible relations. Linear regression was selected as the most appropriate method of analysis.

Scales differ in their range: 15 to 115 for CNS, 0 to 100 for NSc, and 0 to 31 for NIH. The direction of the scales also differs, a normal individual scoring 0 on NIH but 115 on CNS and 100 on NSc. To facilitate graphic analysis, scores were converted into a uniform 0-to-100 range, with 0 representing a severe stroke and 100 a normal score. NSc scores were unchanged. CNS scores had 15 subtracted from the total. The NIH score was multiplied by 100 and divided by 31 before being subtracted from 100.

The data were randomly separated into training (n=246) and test (n=187) sets of data. A mathematical model for conversion between each pair of scales was generated by simple and multiple linear regression using MINITAB statistical software (Minitab Inc). Multiple linear regression models included one other score, age, sex, side of stroke, and OCSP classification. The models obtained from this were used to calculate predicted scores for the test set of data, and prediction error (observed-predicted score). Absolute prediction error, percentage prediction error (observed-predicted values divided by observed values), and percentage absolute prediction error calculated for each pairwise comparison. Prediction errors were analyzed by one-sample t test (two-sided) to determine difference from 0, and comparison between groups was performed by two-way ANOVA with Bonferroni pairwise comparisons using RUMMAGE software (Brigham Young University, Provo, Utah) on an IBM 3080 mainframe computer.

Further exploration of subgroups was carried out to determine whether the prediction was more difficult for clinically distinct groups of patients. The entire data set was used. Hemorrhagic strokes and transient ischemic attacks were removed from the overall group, and regression analysis before and after removal was performed to determine whether there was a significant change in errors in scoring. Finally, the data set was divided according to the OCSP classification, and regression analysis was repeated for each paired conversion within groups. Regression equations were compared by OCSP stroke class using the F-ratio test.

The NIH scale was arbitrarily reweighted to make the motor score predominant (face weakness multiplied by 2, arm and leg weakness multiplied by 5) and to give a total potential score of 58. Complete data were available for 391 patients. Regression analysis was repeated, and prediction errors were recalculated.

Results

The patients had a mean age of 67 years (range, 32 to 96 years). Equal numbers of men and women and of right and left hemipareses were present. When classified by the OCSP classification, there were 21% total anterior circulation strokes, 45% partial anterior circulation strokes, 13% posterior circulation strokes, and 17% lacunar strokes, and 4% unclassifiable.

The models obtained from the training set are given in Table 1, along with coefficients of determination for each pairing (coefficients of determination are identical for each variable in the regression). The prediction intervals calculated for each model are shown in Fig 1. Prediction intervals were notably wide for conversion from the NIH scale to either the CNS or NSc, with a tendency to predict large negative values for lower scores (ie, those with more severe strokes). Prediction intervals for conversion between NSc and CNS were narrow.
TABLE 1. Models for Conversion of Stroke Scales and Coefficients of Determination for Paired Comparisons

<table>
<thead>
<tr>
<th>Scales Predicted</th>
<th>Observed</th>
<th>Model</th>
<th>$R^2$, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS</td>
<td>NSc</td>
<td>CNS = -5.07 + 1.05 NSc</td>
<td>94.7</td>
</tr>
<tr>
<td>CNS</td>
<td>NIH</td>
<td>CNS = -38.8 + 1.36 NIH</td>
<td>87.5</td>
</tr>
<tr>
<td>NSc</td>
<td>CNS</td>
<td>NSc = 7.48 + 0.901 CNS</td>
<td>94.7</td>
</tr>
<tr>
<td>NSc</td>
<td>NIH</td>
<td>NIH = 28.8 + 0.703 NSc</td>
<td>89.0</td>
</tr>
<tr>
<td>NIH</td>
<td>CNS</td>
<td>NIH = 33.5 + 0.646 CNS</td>
<td>87.5</td>
</tr>
<tr>
<td>NIH</td>
<td>NSc</td>
<td>NIH = 7.48 + 0.901 CNS</td>
<td>89.0</td>
</tr>
</tbody>
</table>

CNS indicates Canadian Neurological Scale; NSc, Neurological Score; and NIH, National Institutes of Health stroke scale.

