Improving Trial Power Through Use of Prognosis-Adjusted End Points

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

Background and Purpose—The stroke patient population is heterogeneous, leading to wide variation in outcome caused by differences in age, initial severity, and presence of concomitant disease. Setting an identical recovery target for all patients in intervention trials may conceal individually important therapeutic treatment effects. Instead, a variable end point that takes severity or likely prognosis into account may be more informative.

Methods—We used data from the Glycine Antagonist in Neuroprotection (GAIN) International trial to assess statistical power of various primary end points for intervention trials. We selected prognosis-adjusted cut points based on Barthel Index (BI) or Rankin Scale (RS) using a prognostic model, or assigned a fixed end point within subgroups of patients defined by their Oxford category or National Institutes of Health Stroke Scale (NIHSS) score. We simulated a treatment effect and estimated statistical power with standard formulae.

Results—Assignment of end points using a prognostic model for individual patients increased statistical power, when compared with assigning end points using only the Oxford classification. For the BI, power was increased from 60% to 88% (equivalent to a 49% reduction in sample size if power remains unchanged). With the RS end points, power was increased from 84% to 92% (or a 24% reduction in sample size). Versus a fixed end point for all patients, model-based methods increased power by 22 percentage points for BI≥95 and 14 percentage points for RS≤1 (effective sample size reductions 43% and 34%).

Conclusion—Prognosis-adjusted end points can increase statistical power compared with fixed end points. Assessment is based on realistic goals for individual patients and yet trial results remain generalizable. (Stroke. 2005;36:597-601.)

Key Words: clinical trials ■ stroke, acute

The stroke patient population is heterogeneous. Assigning a common end point for an entire stroke trial population of patients could mask important findings in trial results. For example, a patient with a severe stroke may never reach total independence (Barthel Index1 [BI]≥95) but may still consider being independent from essential care (BI≥60) to be a good outcome. On the other hand, relatively mildly affected patients may only consider themselves to have recovered if they reach total independence. To account for the heterogeneity, the cut point on a recovery scale could be varied for individual patients. This would tailor end points to suit the prognosis or baseline characteristics of patients. Such variation of cut points in a trial would allow patients to be assessed on achievable goals, while ensuring that the trial remains generalizable. Allowing for patients to be assessed on different goals applies the principles of goal attainment scaling,2 which considers the outcome to be favorable if the patient achieves a prespecified objective.

A few trials have already used variable end points to take account of initial severity. The Stroke Treatment with Ancrod Trial (STAT)3 assessed whether patients achieved a BI score of at least 95 or had a score at 3 months at least equal to their prestroke value. The Abciximab in Emergent Stroke Treatment Trial (AbESTT; H.P. Adams, written communication, November 2002) used a dichotomized Rankin Scale (RS)4 secondary end point that split patients into 3 groups using the baseline National Institutes of Health Stroke Scale (NIHSS) score for each patient:5,6 RS≤0 for baseline NIHSS of 0 to 7, RS≤1 for baseline NIHSS of 8 to 14, and RS≤2 for baseline NIHSS >14. Berge and Barer7 also suggested that it may be appropriate to use variable criteria for assessing patients with different stroke severity levels but did not define the criteria to be used.

We have previously described the use of prognosis-adjusted end points for the BI and RS,8 where we divided patients into subgroups using the Oxford classification.9 We chose appropriate cut points for each of the Oxford category by picking a value on the RS and BI that fell close to the median value of the scale within each Oxford category, basing our estimates on placebo data from the Glycine...
TABLE 1. End Points Using Subgroups to Assign Prognosis-Adjusted Cut Points

<table>
<thead>
<tr>
<th>Scale</th>
<th>Subgroup Method</th>
<th>Cut Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>Oxford category LACI: RS ≤ 50</td>
<td>RS ≤ 50</td>
</tr>
<tr>
<td></td>
<td>PACI: RS ≥ 60</td>
<td>PACI ≥ 60</td>
</tr>
<tr>
<td></td>
<td>TACI: RS ≤ 90</td>
<td>TACI ≤ 90</td>
</tr>
<tr>
<td>BI</td>
<td>Oxford category LACI: BI ≤ 95</td>
<td>BI ≤ 95</td>
</tr>
<tr>
<td></td>
<td>PACI: BI ≥ 95</td>
<td>PACI ≥ 95</td>
</tr>
<tr>
<td></td>
<td>TACI:BI ≥ 60</td>
<td>TACI ≥ 60</td>
</tr>
</tbody>
</table>

LACI indicates lacunar infarction; PACI, partial anterior circulation infarction; POCl, posterior circulation infarction; TACI, total anterior circulation infarction.

