Strengthening Acute Stroke Trials Through Optimal Use of Disability End Points

Fiona B. Young, BSc; Kennedy R. Lees, MD, FRCP; Christopher J. Weir, PhD; for the Glycine Antagonist in Neuroprotection (GAIN) International Trial Steering Committee and Investigators

Background and Purpose—Suboptimal choices of primary end point for acute stroke trials may have contributed to inconclusive results. The Barthel Index (BI) and Rankin Scale (RS) have been widely used and analyzed in various ways. We sought to investigate the most powerful end point for use in acute stroke trials.

Methods—Data from the Glycine Antagonist in Neuroprotection (GAIN) International Trial were used to simulate 24,000 clinical trials exploring various patterns and magnitudes of treatment effect and thus to estimate the statistical power for a range of end points based on the BI or RS.

Results—RS end points were more powerful than BI end points. End points dichotomized toward the favorable extreme of either scale or adjusted according to baseline prognosis (“patient-specific” end point) were among the most powerful. Combining RS and BI in a “global” end point was also successful. Improvements in statistical power indicated that using a RS end point instead of BI ≥60 could reduce the sample size by up to 84% (95% CI, 80% to 87%), 73% (95% CI, 68% to 79%) for a patient-specific BI end point, or 81% (95% CI, 76% to 85%) for a global end point.

Conclusions—The RS and global end points are preferable to BI end points; the position of the cut point is also important. Better choices of end point substantially strengthen trial power for a given trial size or allow reduced sample sizes without loss of statistical power. (Stroke. 2003;34:2676-2680.)

Key Words: clinical trials ■ end point determination ■ neuroprotection ■ stroke, acute ■ thrombolysis
zation at this mild end of the scale disadvantages the contribution of more severely affected patients; on average, their outcomes will be much poorer, and small but valuable improvements caused by treatment would not be measured. To allow both mildly and moderately severely affected patients to contribute to the significance test, a second cut point can be added, forming a trichotomized analysis.

Another approach is to use a global end point, simultaneously incorporating outcome measures from different domains such as handicap and activities of daily living. This is conceptually appealing because no single outcome measure describes all dimensions of recovery from stroke, yet it has received limited attention to date. The statistical power of a global end point should be greater than or equal to that of an individual end point\(^1^{10}\) but may be weakened with the inclusion of a scale less influenced by a treatment.

There is considerable heterogeneity in stroke severity; using an end point with a fixed cut point may render many patients uninformative. It may be appropriate to group patients according to clinical presentation and to vary cut points according to group. This “patient-specific” end point would give a more realistic assessment of a treatment effect and allow all patients to contribute to the results of the trial.

This article explores the optimal primary end points incorporating the BI and RS. We assessed a selection of end points used in published trials as well as patient-specific and global end points. We sought to establish which end point would perform best under likely trial circumstances. The Oxford classification\(^{11}\) was used to categorize patients by clinical presentation in this study.

**Methods**

Our statistical approach is described in an appendix to this article (available online at http://stroke.ahajournals.org). Briefly, we based our work on the patients from the Glycine Antagonist in Neuroprotection (GAIN) International Trial\(^8\) data set. The GAIN trial was neutral; however, to avoid any bias, only the placebo patients were used. We generated 24,000 clinical trials, each with 1400 patients split between active treatment and placebo groups (700 per group), representing 33.6 million randomized patients. Within each trial, patients were simulated by randomly sampling with replacement from the GAIN data. The characteristics of every simulated patient were based on a real example from the GAIN trial, preserving the correlation between the National Institutes of Health Stroke Scale (NIHSS)\(^{12}\) Oxford classification, and final outcome described by RS and BI. The placebo and treatment groups were generated slightly differently, so that the simulated treatment group was forced to have slightly milder stroke as assessed by NIHSS at baseline. The difference between the average NIHSS score for the 2 groups varied from 0 through 4 points (described as treatment level), but for clarity our results concentrate on the 2-point difference. This treatment level is equivalent to a relative risk reduction in being dead or disabled of 9%, an absolute risk reduction of 4%, or an odds ratio of 1.19, with the use of BI ≥60.

