Should Stroke Trials Adjust Functional Outcome for Baseline Prognostic Factors?

The Optimising the Analysis of Stroke Trials (OAST) Collaboration

Background and Purpose—Many stroke trials have provided neutral results. Suboptimal statistical analyses may be failing to detect effective interventions. Adjusting outcomes for baseline prognostic factors in the analysis may improve the efficiency of analysis of outcomes.

Methods—Data from 23 stroke trials (25,674 patients) assessing functional outcome were included. The prognostic variables considered were age, sex, and baseline severity. Unadjusted and adjusted ordinal logistic regression models were compared using simulated data from each trial (10,000 simulations per trial). Three levels of treatment effect were assessed with ORs of 0.95, 0.74, and 0.57. The reduction in sample size gained from using the adjusted models, as compared with an unadjusted model, was then calculated as a reflection of the increase in statistical power.

Results—Adjusting for baseline factors led to a reduction in sample size, which was similar across all three treatment effects (median percentage reduction, interquartile range): OR = 0.95: 35.3% (21.0 to 42.1); OR = 0.74: 38.4% (29.4 to 42.7); and OR = 0.57: 38.4% (27.4 to 42.2). As the treatment effect increased, the proportion of simulations in which the treatment effect for the adjusted model was greater than for the unadjusted model also increased.

Conclusion—Adjusting for prognostic factors in stroke trials can reduce sample size by at least 20% to 30% (the lower interquartile range) for a given power. Conversely, trialists may want to power for an unadjusted analysis and then increase statistical power by adjusting for prognostic factors.

Key Words: clinical trials ■ statistical analysis ■ stroke

Randomized, controlled trials have greatly improved the care and outcome of patients with acute stroke, although the number of trials carried out and patients included is not reflected in the number of beneficial treatments. By the end of 2001, over 83,000 patients with acute stroke had been included in over 189 trials; the majority of these studies provided neutral results with only alteplase, hemicraniectomy, aspirin, and stroke units showing beneficial effects.2-5 There are many possible reasons for the failure of these trials, including the relevance of preclinical findings to clinical stroke,6 inadequate sample size,1 and the choice of primary outcome and its statistical analysis.7

Results from the Optimising the Analysis of Stroke Trials (OAST) Collaboration have shown that the analysis of stroke trials can be improved by maintaining the inherent ordering of functional outcome using ordinal logistic regression rather than collapsing data into 2 or more groups.7,8 Ordinal logistic regression allows adjustment for covariates; adjustment may further increase statistical power if imbalances in prognostic variables are present and by reducing the variability in the data so that more precise comparisons of treatment can be made.9

When considering an adjusted analysis, the choice of covariates is of prime importance; 3 main methods have been proposed for selecting covariates10: variables that are known to be imbalanced across the treatment groups, prognostic factors that are related to the primary outcome, and a combination of adjusting for those variables that are both related to outcome and imbalanced across treatment groups. Senn suggested that the latter approach may be the most sensible because the reliability of unadjusted tests is affected by both the correlation between the outcome and covariate and the level of imbalance.11 However, accounting for imbalances requires a post hoc decision and therefore is not practical in clinical trials in which models have to be specified in the statistical analysis plan before database closure, lock, and analysis.

Several post hoc studies have examined adjustment for prognostic variables when using functional outcome scales. Reanalysis of data from the National Institute of Neurological Diseases and Stroke trial of alteplase12 using a logistic regression model adjusted for an estimate of prior risk found a 13% reduction in the sample size.13 A study using data from brain injury trials measuring outcome on the Glasgow Outcome Scale found that covariate adjustment led to a 25% reduction in sample size when using logistic regression.14 Other studies have found similar reductions in sample size...
with time to event analyses. However, no studies to date have looked at the effect of adjustment on ordinal logistic regression, which is a more powerful method of analyzing data from functional outcome scales than logistic regression based on dichotomized data. Furthermore, none of these studies discussed the inherent differences between adjusted and unadjusted models. Adjusted models are conditional on the covariates included in the model and therefore interpretation of the results is at the patient level, whereas unadjusted models (which do not account for covariates) have a population-level interpretation.

The aim of this analysis was to assess whether stroke trials using ordinal logistic regression should routinely adjust for important prognostic factors in their primary analyses.

**Methods**

**Optimizing the Analysis of Stroke Trials**

A detailed description of the OAST methodology has been published. In brief, individual patient data were sought from randomized, controlled trials assessing functional outcome after stroke for interventions that were either positive or negative (not neutral) according to the trial publication or for interventions known to be beneficial or harmful from meta-analysis; neutral trials in a neutral meta-analysis were excluded. Trials of thrombolytic agents were not included because their analysis does not benefit from ordinal analyses.

