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

Most clinical trials in acute stroke have been unsuccessful in demonstrating a positive therapeutic effect. Neuroprotective trials have likely been underpowered to detect subtle but clinically important treatment effects. Statistical power is the probability that a statistical test identifies a significant treatment effect (where one truly exists) at a given significance level and sample size. Inappropriate choice of cut point for analysis of the outcome scale(s) may be one of several factors contributing to a lack of statistical power. Efforts must be made to optimize the analysis of clinical trials for both ethical and practical reasons.

A variety of primary end points have been used in acute stroke trials. The Barthel Index (BI) and the modified Rankin Scale (RS) have been the most commonly used disability outcome measures. The BI is a 10-item scale in which disability is assessed on various aspects of self-care, such as dressing and toilet use. It has a maximum score of 100 (fully independent, physically functioning). The RS is a 6-point scale in which disability is assessed on various aspects of self-care, such as dressing and toilet use. It has a maximum score of 100 (fully independent, physically functioning).

To date, functional outcome scores have usually been dichotomized as favorable versus unfavorable, although there is little consensus on the optimal cut point, and selection of this is often arbitrary. The most commonly used end point in published trials has been the BI cut point at 60, at which a patient is thought to be capable of independence from full-time care. BI cut points have ranged from 55 to 100. The RS has been used less frequently, although outcomes of ≤2 (slight or no disability) and ≤1 (no significant disability) have been utilized. A trichotomized BI end point (split into 3 categories) has also been used.

The BI has a U-shaped distribution, in which patient outcomes cluster at the extremes. The quarter of patients who die are arbitrarily scored 0; the 40% who recover are scored 95 or 100. Since the remaining third have BI scores distributed between 5 and 90, any cut point selected within this range will have a small number of patients populating the adjacent categories: as few as 5% of the patients may lie 5 or 10 points below a cut point of 60. If it is assumed that a drug treatment effect will improve patients by only 1 or 2 BI points, however, the RS may be less reproducible because of its relative lack of structure.

To date, functional outcome scores have usually been dichotomized as favorable versus unfavorable, although there is little consensus on the optimal cut point, and selection of this is often arbitrary. The most commonly used end point in published trials has been the BI cut point at 60, at which a patient is thought to be capable of independence from full-time care. BI cut points have ranged from 55 to 100. The RS has been used less frequently, although outcomes of ≤2 (slight or no disability) and ≤1 (no significant disability) have been utilized. A trichotomized BI end point (split into 3 categories) has also been used.

The BI has a U-shaped distribution, in which patient outcomes cluster at the extremes. The quarter of patients who die are arbitrarily scored 0; the 40% who recover are scored 95 or 100. Since the remaining third have BI scores distributed between 5 and 90, any cut point selected within this range will have a small number of patients populating the adjacent categories: as few as 5% of the patients may lie 5 or 10 points below a cut point of 60. If it is assumed that a drug treatment effect will improve patients by only 1 or 2 BI points, however, the RS may be less reproducible because of its relative lack of structure.
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 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 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 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 1,400 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) and 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 thrombolyis (TP1); and finally, an effect that was dependent on patient characteristics, with deterioration in some patients (TP2). In summary, there were 24,000 trials: 1,500 simulated trials

### 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).

<table>
<thead>
<tr>
<th>Scale</th>
<th>End Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barthel 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></td>
<td>Patient-specific dichotomy</td>
</tr>
<tr>
<td>(LACI, PACI and POCl ≥95; TACI ≥60)</td>
<td></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></td>
<td>Patient-specific dichotomy</td>
</tr>
<tr>
<td>(LACI and POCl ≤1, PACI and TACI ≤2)</td>
<td></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>(BI: LACI, PACI and POCl ≥95; TACI ≥60)</td>
<td></td>
</tr>
<tr>
<td>(RS: LACI and POCl ≤1; PACI and TACI ≤2)</td>
<td></td>
</tr>
</tbody>
</table>

BI indicates Barthel Index; RS, Rankin Scale; LACI, lacunar infarction; PACI, partial anterior circulation infarct; POCl, posterior circulation infarct; TACI, total anterior circulation infarct.
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.

Dichotomization may be less sensitive than trichotomized end points, global end points, or patient-specific end points. Berge and Barer supported the separate definition of favorable outcomes for subgroups of patients to maximize the power of stroke trials. Patient-specific end points would ensure that trial results are generalizable across a wide range of stroke severity. 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
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%.

It is not only the choice of end point that can influence the power of a clinical trial: validity of outcome measures and restrictive entry criteria may also be factors. STAT and NINDS both restricted time to treatment to 3 hours. PROACT II restricted entry to patients with proven middle cerebral artery occlusion.

We have demonstrated the disadvantage of BI ≥60 as a primary end point. Trials that are currently in progress should consider revisions to their statistical analysis plan before unblinding takes place to optimize statistical power. Such a decision has recently been announced by the international Intravenous Magnesium Efficacy in Stroke (IMAGES) trial group.

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.

**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