The above “fixed” effect was our most basic approach since it assumes that treatment is uniformly effective in all patients. Consequently, we also simulated effects in which benefit from treatment was dependent on certain patient characteristics, such as age and sex (neuroprotective effect, denoted NP); in which a uniform benefit was offset in a randomly selected subgroup by deterioration to mimic the effect of thrombolysis (TP1); and finally, an effect that was dependent on patient characteristics, with deterioration in some patients (TP2). In summary, there were 24,000 trials: 1500 simulated trials for each of 4 treatment effects and 4 treatment levels, with every trial involving 1400 patients.

**End Points**

Published cut points were used when we dichotomized or trichotomized the BI and/or RS (Table 1). We also explored patient-specific cut points, in which we specified different thresholds for favorable outcome according to baseline prognosis, using the Oxford classification to group patients. We chose thresholds that were close to the median value of BI or RS achieved by each Oxford classification category in the original GAIN trial.

**Estimation of Statistical Power**

We analyzed the simulated trials via Pearson’s \(x^2\) test for dichotomized end points and the Cochran-Mantel-Haenszel \(x^2\) test\(^{13}\) for trichotomized end points. The global end points were analyzed via generalized estimating equations.\(^{14}\) A bootstrap approach was used to calculate CIs for the power.

The end points were compared by calculating the sample size that would be required to maintain the same statistical power when 1 end point was chosen in preference to BI ≥60 with the use of standard sample size equations.\(^{15}^{–}^{17}\) If an end point were more powerful, the required sample size expressed as a percentage would be <100%. For an overall comparison of the end points, binary logistic regression was used to model the proportion of significant trials, adjusted for treatment effect size.

**Results**

The pattern of results we observed was similar across all treatment effect patterns for both the RS and BI end points (Table 2). The NP effect and the TP2 effect could be detected with the lowest power. The BI ≥60 dichotomy was consistently the least powerful end point. Among the remaining BI end points, the ≥95 dichotomy and the patient-specific dichotomized end points were equally the most powerful (Figure). The RS end points followed a less consistent pattern. The RS ≤2 end point was the least powerful for all treatment effect patterns; end points incorporating RS ≤3 were no better (data not shown). Depending on the treatment effect pattern, the RS ≤1, the RS ≤1 and ≥2 trichotomy, or the dichotomized patient-specific end points would give a more realistic assessment of a treatment effect and allow all patients to contribute to the results of the trial.


### Table 1. End Points Assessed

<table>
<thead>
<tr>
<th>Scale</th>
<th>End Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthe Index</td>
<td>≥60 dichotomy</td>
</tr>
<tr>
<td></td>
<td>≥95 dichotomy</td>
</tr>
<tr>
<td></td>
<td>≥60 and ≥95 trichotomy</td>
</tr>
<tr>
<td>Patient-specific dichotomy</td>
<td>(LACI, PACI and POCI ≥95; TACI ≥60)</td>
</tr>
<tr>
<td>Rankin Scale</td>
<td>≤2 dichotomy</td>
</tr>
<tr>
<td></td>
<td>≤1 dichotomy</td>
</tr>
<tr>
<td></td>
<td>≤1 and ≥2 trichotomy</td>
</tr>
<tr>
<td>Patient-specific dichotomy</td>
<td>(LACI and POCI ≤1; PACI and TACI ≤2)</td>
</tr>
<tr>
<td>Global outcome</td>
<td>(1) Dichotomy (BI ≥95 and RS ≤1)</td>
</tr>
<tr>
<td></td>
<td>(2) Patient-specific dichotomy</td>
</tr>
<tr>
<td></td>
<td>(BI: LACI, PACI and POCI ≥95; TACI ≥60)</td>
</tr>
<tr>
<td></td>
<td>(RS: LACI and POCI ≤1; PACI and TACI ≤2)</td>
</tr>
</tbody>
</table>

 BI indicates Barthel Index; RS, Rankin Scale; LACI, lacunar infarction; PACI, partial anterior circulation infarct; POCI, posterior circulation infarct; TACI, total anterior circulation infarct.
point was the most powerful. The range of power was narrower for the RS end points than for the BI end points.

