Seven-Day NIHSS Is a Sensitive Outcome Measure for Exploratory Clinical Trials in Acute Stroke
Evidence From the Virtual International Stroke Trials Archive

Daniel M. Kerr, BSc; Rachael L. Fulton, MSc; Kennedy R. Lees, MD, FRCP; for the VISTA Collaborators

Background and Purpose—Clinical trials in stroke typically measure outcome after 90 days. Earlier outcome assessment would reduce costs and may increase power. We aimed to compare the sensitivity of 4 end points (modified Rankin Scale [mRS] at 30 and 90 days, and National Institutes of Health Stroke Scale [NIHSS] at 7 and 90 days, analyzed as ordinal measures) to detect the established treatment effect of recombinant tissue-type plasminogen activator (rtPA).

Methods—Within the Virtual International Stroke Trials Archive, we compared rtPA–treated patients with untreated control subjects using a multiple resampling approach. From our total sample we drew 10 000 random samples of unique patients, constraining the sample sizes in treated and untreated groups to be equal. In each of these samples we tested for the treatment effect of rtPA by each of the 4 studied end points. The percentage of samples yielding significant results approximates the power of each end point at a given sample size. This process was repeated across a range of sample sizes, to determine the relationship between sample size and power for each of the 4 end points.

Results—For our 4 end points of mRS at 30 and 90 days, and NIHSS at 7 and 90 days the smallest sample sizes required to generate statistical power >80% were 620, 480, 370, and 420, respectively, making 7-day NIHSS the most sensitive end point. These results were supported by dichotomized analyses.

Conclusions—Seven-day NIHSS score appears a sensitive end point that should be validated in randomized trial datasets for use in exploratory stroke trials. (Stroke. 2012;43:1401-1403.)

Key Words: clinical trials ■ stroke management ■ thrombolysis

Acute stroke trials typically record outcomes after 90 days. Prolonged follow-up increases costs, and risks weakening conclusions due to patient attrition. To enhance efficiency of exploratory trials, there may be benefit from reliance on earlier assessment of outcomes, that is, after 7 or 30 days. This may also minimize confounding from unrelated adverse events. With established efficacy of recombinant tissue-type plasminogen activator (rtPA) on the modified Rankin Scale (mRS) at 90 days, we have an opportunity to explore alternative approaches.

The National Institute of Neurological Disorders and Stroke (NINDS) Thrombolysis Trial Group proposed 24-hour change in National Institutes of Health Stroke Scale (NIHSS) as an end point for part 1 of their trial but 90-day global outcome for part 2; however, they concluded in 2000 that early NIHSS measures may be more sensitive than 90-day measures. This evolution between the NINDS trials, the conflicting rank order of power in subsequent end point analysis work, and a general acceptance that sustained functional benefit should be demonstrated in pivotal trials may together account for the limited implementation of their suggestion.

Nevertheless, if earlier outcome measures are more sensitive to treatment effects than 90-day mRS, they should be implemented in exploratory trials. We investigated the sensitivity of 7-day NIHSS and 30-day mRS to the established treatment effect of rtPA on 90-day mRS, to inform their validity as end points in clinical trials in stroke.

Methods
We compared rtPA–treated patients against untreated control subjects, using data from a previously described trials archive, Virtual International Stroke Trials Archive (VISTA). From this large dataset of treated and untreated patients, we repeatedly created random samples containing equal numbers of treated and untreated patients. We did this across a wide range of sample sizes, and for each sample size we repeated the procedure 10 000 times. Each time
we created a treated and untreated group, we tested for a difference in outcomes on a series of measures: 7-day NIHSS, 30-day mRS, and 90-day NIHSS or mRS. By counting the number of such “trials” that declared a significant treatment difference, we could assess the power to detect a typical thrombolysis treatment effect across a range of sample sizes.

Our primary analysis used ordinal logistic regression. All analyses were adjusted for age and baseline NIHSS score. Full methodology is described in the online-only Data Supplement material (http://stroke.ahajournals.org).

Results

From VISTA, we obtained data on 7886 patients, of whom 4712 met the data requirements for inclusion; 1934 (41.0%) were treated with rtPA. There were important imbalances in baseline demographics (online-only Data Supplement Table 1).

The relationships between sample size and power for each end point are shown in the Figure. Secondary end points based on dichotomization are displayed in online-only Data Supplement Figures 1 and 2. The Table shows the minimum sample sizes required for each end point to achieve statistical power at typical choices of 80% or 90%.

Seven-day NIHSS was confirmed to be the most sensitive outcome measure and 30-day mRS the least. For the 90-day outcome measures, the NIHSS performed modestly better than the mRS, though this difference was negligible on dichotomized analysis.

Discussion

We found NIHSS as measured at 7 days to be the outcome measure most sensitive to the treatment effect of rtPA. This is consistent with conclusions from independent datasets. Nevertheless, potential confounding influences should be considered.

Allocation to the rtPA and control groups in our study was nonrandomized and based on a range of clinical factors, some undocumented. Also, outcome assessors, although masked to investigational treatment allocation, were not blinded to use of alteplase. Some reassurance derives from correspondence of the overall treatment effect, as measured by 90-day mRS, with results of the pooled, randomized trials. Given these limitations, our results deserve external validation using randomized trial data.