Note that since only one independent variable is used, coefficients of determination of Y on X and X on Y are identical.

Absolute prediction errors were significantly greater for conversion from the NIH scale to either CNS or NSc. Percentage absolute prediction error was 30±50% (mean±SD) for predicting NSc from NIH and 29±58% for CNS from NIH. Thus, for an individual with a known NIH score, the variation in CNS or NSc predicted using the models obtained here will be substantial. This contrasts with percentage absolute prediction error of 16±22% for conversion from CNS to NSc. As is evident from Fig 2, the errors for predicting the NIH score from CNS or NSc are greater than between CNS and NSc but significantly less than prediction errors for the opposite direction of conversion.

Elimination of all those cases subsequently found to have hemorrhage on computed tomographic scanning (n=49) or transient ischemic attacks (n=58) did not alter the coefficients of determination ($R^2$). Multiple linear regression taking into account age, sex, stroke side, and OCSP classification (all proposed as potential confounding variables) did not improve $R^2$.

Coefficients of determination were influenced by subdivision according to OCSP classification (Table 2). The population that accounted for these differences comprised patients with total anterior circulation strokes and dysphasia (NSc-CNS, $P=0.01$; CNS-NIH, $P=0.01$; NSc-NIH, $P=0.002$); regression lines did not differ significantly for nondysphasic total anterior circulation strokes compared with the population as a whole (NSc-CNS, $P=0.96$; CNS-NIH, $P=0.06$; NSc-NIH, $P=0.18$). Reweighting the NIH scale improved $R^2$ values for interconversion (CNS-NSc, $R^2=95.5%$; CNS-NIH, $R^2=93.1%$; NSc-NIH, $R^2=94.7%$) and reduced the absolute prediction errors. However, while prediction errors were similar for NSc-NIH conversions (NSc from NIH, $5.5±5.1$) and CNS-NSc conversions (NSc from CNS, $5.3±4.4$), the error for conversion from NIH to CNS remained significantly greater (absolute prediction error, 6.7±6.4; $P=0.01$).

### Discussion

We have shown that there are unacceptably large errors in the prediction of either the NSc or the CNS score from the NIH scale. Although it was possible to improve interconversion error by reweighting the NIH scale to make it a predominantly motor score, significant differences in prediction errors persisted. The problems with mathematical interconversion are not surprising, since the scales compared were designed from different principles. The CNS was designed to be used in any stroke type and to be administered by medical or nursing staff. It has been validated for intrarater reliability and for relevance to functional outcome. The middle cerebral artery NSc of Orgogozo and Dartigues was designed for assessment of middle cerebral artery infarction only and has been less extensively validated than the CNS. These two scales are similarly...
weighted, with the greater part of the scale assessing motor impairment. The NIH stroke scale was designed around a traditional neurological examination and is effectively confined to use by medical staff. It was validated against the calculated volume of cerebral infarction on computed tomographic scan rather than clinical outcome. Weighting differs greatly from the CNS or NSc, such that motor assessment constitutes only a minor part of the scale. The NIH scale is in fact not a true ordinal scale of impairment, since some items are mutually exclusive and nonassessable items must score zero. In practice, this leads to a ceiling on the scores that is well below the theoretical maximum (eg, a comatose, hemiplegic patient who is mute and has forced eye deviation scores of only 21/31 because other items are untestable). This "ceiling effect" may account for the finding that conversion from CNS or NSc to NIH has a much smaller prediction error than conversion in the opposite direction. Since in practice the majority of NIH scores are in the range of 0 to 23, NIH scores predicted from CNS or NSc, where the scores cover the full range of theoretical values, will tend to be in the middle range of the NIH scale. Conversely, the regression for conversion from NIH to NSc or CNS will attempt to explain extreme scores on NSc or CNS from a scale that inherently lacks the full range of theoretical values. The prediction errors will therefore be smaller on converting to the NIH than converting from the NIH. Conversion between the middle cerebral scores and the NIH scale may be possible with acceptable error in one direction only, but the significantly greater errors in conversion in the opposite direction suggest that there are intrinsic differences between these scales that prevent reliable clinical use.