Both the dichotomized and patient-specific global end points were more powerful than the BI end points for all treatment effect patterns but not always more powerful than RS \( \geq 1 \) or the RS \( \leq 2 \) trichotomy. Generally, the patient-specific global end point was less powerful than the dichotomized global end point.

Table 3 compares the end points in terms of required sample sizes relative to BI \( \geq 60 \). For the BI end points, the greatest sample size reduction was obtained under the TP2 effect and the patient-specific end point or the BI \( \geq 95 \) end point. The RS end points generally had larger sample size reductions. Either of the global end points could reduce the sample sizes even further, depending on the underlying treatment effect pattern.

Overall, the RS end points were more powerful than the BI end points (Table 4). The odds of achieving a statistically significant result increased by 89% under a fixed treatment effect if a RS end point were used instead of a BI end point.

### Discussion

Our results have important implications for the choice of primary end point in acute stroke trials. Primary end points that include the RS are more powerful than those based on the BI. The position of the cut point on these scales is also of great importance; end points dichotomized toward the favorable extreme were more powerful. The patient-specific BI and the trichotomized RS also performed well.

Our analyses were performed with a range of treatment effects, and our findings are reasonably consistent across a likely range of trial conditions. However, since all analyses used the GAIN International database, applying the end points to an independent data set may be informative.

Broderick and colleagues\(^{18}\) used National Institute of Neurological Disorders and Stroke (NINDS) trial data and established that the RS dichotomized at \( \geq 1 \) was the most effective in differentiating between the treatment groups in that trial. The BI dichotomized at \( \geq 95 \) was also effective. However, since such an analysis is data dependent, it may not be generalizable. An analysis that relies solely on choosing positive end points from a selection of trials in which putative effects may have been seen is subject to selection bias and random variability. Our method involves assumptions about the generation of the treatment effect (since it assumes that outcome at 90 days is related to the underlying data). Our cut points were chosen on the basis of the Oxford classification category; further work is required to assess more appropriate methods of selecting the cut points.

The inclusion of only the BI and RS in the global end points may have restricted the power. These outcome measures are highly correlated; the full potential of a global end point to assess many different dimensions of recovery was not exploited. The

### Table 2. Statistical Power Obtained for Each End Point

<table>
<thead>
<tr>
<th>Scale</th>
<th>End Point</th>
<th>NP</th>
<th>TP1</th>
<th>TP2</th>
<th>Fixed*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthel Index</td>
<td>( \geq 60 )</td>
<td>0.353 (0.329, 0.378)</td>
<td>0.135 (0.118, 0.153)</td>
<td>0.170 (0.151, 0.189)</td>
<td>0.103 (0.087, 0.118)</td>
</tr>
<tr>
<td></td>
<td>( \geq 95 )</td>
<td>0.639 (0.614, 0.663)</td>
<td>0.263 (0.241, 0.286)</td>
<td>0.431 (0.406, 0.456)</td>
<td>0.241 (0.219, 0.262)</td>
</tr>
<tr>
<td></td>
<td>( \geq 60 ) and ( \geq 95 )</td>
<td>0.565 (0.540, 0.590)</td>
<td>0.212 (0.191, 0.233)</td>
<td>0.309 (0.286, 0.333)</td>
<td>0.169 (0.150, 0.188)</td>
</tr>
<tr>
<td></td>
<td>PS dichotomy</td>
<td>0.575 (0.550, 0.600)</td>
<td>0.265 (0.243, 0.288)</td>
<td>0.357 (0.333, 0.382)</td>
<td>0.261 (0.238, 0.283)</td>
</tr>
<tr>
<td>Rankin Scale</td>
<td>( \leq 2 )</td>
<td>0.703 (0.680, 0.726)</td>
<td>0.233 (0.212, 0.255)</td>
<td>0.455 (0.429, 0.480)</td>
<td>0.259 (0.237, 0.282)</td>
</tr>
<tr>
<td></td>
<td>( \leq 1 )</td>
<td>0.760 (0.738, 0.782)</td>
<td>0.270 (0.248, 0.292)</td>
<td>0.613 (0.588, 0.637)</td>
<td>0.412 (0.387, 0.437)</td>
</tr>
<tr>
<td></td>
<td>( \leq 1 ) and ( \leq 2 )</td>
<td>0.779 (0.758, 0.800)</td>
<td>0.272 (0.249, 0.295)</td>
<td>0.575 (0.550, 0.600)</td>
<td>0.369 (0.345, 0.394)</td>
</tr>
<tr>
<td></td>
<td>PS dichotomy</td>
<td>0.735 (0.713, 0.758)</td>
<td>0.277 (0.254, 0.299)</td>
<td>0.559 (0.534, 0.584)</td>
<td>0.363 (0.338, 0.387)</td>
</tr>
<tr>
<td>Global end point</td>
<td>Global dichotomy</td>
<td>0.767 (0.746, 0.789)</td>
<td>0.291 (0.268, 0.314)</td>
<td>0.577 (0.552, 0.602)</td>
<td>0.386 (0.361, 0.411)</td>
</tr>
<tr>
<td></td>
<td>Global PS dichotomy</td>
<td>0.715 (0.692, 0.738)</td>
<td>0.294 (0.271, 0.317)</td>
<td>0.505 (0.480, 0.531)</td>
<td>0.365 (0.340, 0.389)</td>
</tr>
</tbody>
</table>