The VISTA analysis presented in the present report complements reports by Young et al. and Broderick et al. Young et al found dichotomized NIHSS end points more sensitive to a simulated treatment effect than various dichotomized disability end points. Broderick et al reported 24-hour dichotomized NIHSS to be the most powerful end point. Our analysis is based on a sample that is 8-fold larger and could adjust for known baseline prognostic factors such as age and NIHSS.

The NIHSS reflects impairment, the clinical domain in which the effects of acute stroke therapies are likely to be most marked. In contrast, the mRS covers a broader domain and is influenced by extraneous factors that acute stroke therapies may not influence. Restricting such background noise may improve sensitivity to acute treatment effects.

Table. Minimum Sample Size Required to Achieve Power of 80% for Each Outcome Measure Using Ordinal and Dichotomized Methods

<table>
<thead>
<tr>
<th>End Point</th>
<th>Ordinal</th>
<th>Dichotomized</th>
</tr>
</thead>
<tbody>
<tr>
<td>80% Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-Day mRS</td>
<td>480</td>
<td>610</td>
</tr>
<tr>
<td>30-Day mRS</td>
<td>620</td>
<td>860</td>
</tr>
<tr>
<td>90-Day NIHSS</td>
<td>420</td>
<td>600</td>
</tr>
<tr>
<td>7-Day NIHSS</td>
<td>370</td>
<td>310</td>
</tr>
<tr>
<td>90% Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90-Day mRS</td>
<td>640</td>
<td>780</td>
</tr>
<tr>
<td>30-Day mRS</td>
<td>800</td>
<td>&gt;1000</td>
</tr>
<tr>
<td>90-Day NIHSS</td>
<td>540</td>
<td>770</td>
</tr>
<tr>
<td>7-Day NIHSS</td>
<td>480</td>
<td>410</td>
</tr>
</tbody>
</table>

mRS indicates modified Rankin Scale; NIHSS, National Institutes of Health Stroke Scale.

For each outcome measure, the most sensitive dichotomy was used. The dichotomized methods used are discussed in the online-only Data Supplement Methods.
This interpretation may also contribute to the greater sensitivity of the NIHSS as measured at 7 days compared with 90 days. Extraneous factors may have a greater influence on 90-day than 7-day NIHSS, with the latter better reflecting treatment effects in isolation from the myriad factors which come into play after discharge.

For pivotal trials in acute stroke, the European Medicines Agency (EMA) supports neurological scales mainly as a supplement to ordinal analysis of an activity scale (primarily mRS). However, for early-phase, exploratory research, 7-day NIHSS score has much to recommend it as an end point. It is sensitive to treatment effects, requiring comparatively low sample sizes to achieve desirable levels of statistical power. Moreover, by recording outcome at 7 days, when most participants will be in hospital, losses to follow-up will be minimized and unrelated adverse events should have less influence on detection of adverse reactions. Trial management decisions, such as dose escalation between patient cohorts, can be expedited and costs contained.

In summary, 7-day NIHSS score appears to be an ideal end point for the early exploratory testing of novel agents. Promising agents could then be validated in larger phase III trials using the mRS at 90 days to inform licensing and purchasing decisions, with validation of the 7-day NIHSS end point on the same sample to permit use of prior data as the necessary supporting evidence for regulatory submissions.

### Appendix


### Sources of Funding

R.L. Fulton is supported by studentships from Wyeth/Pfizer and J&J.

### Disclosures

K.R. Lees is Associate Director of the NIHR SRN and chaired the independent data monitoring committee for the ECASS-III trial, chairs the VISTA collaboration, and serves on the Steering Committee for the SITS collaboration. He has received fees and expenses from Boehringer Ingelheim for committee work and lectures.

### References

Seven-Day NIHSS Is a Sensitive Outcome Measure for Exploratory Clinical Trials in Acute Stroke: Evidence From the Virtual International Stroke Trials Archive
Daniel M. Kerr, Rachael L. Fulton, Kennedy R. Lees and for the VISTA Collaborators

Stroke. 2012;43:1401-1403; originally published online February 2, 2012;
doi: 10.1161/STROKEAHA.111.644484

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/43/5/1401

Data Supplement (unedited) at:
http://stroke.ahajournals.org/content/suppl/2012/02/10/STROKEAHA.111.644484.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/
ONLINE SUPPLEMENT

7 Day NIHSS is a Sensitive Outcome Measure for Exploratory Clinical Trials in Acute Stroke: Evidence from the Virtual International Stroke Trials Archive

Daniel M Kerr BSc, Rachael L Fulton MSc, Kennedy R Lees MD, FRCP, for the VISTA Collaborators*