Antagonist in Neuroprotection (GAIN) International trial.10 To enhance the selection of end points, we now consider the use of factors that are more closely related to baseline severity and prognosis, because assignment of cut points using an approach that incorporates a prognostic model may be potentially more powerful.

We aimed to assess a range of prognosis-adjusted end points. We examined cut points for both the RS and BI that were assigned either using subgroups of patients or using prognostic models. These prognosis-adjusted end points were compared according to their statistical power estimated under likely trial circumstances and were also contrasted with power estimates for the traditional fixed end points of RS ≤ 1 and BI ≥ 95.

Methods

End Points Assessed

The end points that we assessed can be considered in 2 categories: those in which cut points were assigned to subgroups of patients and those that used a prognostic model to assign cut points to patients. The subgroup approach assigned cut points based on the pretreatment assessment of either the Oxford Classification or the NIHSS (Table 1). Our other approach used a prognostic model to assign the cut points to individual patients. We built separate models for each of the RS and BI, using ordinal logistic regression.11 We initially split each scale into 3 categories: 0 to 55, 60 to 90, and 95 to 100 for the BI and 0 to 1, 2, and 3 to 6 for the RS. Later, we considered more complex models with extra response categories. We used stepwise regression: variables entered into the model for selection included total baseline NIHSS, age, Oxford classification, individual components of the baseline NIHSS, and various risk factors. To take account of the side of the stroke, we added factors for the worst leg and arm scores from the NIHSS. We continued to add variables to the model for as long as inclusion of a variable significantly improved the amount of variation explained by the model. However, for both outcome scales we also developed simple models that incorporated only age and baseline NIHSS.

We developed all models using the GAIN International trial10 placebo data. There was no significant treatment effect observed in the GAIN trial, so the final prognostic models were tested on the GAIN investigational treatment group data to check for generalizability and accuracy.

Our prognostic models predicted the cumulative probability of a given patient being in each outcome category. Initially, we set the probability threshold at 50% (patients were assigned cut points that they had at least a 50% chance of achieving), but later we considered alternative thresholds to find the optimal cut point. For example, using a 3 category BI model with categories set to 0 to 55, 60 to 90, and ≥ 95 and with a probability threshold of 50%, the highest outcome category would be predicted if a given patient had at least a probability of 50% of achieving a score of 95 or more. If the probability of achieving a score of 95 or more was < 50% but the probability of obtaining a score of 60 or more was at least 50%, then the outcome category of the patient was predicted to be 60 to 90. If neither of these constraints was achieved, then the outcome was predicted as 0 to 55. We took the lower bound of these ranges for the cut point (ie, 60 for a patient predicted to lie between 60 and 90).

We also investigated alternative probability thresholds, ranging from 45% to 5%. Assessing patients on end points that are toward the favorable end of the outcome scale had been more powerful in a previous study, and hence if the probability threshold was reduced from 50%, then patients would be assessed on a cut point that would be more difficult to attain (movement of cut point toward the most favorable extreme of the scales).

Simulation of Treatment Effect and Estimation of Statistical Power

We simulated a treatment effect which assumed that all patients would derive equal benefit from the treatment. In practice this is unlikely to be a valid assumption; however, our previous work has shown that more complex treatment patterns tend to alter the magnitude but not the direction of the conclusions. Our method consisted of applying treatment effects to the GAIN International placebo data. The treatment effect in terms of an odds ratio was estimated using the data from a previous simulation study. Each simulated clinical trial in the study consisted of 2 treatment groups. Patients entered into the simulated placebo groups simply were randomly selected with replacement from the GAIN data. In contrast, patients entered into the simulated treatment groups were selected so as to have had a milder stroke on average than the placebo group (ie, they had lower baseline NIHSS scores). We assumed that this would confer a greater chance of favorable outcome at 90 days. We believe that this is the closest artificial treatment effect that we can generate to mimic a scenario in which a neuroprotectant or thrombolytic has an early effect to limit infarct extent.

TABLE 2. Statistical Power of Barthel Index Patient Specific End Points

<table>
<thead>
<tr>
<th>End Point</th>
<th>Treatment Effect Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Oxford category*</td>
<td>0.212 (0.198, 0.229)</td>
</tr>
<tr>
<td>50% probability threshold</td>
<td>0.133 (0.121, 0.142)</td>
</tr>
<tr>
<td>Model BI1</td>
<td>0.239 (0.221, 0.254)</td>
</tr>
<tr>
<td>Model BI2</td>
<td>0.239 (0.221, 0.254)</td>
</tr>
</tbody>
</table>

Parentheses contain 95% CI for the power.