**Trial Data**

Data on demographics (age, sex) and stroke severity (National Institutes of Health Stroke Scale [NIHSS], Orgogozo Stroke Scale, Unified Neurological Stroke Scale, or other similar measures) and functional outcome (Barthel Index [BI], modified Rankin Scale [mRS], “3 question” scale [3Q]—a derivative of mRS) variables were collected for each trial. Scores for functional outcome data used standard definitions for BI (22 levels ranging from no disability [BI/0] through severe disability [BI/5] to death [BI/5]); mRS (7 levels ranging from no disability/dependency [mRS/0] through severe dependency [mRS/5] to death [mRS/6]); and 3Q scale (independent [3Q/4], mild/moderate dependency [3Q/3], severe dependency [3Q/2], and death [3Q/1]). mRS was chosen when more than one scale had been used to assess functional outcome. Although the 22-level BI tends to have sparse data at the center of the scale, we have shown previously that ordinal logistic regression of raw data is a valid approach.

**Statistical Methods**

All analyses were carried out in Stata (Version 8). Statistical significance relates to $P<0.05$. Ordinal logistic regression was used to assess the relationship between each covariate and outcome within each trial. Although statistical testing for baseline imbalances should be discouraged in the primary publication, this was carried out in this study so that the effect of imbalance on adjustment could be assessed. Baseline imbalances between each covariate and treatment were assessed using the $t$ test for age and severity and the $\chi^2$ test for sex.
Figure. A–C, Relationship among age (A, n=9), severity (B, n=6), and sex (C, n=9) and outcome (mRS); data show mean and SD.
Models
Two models were compared; the first model contained treatment assignment only (unadjusted model), whereas the second contained sex, age, and baseline severity as well as treatment assignment (adjusted model). These 3 comprise key demographic and clinical variables and have been shown to influence outcome in stroke\textsuperscript{14} and were the only prognostic variables available for all the included trials.

Simulations
Although some included trials were individually significant on their assessment of functional outcome, others were neutral but included because they tested effective or hazardous treatments (as determined in published meta-analyses). We therefore simulated significant treatment benefits with 3 levels of effect (coefficients of \(-0.56, -0.30, \text{ or } -0.05\) relating to unadjusted ORs of 0.57, 0.74, or 0.95\textsuperscript{14}, respectively). By reference, trials of hemicraniectomy\textsuperscript{21}, thrombolysis\textsuperscript{15}, stroke units\textsuperscript{22}, and aspirin\textsuperscript{23} achieved ORs of 0.24, 0.63, 0.60, and 0.94, respectively. For consistency across studies, BI and 3Q scales were reversed so that higher scores related to a worse state of outcome, like with the mRS; hence, an OR \(<1\) reflects a positive treatment effect across all trials and scales. Simulations were based on the method proposed by Hernandez et al for logistic regression\textsuperscript{18} but extended for outcomes of an ordinal nature by using ordinal logistic regression, which assumes a common OR across the whole scale (proportionality of odds).\textsuperscript{19}

Within each included trial, an ordinal logistic regression model containing age, sex, and severity, with functional outcome as the dependent variable, was used to estimate the probability of having an unfavorable functional outcome for each patient. Patients were then randomly assigned to either the active treatment or control groups (with groups of the same size as the original trial). The specified treatment effect was then added making a better outcome more likely in the active treatment group by a set amount. A new outcome containing age, sex, and severity, with functional outcome as the dependent variable, was used to estimate the probability of having an unfavorable functional outcome for each patient.

Reduction in Sample Size
The reduction in sample size was used to assess the increase in power gained from adjustment. The Z scores from the unadjusted and adjusted models were compared and the reduction in sample size calculated using:

\[\text{Reduction} = 100 - 100 \times \frac{\text{Mean Z score unadjusted}}{\text{Mean Z score adjusted}}\]

**Table 2. The Median (Interquartile Range) ORs Obtained From Unadjusted and Adjusted Models and the Reduction in Sample Size Gained From Using an Adjusted Analysis Compared With an Unadjusted Analysis**