The assessments of ataxia, sensory loss, inattention/neglect, and visual fields required by the scale necessitate a considerable degree of comprehension on the part of the patient and are often not assessable. The NIH scale deliberately includes these items to permit assessment of posterior circulation stroke, often cited as its main advantage over other scales. We have demonstrated that coefficients of determination are poorest not for the posterior circulation strokes, however, but for large middle cerebral artery stroke in which there is a comprehension deficit. Coefficients of determination for conversion to CNS or NSc from NIH are as good for posterior circulation strokes as for lacunar or branch middle cerebral strokes.

This carries important implications for clinical trials. Trialists have sought to include only patients with a significant measurable deficit but with a high likelihood of survival, based on the premise that the potential to demonstrate improvement is greatest in this "middle" group of stroke patients. Many trials therefore include only conscious patients with middle cerebral artery territory deficits. Thus, there is a bias toward the inclusion in trials of those patients for whom we have demonstrated the least reliable interconversion of different scales.

The pitfalls in converting stroke scales of different designs that we have shown would ideally be irrelevant in the assessment of treatment efficacy, since scales should not be used as a trial end point. However, as discussed above, multiple different scales have been subjected to meta-analysis using only mathematical manipulation of total scores in instances where multiple small trials have been equivocal. This approach cannot be considered valid. It will also not be possible to subgroup patients on the basis of apparent severity as determined on different scales. Although longitudinal collection of scores with comparison of interconversion at different time points would have added worthwhile information to our study, in practice meta-analyses have converted all scales at discrete time points and examined the change in the converted value over time. Change in individual scales over time should be addressed in the future.

Current therapeutic trials use a number of different scales; for example, the National Institutes of Neurological Disease and Stroke trial uses NIH, the Australian Streptokinase Trial uses the CNS in the pilot and the NIH in the main study, and phase-three trials of the
free radical scavenger tirilazad use NIH in the United States and middle cerebral artery scales (NSc and Scandinavian) in Europe. The multiple current phase-two studies of potential neuroprotective agents use a similar number of separate scales.

Since it seems improbable that the various proponents of these scales will arrive at a consensus for future trials, is there a possible way forward? The NIH scale is in such widespread use that recognition of the difficulty in converting scores from this scale will not be likely to lead to its abandonment. Adjustment of the weighting of individual items in the NIH scale will not remove the fundamental difficulties inherent in its design, which we have discussed. The CNS and the NSc are sufficiently similar that they are almost interchangeable. The similarity between these and other scales has been recognized by Orgogozo et al, who proposed a unified scoring form. This lists the acute component of the Scandinavian Stroke Scale beside the NSc in an attempt to permit simultaneous collection of two different scores. There are problems with this approach, however. First, the scores are so similar that interconversion should in any case be possible with acceptable error, as we have demonstrated. Second, the equivalence of distinct descriptions of neurological deficits is assumed. Distinct ordinal scores are separated on the basis of a list of subjective descriptive terms. This format has surprisingly been accepted for use in several large trials, under the title of the "Unified Scale" despite not having been validated for interrater reliability or compared with independently obtained scores on the two scales it seeks to unite.

Ideally, future stroke trials will be of sufficient size to obviate the current perceived need for scales. The large number of new pharmaceutical compounds currently in development for treatment of cerebral ischemia will, however, ensure that small trials with restricted entry criteria and a reliance on stroke scales will be with us for many years to come. If we are to continue to use scales, the task will be to reconcile the NIH scale with the "middle cerebral scales." Proposing a further scale would be counterproductive. We recommend an alternative approach, whereby a clinical algorithm is generated to specify the minimum clinical examination required to score a patient on multiple scales. This should not simply amount to the grouping of motor deficits under a broad list of roughly equivalent terms, as the Orgogozo unified system attempts, but must specify well-defined means of objective assessment that will reliably predict an independently obtained score on multiple different scales. The routine collection of these data would improve the validity of extrapolating results from one patient group to another.

References

Interconversion of stroke scales. Implications for therapeutic trials.
K W Muir, D G Grosset and K R Lees

Stroke. 1994;25:1366-1370
doi: 10.1161/01.STR.25.7.1366

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/25/7/1366