*The Oxford category end point constituted BI ≥ 95 for LACI, PACI, and POCl; BI ≥ 60 for TACI.
The simulated clinical trials were used to estimate the difference between groups for each given treatment level in terms of an odds ratio. Treatment level could be defined as the difference in baseline NIHSS score that we had artificially generated (0, 1, 2, or 3 points). We then examined the 3-month outcomes of the patients and estimated differences between the active treatment and placebo groups for each end point. Bootstrap CIs12 for the odds ratios were constructed using 1000 replications. We calculated the statistical power of each end point and its 95% CI using standard formulae.13 We could then compare end points on the basis of this estimate of statistical power. We also calculated any potential reduction in sample size that we could introduce without reduction of power, assuming that we used the revised end point instead of a fixed end point that incorporated BI≥95 or RS≤1.

**Results**

**Barthel Index End Points**

From stepwise ordinal logistic regression, we found baseline NIHSS, age, presence of diabetes, the gaze component of the NIHSS, and the worst leg NIHSS motor score to be the best predictors of outcome at 90 days (we labeled this model BI1). We called our simple robust BI model that included only baseline NIHSS and age, BI2.

Compared with using the Oxford classification to assign cut points to patients, only model BI2 increased the statistical power (Table 2). We found no further improvement in power through increasing the model complexity.

**Rankin Scale End Points**

From stepwise regression, we found baseline NIHSS, age, and worst leg score to be closely related to RS outcome. We called this model RS1, and a further model that included only baseline NIHSS and age, RS2.

On average, the RS patient specific end points had higher statistical power than the BI end points (Table 3). Subgrouping the patients by Oxford category delivered the lowest statistical power, whereas subgrouping by NIHSS produced the highest power. When we used a model instead of subgroups, we found that we achieved good power through the simple approach of RS2 that controlled only for age and baseline NIHSS.

**Optimal Probability Threshold for Model-Based End Points**

For the analyses above, we assigned target outcomes to patients such that there would be a 50% chance that they would attain the chosen recovery target. The statistical power of our model-based prognosis-adjusted end points was further improved when the probability thresholds were optimized (Table 4). For the BI model, greatest power was obtained with a probability threshold of 25%, whereas for the RS model the optimal probability threshold was 30%. This suggests that assessing patients on stricter criteria for recovery results in a more powerful end point. The power that we obtained by using the RS2 model at its optimal probability threshold exceeded that obtained when patients were subgrouped by NIHSS: 0.921 (95% CI: 0.913, 0.928) compared with 0.882 (95% CI: 0.871, 0.893).

**Comparison of Prognosis-Adjusted End Points to Fixed End Points**

We compared the prognosis-adjusted end points to what could be considered the best fixed end points (BI≥95 or RS≤1). With a treatment level of 2, the BI≥95 and RS≤1 fixed end points obtained statistical power of 0.657 (95% CI: 0.639, 0.676) and 0.784 (95% CI: 0.770, 0.800), respectively. These values are inferior to those obtained with the model-based or NIHSS subgroup-based prognosis-adjusted end points.

**Comparison of End Points in Terms of a Relative Sample Size**

Finally, we compared selected end points in terms of a relative sample size (Table 5). All of the model-based end points used the optimal probability thresholds discussed in the previous section. For the BI end points, the sample size

### Table 3. Statistical Power of Rankin Scale Patient Specific End Points

<table>
<thead>
<tr>
<th>End point</th>
<th>Threshold</th>
<th>Power (Model BI2)</th>
<th>Power (Model RS2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxford category*</td>
<td>0.222</td>
<td>0.840 (0.828, 0.853)</td>
<td>0.996 (0.995, 0.997)</td>
</tr>
<tr>
<td>NIHSS†</td>
<td>0.311</td>
<td>0.882 (0.871, 0.893)</td>
<td>0.995 (0.994, 0.996)</td>
</tr>
<tr>
<td>50% probability threshold</td>
<td>Model RS1</td>
<td>0.248 (0.231, 0.266)</td>
<td>1.000 (1.000, 1.000)</td>
</tr>
<tr>
<td></td>
<td>Model RS2</td>
<td>0.298 (0.279, 0.317)</td>
<td>0.998 (0.997, 0.999)</td>
</tr>
</tbody>
</table>

Parentheses contain 95% CI for the power.