<table>
<thead>
<tr>
<th>Artificial Treatment Effect OR</th>
<th>Unadjusted Model</th>
<th>Adjusted Model</th>
<th>Reduction in Sample Size, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>0.96 (0.96–0.96)</td>
<td>0.95 (0.95–0.95)</td>
<td>35.3 (21.0–42.1)</td>
</tr>
<tr>
<td>0.74</td>
<td>0.79 (0.78–0.80)</td>
<td>0.73 (0.73–0.74)</td>
<td>38.4 (29.4–42.7)</td>
</tr>
<tr>
<td>0.57</td>
<td>0.65 (0.63–0.66)</td>
<td>0.57 (0.56–0.57)</td>
<td>38.4 (27.4–42.2)</td>
</tr>
</tbody>
</table>

**Subgroup Analysis**
Subgroup analyses were performed by assessing the reduction in sample size for differing trial characteristics: functional outcome scale (mRS, BI, 3Q), severity scale (NIHSS, other scale), and trial size (\(\leq 250, >250\)).

**Results**
The present data set compared individual patient data from 23 trials (20 from the original OAST data set\textsuperscript{7} 3 new trials\textsuperscript{21,25,26}) including 25 674 patients. The characteristics of the trials included are given in Table 1. Thirteen trials measured outcome using the BI, 9 used the mRS, and one used the 3Q scale. Fourteen trials measured baseline severity using the NIHSS with others using another measure such as the Orgogozo Stroke Scale. Trial sizes ranged from 32 to 19 435 patients (median, 259).

**Relationship of Covariates With Functional Outcome**
A highly statistically significant \((P<0.0001)\) relationship between severity and functional outcome was found for all trials with greater baseline severity leading to worse functional outcome. The majority of trials (29 of 30) showed a significant relationship between age and outcome, and 6 showed a significant relationship with sex. The Figure shows these relationships graphically in those trials that measured outcome using the mRS.

**Baseline Imbalances in Covariates**
Statistically significant differences in baseline covariates were only seen in 3 of the included trial data sets, one for age (in the Acute Stroke Studies Involving Selfotel Treatment (ASSIST) 07 trial the treatment groups differed by 3.6 years, a difference that has borderline biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance) and 2 for stroke severity (a difference in the trial specific measure of severity of 0.14 points is probably not of biological significance).

**Reduction in Sample Size**
Table 2 shows the median reduction in sample size for the 3 levels of treatment effect. Trial sample size was reduced by 35% to 38% when covariates were introduced and was

**Table 3. Comparison of Z Scores and Treatment Coefficients From the Adjusted Models With Those From the Unadjusted Models**

<table>
<thead>
<tr>
<th>Treatment Effect</th>
<th>0.95</th>
<th>0.74</th>
<th>0.57</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z score, adjusted&gt;unadjusted (%)</td>
<td>52.5 (51.5–53.6)</td>
<td>65.0 (59.8–69.6)</td>
<td>75.8 (67.6–82.7)</td>
</tr>
<tr>
<td>Treatment coefficient, adjusted&gt;unadjusted (%)</td>
<td>52.8 (52.0–54.3)</td>
<td>67.3 (62.6–73.5)</td>
<td>79.2 (71.8–87.6)</td>
</tr>
</tbody>
</table>

*Data given as median percentage and interquartile range.*
Table 4. Reduction in Sample Size (%) for Trials Subgrouped by Type of Severity Scale, Functional Outcome Scale, and Sample Size; Median and Interquartile Range

<table>
<thead>
<tr>
<th>No. of Trials</th>
<th>Treatment Effect (OR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td>Overall</td>
<td>23</td>
</tr>
<tr>
<td>Outcome scale</td>
<td></td>
</tr>
<tr>
<td>mRS</td>
<td>9</td>
</tr>
<tr>
<td>BI</td>
<td>13</td>
</tr>
<tr>
<td>Three questions</td>
<td>1</td>
</tr>
<tr>
<td>Severity scale</td>
<td></td>
</tr>
<tr>
<td>NIHSS</td>
<td>14</td>
</tr>
<tr>
<td>Other scale</td>
<td>9</td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
</tr>
<tr>
<td>&lt;250</td>
<td>11</td>
</tr>
<tr>
<td>≥250</td>
<td>12</td>
</tr>
</tbody>
</table>

Subgroup Analysis

The results from the subgroup analyses are shown in Table 4. The biggest reduction in sample size was seen for trials using the BI (40%) as compared with 21% to 29% for mRS and 20% for 3Q. Trials using the NIHSS as a measure of stroke severity also had a greater reduction in sample size (37% to 39%) than those using other severity scales (29% to 30%). No difference was seen between smaller (<250 patients) and larger (≥250 patients) studies. However, different studies used different measures of severity and outcome and it was not possible to compare directly the relative benefits of using any particular scale.