Supplemental Methods

We undertook a controlled comparison of r-tPA treated patients versus untreated controls from VISTA. The characteristics of this database have been described previously\(^1\). Briefly it is a large international database containing individual data on participants from a range of clinical trials in acute stroke. VISTA operating procedures preclude reanalysis of randomised controlled trials of r-tPA for the thrombolysis effect. Instead we considered participants from trials of putative neuroprotectant agents. In these trials up to 50% of patients received r-tPA as part of standard care alongside the trial drug. The decision to administer r-tPA was not randomised, but comparison of r-tPA treated patients with untreated controls offers an approximation of r-tPA related treatment effects that have previously appeared to give a valid estimate. All included trials recorded informed consent from participants; as VISTA data are anonymised further ethical review was not required under institutional guidelines. From VISTA we considered patients with data on receipt of r-tPA, mRS at 30 and 90 days, and NIHSS score at 7 and 90 days. If patients lacked a 90 day mRS value and/or NIHSS score but met all the other data requirements and were known to be alive at 90 days, we carried forward their last recorded mRS and/or NIHSS value. If they were known to be dead at 90 days we ascribed a mRS of 6 and a NIHSS score of 42.

Baseline patient characteristics were compared using Mann Whitney \(U\) tests for continuous variables and \(X^2\) tests for categorical variables. Multiple re-sampling was performed to assess the sensitivity of each variation of the two outcome measures to the treatment effect of r-tPA. From our VISTA dataset 10,000 random samples of patients were drawn at each of a series of sample sizes ranging from 1000 downwards in decrements of 10. Each sample consisted of equal numbers of r-tPA treated and control patients. For every sample size, the percentage of trials yielding a statistically significant treatment effect was recorded. This percentage approximates the statistical power of each outcome measure to detect r-tPA treatment effects at that sample size.

In testing for the treatment effect of r-tPA our primary analysis used ordinal logistic regression. This compares shifts across all categories of the mRS or NIHSS rather
than arbitrarily privileging one health state as in dichotomised analysis. For the purposes of this analysis NIHSS scores were grouped into the following categories: 0 (no measurable deficit), 1-4, 5-8, 9-12, 13-16, 17-20, 21-24, and ≥25.

The EMEA Points to Consider for stroke trial design advises that ordinal analysis may be supported by supplementary analysis of secondary outcome measures. In this regard we undertook dichotomised analysis using binary logistic regression. The power of dichotomous endpoints depends on case-mix and the specific treatment effect studied. To find the optimal dichotomised endpoint for each outcome measure, a range of dichotomies were included in our simulations. For mRS outcome measures we used excellent (mRS 0-1 vs 2-6) and good (mRS 0-2 vs 3-6) functional outcomes; for NIHSS we used the dichotomies 0-1, 0-1 or improvement by 4 or more points from baseline, 0-1 or improvement by 8 or more points from baseline, and 0-1 or improvement by 11 or more points from baseline. The most sensitive dichotomy for each outcome measure was used for comparisons. For 90 and 30 day mRS these were good functional outcomes (mRS 0-2) and excellent functional outcomes (mRS 0-1) respectively. For 90 day and 7 day NIHSS these were 0-1 or improvement of 11 or more points from baseline, and 0-1 or improvement of 8 or more points from baseline, respectively.

In all analyses we adjusted for age and baseline NIHSS score. These are established as the most important predictors of stroke mortality and functional outcomes. Sensitivity analyses performed on the total sample found adjustment for additional variables to have negligible influence on the results.

Simulations were performed using R version 2.12.1, all other analyses used PASW 18. A p-value <0.05 was considered significant.
Table S1 Baseline Characteristics of Patients

<table>
<thead>
<tr>
<th></th>
<th>r-tPA (n= 1934)</th>
<th>Control (n=2778)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>67.2 (12.8)</td>
<td>68.9 (12.2)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>69.5 (18)</td>
<td>71 (17)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Male Sex (%)</strong></td>
<td>1094 (56.6)</td>
<td>1485 (53.5)</td>
<td>0.035</td>
</tr>
<tr>
<td><strong>History of Atrial Fibrillation (%)</strong></td>
<td>440 (22.8)</td>
<td>707 (25.4)</td>
<td>0.034</td>
</tr>
<tr>
<td><strong>History of Diabetes (%)</strong></td>
<td>357 (18.5)</td>
<td>677 (24.4)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>History of Hypertension (%)</strong></td>
<td>1339 (69.2)</td>
<td>2113 (76.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>History of Heart Failure (%)</strong></td>
<td>150 (7.8)</td>
<td>216 (7.9)</td>
<td>0.98</td>
</tr>
<tr>
<td><strong>History of Myocardial Infarction (%)</strong></td>
<td>254 (13.1)</td>
<td>322 (11.6)</td>
<td>0.112</td>
</tr>
<tr>
<td><strong>Prior Stroke (%)</strong></td>
<td>242 (12.5%)</td>
<td>623 (22.4%)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td><strong>Baseline NIHSS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>13.6 (5.0)</td>
<td>11.7 (5.1)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>13 (8)</td>
<td>10 (8)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Figure S1

Figure S1 Line graph showing the relationship between sample size and statistical power for each of the four outcome measures using dichotomised analysis. For each outcome measure the most sensitive dichotomy was used.
Figure S2 Line graph showing the relationship between sample size and statistical power for each of the 7 day NIHSS endpoints studied.
Supplemental References