*The Oxford category end point contained RS=1 for LACI and POCI; RS≥2 for PACI and TACI.
†End point based on NIHSS (NIHSS 4 to 7: RS=1; NIHSS 8 to 14: RS=2; NIHSS≥15: RS=2).

### Table 4. Effect of Probability Threshold on Statistical Power

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Power (Model BI2)</th>
<th>Power (Model RS2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.669 (0.650, 0.689)</td>
<td>0.766 (0.750, 0.782)</td>
</tr>
<tr>
<td>0.10</td>
<td>0.700 (0.682, 0.719)</td>
<td>0.802 (0.787, 0.815)</td>
</tr>
<tr>
<td>0.15</td>
<td>0.763 (0.747, 0.781)</td>
<td>0.779 (0.761, 0.795)</td>
</tr>
<tr>
<td>0.20</td>
<td>0.806 (0.790, 0.819)</td>
<td>0.771 (0.754, 0.789)</td>
</tr>
<tr>
<td>0.25</td>
<td>0.879 (0.867, 0.890)</td>
<td>0.878 (0.867, 0.888)</td>
</tr>
<tr>
<td>0.30</td>
<td>0.848 (0.835, 0.861)</td>
<td>0.921 (0.913, 0.928)</td>
</tr>
<tr>
<td>0.35</td>
<td>0.784 (0.767, 0.798)</td>
<td>0.876 (0.864, 0.886)</td>
</tr>
<tr>
<td>0.40</td>
<td>0.638 (0.619, 0.660)</td>
<td>0.848 (0.837, 0.860)</td>
</tr>
<tr>
<td>0.45</td>
<td>0.670 (0.651, 0.690)</td>
<td>0.859 (0.847, 0.869)</td>
</tr>
<tr>
<td>0.50</td>
<td>0.701 (0.686, 0.720)</td>
<td>0.797 (0.784, 0.812)</td>
</tr>
</tbody>
</table>

Parentheses contain 95% CI for the power.
could be reduced by 43% (95% CI: 41%, 45%) if the model BI2 was used instead of the BI≥95 dichotomized end point. Using the BI2 model end point rather than the Oxford category end point would allow an effective sample size reduction of 50% (95% CI: 48%, 52%). For the RS end points, using the RS2 model to assign cut points could reduce the sample size by 34% (95% CI: 32%, 36%) compared with the RS≤1 dichotomy. If the NIHSS subgroup end point was used instead of the RS≤1 dichotomy, the sample size could be reduced by 24% (95% CI: 22%, 28%). Using the Oxford category to subgroup patients could result in reductions in sample size of 14% (95% CI: 11%, 17%).

**Discussion**

We have shown that prognosis-adjusted end points may be preferable to fixed end points. The prognosis-adjustment approach can be enhanced by assigning cut points based on individually estimated patient prognosis rather than categorization into subgroups. Our simple model-based approach to assign appropriate cut points for patients using age and baseline NIHSS to estimate prognosis produced the greatest power advantage for the BI end points. This is partly because the BI starts with low inherent power. Splitting the patients into groups using the NIHSS (as used in the AbESTT) before assigning RS end points proved effective but so was our model-based approach that considered the age and baseline NIHSS of the patients. These prognosis-adjusted end points were almost invariably more favorable than the best traditional fixed end points.

Our model-based method performed better than the approach in which prognosis was identified from NIHSS or Oxford subgroups. Although the NIHSS gives a reasonable measure of initial severity, it disregards factors such as age that also influence prognosis. The Oxford classification gives an even more crude grading of severity and as a guide to prognosis it is more suitable for epidemiological purposes than for prediction of outcome in individuals.

Although our model-based approach was slightly more powerful than the NIHSS subgroup method used in AbESTT, the absolute advantage was marginal and the simplicity of the AbESTT method may outweigh this advantage. Such an end point will consequently be easier to understand in the context of clinical trials.

We used the placebo patients from the GAIN International trial. The GAIN trial showed no effect of gavestinel. It would be informative also to use data from other sources to validate the results. Our treatment effect is artificially generated. We hope that a successful thrombolytic or neuroprotectant would have an almost immediate effect in limiting infarct extent and thus initial severity, but there can be no guarantee that this would hold true. Only 1 pattern of treatment effect was considered: a fixed effect where all patients were assumed to improve by the same magnitude. Even though there is little knowledge of the actual true effect of most stroke interventions, this is unlikely to be clinically valid; however, our previous work has shown that more complex treatment patterns tend to alter the magnitude but not the direction of the conclusions. Also, other factors can influence the statistical power of a trial: patient selection, sample size, and time to treatment are other possible factors that should be considered. Elsewhere, we have also recently proposed that age and baseline NIHSS should be used together when considering eligibility for acute stroke studies.