Power levels for a treatment effect equivalent to a 2-point shift in baseline NIHSS. For BI \( \geq 60 \) this is approximately OR=1.19, RRR=9%, ARR=4%. Parentheses contain the 95% confidence interval for the power.

PS indicates patient specific; NP, neuroprotective treatment effect; TP1 and TP2, thrombolytic treatment effects.

* Power levels achieved with a fixed 3-point shift in baseline NIHSS, for the BI \( \geq 60 \) end point this is approximately OR=1.42, RRR=18%, ARR=9%.
inclusion of other outcome measures such as the NIHSS may further improve the power, as used in the NINDS trial. However, some regulatory authorities, such as the European Medicines Evaluation Authority, have been reluctant to consider a global end point that combines diverse outcome measures.

Most stroke trials are powered to detect an absolute risk reduction of 10%. This study used a treatment effect level that was equivalent to an absolute risk reduction of 4% (BI ≥60, fixed effect). We believe that this is a more realistic effect of a stroke intervention. This has resulted in levels of statistical power substantially below 80%, suggesting that the sample size of 1400 is too small. The final column in Table 2 shows the power that was achieved when a 3-point decrease in baseline NIHSS was applied (absolute effect of 9%). With this larger treatment effect, the power for all end points exceeds 80%, and it is sufficient to render a trial informative, but it may be a prerequisite. Substantial and significant increases in power are observed when a dichotomized end point cut at the favorable extreme of the BI or RS, a patient-specific end point, or a global end point is used. On average, RS end points appear more powerful than BI end points, whether analyzed alone or as part of a global end point.

In conclusion, this study has shown that many clinical trials in acute stroke have not used an optimal primary end point, which may have led to inconclusive results. Statistical power is not sufficient to render a trial informative, but it may be a prerequisite. Substantial and significant increases in power are observed when a dichotomized end point cut at the favorable extreme of the BI or RS, a patient-specific end point, or a global end point is used. On average, RS end points appear more powerful than BI end points, whether analyzed alone or as part of a global end point.

Acknowledgments

This study was supported by a collaborative studentship from the Medical Research Council and Pfizer to F.B. Young and by a Medical Research Council career development fellowship to Dr Weir. The GAIN trial was sponsored by GlaxoWellcome (now GlaxoSmithKline). GlaxoSmithKline had no involvement in this analysis or article.

References


Strengthening Acute Stroke Trials Through Optimal Use of Disability End Points
Fiona B. Young, Kennedy R. Lees and Christopher J. Weir
for the Glycine Antagonist in Neuroprotection (GAIN) International Trial Steering Committee
and Investigators

Stroke. 2003;34:2676-2680; originally published online October 16, 2003;
doi: 10.1161/01.STR.0000096210.36741.E7
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2003 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/34/11/2676

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/