We have investigated a range of prognosis-adjusted end points and found that adjusting the cut points for patients depending on prognosis offers analytical power advantages over use of a single fixed end point. Our optimal method of assigning cut points used a prognostic model that considered age and baseline NIHSS, though simply subgrouping according to baseline NIHSS was more straightforward and almost as effective. Prognosis-adjusted end points allow patients to be assessed on realistic achievable goals, while allowing the clinical trial results to be generalizable. Maximizing trial power makes development of treatment for stroke more attainable and less expensive.

**Acknowledgments**

The GAIN trial was sponsored by GlaxoWellcome (now GlaxoSmithKline). GlaxoSmithKline had no involvement in this analysis or article. F.B.Y. is supported by a collaborative studentship from the Medical Research Council and Pfizer. C.J.W. is funded by a Medical Research Council career development fellowship.

---

**TABLE 5. Comparison of End Points in Terms of a Relative Sample Size**

<table>
<thead>
<tr>
<th>Prognosis-Adjusted End Point</th>
<th>Reference End Point</th>
<th>Relative Sample Size*</th>
<th>Reduction in Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI2†</td>
<td>Oxford category‡</td>
<td>351 (339, 363)</td>
<td>49.9% (48.1%, 51.6%)</td>
</tr>
<tr>
<td>RS2†</td>
<td>Oxford category§</td>
<td>602 (582, 623)</td>
<td>14.0% (11.0%, 16.9%)</td>
</tr>
<tr>
<td>NIHSS§</td>
<td>Oxford category§</td>
<td>616 (595, 638)</td>
<td>12.0% (8.8%, 15.0%)</td>
</tr>
</tbody>
</table>

Based on a treatment effect level of 2, table shows the sample size required to maintain statistical power at the level of the reference end point, based on a sample size of 700 patients per treatment group. Parentheses contain 95% CI for the power.

*Compared to a trial based on reference end point with 700 patients per group.
†Both model-based end points used the optimal probability thresholds.
‡The Oxford category end point constituted BI=95 for LACI, PACI, and POCl; BI≤60 for TACI.
§The Oxford category end point contained RS≤1 for LACI and POCl; RS≤2 for PACI and TACI.
||The NIHSS subgroup end point consisted of RS≤0 for NIHSS 4 to 7; RS≤1 for NIHSS 8 to 14; and RS≤2 for NIHSS≥15.
References

1. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. 
2. Rockwood K, Joyce B, and Stolee P. Use of goal attainment scaling in 
   measuring clinically important change in cognitive rehabilitation patients. 
3. Sherman DG, Atkinson RP, Chippendale T, Levin KA, Ng K, Futrell N, 
   Hsu CY, Levy DE; for the STAT participants. Intravenous Ancrod for 
   treatment of acute ischemic stroke. The STAT study: a randomized 
4. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Inter-
   observer agreement for the assessment of handicap in stroke patients. 
5. Brott T, Adams HP, Olinger CP, Marler JR, Barsan WG, Biller J, Spilker 
   Measurements of acute cerebral infarction: a clinical examination scale. 
   Adjusting favourable outcomes following treatment of acute ischemic 
   stroke as influenced by baseline severity of neurological impairments. 
   *Cerebrovasc Dis.* 2003;16. (Abstract.)
7. Berge E, Barer D. Could stroke trials be missing important treatment 
8. Young FB, Lees KR, Weir CJ. Strengthening acute stroke trials through 
9. Bamford J, Sandercock P, Dennis M, Burn J, Warlow C. Classification 
   and natural history of clinically identifiable subtypes of cerebral 
    Orgogozo JM, Whitehead J; for the GAIN International Investigators. 
    Glycine antagonist (gavestinel) in neuroprotection (GAIN International) 
    in patients with acute stroke: a randomised controlled trial. 
12. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application.* 
13. Woodward M. Formulae for sample size, power and minimum detectable 
14. Dorman PJ, Sandercock P. Considerations in the design of clinical trials 
16. Weir CJ, Kaste M, Lees KR. Targeting neuroprotection clinical trials in 
    ischemic stroke patients with potential to benefit from therapy. 
Improving Trial Power Through Use of Prognosis-Adjusted End Points
Fiona B. Young, Kennedy R. Lees and Christopher J. Weir
for the GAIN International Trial Steering Committee and Investigators

Stroke. 2005;36:597-601; originally published online January 20, 2005;
doi: 10.1161/01.STR.0000154856.42135.85
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 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/36/3/597