Discussion

The increasing number and size of stroke trials and failure to identify effective acute treatments are threatening the viability of future studies. Any method that reduces sample size (and hence, the cost and duration of trials) or increases statistical power, thereby improving the likelihood of finding effective interventions, will be welcome. This study shows that the efficiency of analyses of functional outcome in stroke trials is improved when outcome is adjusted for 3 prognostic factors: age, sex, and stroke severity. Such inclusion of covariates allows a substantial reduction in sample size to be achieved, in this case by approximately one fourth (the lower end of the interquartile range), for a given power; conversely, statistical power can be increased for a given sample size. Maintaining sample size and increasing the statistical power has the added benefit of improving the robustness of subgroup analyses. Importantly, covariate adjustment appeared to be effective irrespective of the scales used to measure baseline severity and functional outcome. (Of note, the design of stroke trials and calculation of sample size has been addressed in an earlier OAST study.)

Other studies have shown that adjustment for baseline covariates improves statistical power. International Mission on Prognosis and Clinical Trial design in TBI (IMPACT) assessed ways of improving the design and analysis of brain injury trials and found that adjustment for 7 predictors of outcome reduced sample size by approximately 16% to 23% when analyzed using logistic regression on a dichotomized Glasgow Outcome Scale. Similar results have been reported for time to event analyses using the Cox proportional hazards model. However, this OAST analysis is the only study looking at the effect of adjustment on ordinal logistic regression and assessment of potential benefits on sample size.

Adjustment addresses imbalances in baseline prognostic factors, which occur by chance with simple randomization. Historically, the interpretation of several stroke trials has been confounded by imbalances at baseline. For example, the 20 000 patient International Stroke Trial was neutral in its primary univariate analysis but positive after adjustment with a model predictive of outcome. Similarly, the Stroke-Acute Ischemic NXY Treatment I (SAINT-I) trial was positive when adjusted for prognostic factors but neutral when analyzed without covariate adjustment (unpublished data). Such imbalances in baseline factors may be reduced using stratification or adaptive randomization (minimization); the latter technique also moderately improves statistical power.

Adjustment for covariates increases the precision of the estimated treatment effect and changes the interpretation of the results, because these are now conditional on the chosen covariates. It is therefore crucial that adjustment is considered at the protocol development stage of setting up a clinical trial and that the covariates are chosen and stated a priori; the decision to include covariates, and which ones, at the time of analysis would be incorrect and results in misleading data-driven analyses.

There are several limitations to the present analysis. First, only 20 of the original 55 OAST data sets could be used
because many studies did not share baseline data. Although this is unlikely to have changed the quantitative findings, it will have reduced the power of the analyses. In this respect, it is vitally important that trialists, both academic and commercial, share data after publication of the main trial paper for use in other projects (such as OAST and Virtual International Stroke Trials Archive [VISTA]\(^\text{29}\)) so that its value is maximized. Second, only 3 covariates (age, sex, severity) were used so as to maximize the number of included data sets. However, this limitation is not important because, although there are many baseline characteristics that have prognostic significance (eg, atrial fibrillation, temperature, blood pressure, and serum glucose), severity has been shown consistently to be the most powerful predictive factor and explains most of the variation in covariate-adjusted analyses (as shown here). Third, beneficial effects on study power/sample size may not translate to other clinical areas; stroke is unusual in having such a strong predictor of outcome in the form of baseline severity and, as such, the reduction in sample size gained by adjusting for covariates will be greatly influenced by the strength of the relationship between severity and outcome. Fourth, methods of analysis that assess shifts in outcome over the entire distribution, although popular with statisticians.\(^30\) Additionally, further work needs to address what magnitude of shift in outcome is meaningful to patients, healthcare professionals, and health funders. Last, we have assumed no interaction between the treatment and covariates, but interactions should be considered when carrying out an adjusted analysis.

In summary, trialists should consider using key prognostic variables in the analysis of functional outcome in stroke trials when using ordinal analyses. This will allow trials to be smaller for a given statistical power or to achieve greater statistical power for a given sample size. Nevertheless, existing knowledge that covariate-adjusted logistic regression is more powerful than unadjusted analyses has not led to all trials moving to this approach, perhaps because of uncertainty about the interpretation and presentation of trial results based on adjusted analyses. Hence, in practical terms, trialists may, at least in the short term, want to power their study for an unadjusted analysis and then analyze the completed trial with adjustment for covariates, thereby increasing the statistical power but maintaining a large enough sample size to carry out an unadjusted analysis as a secondary end point. Nevertheless, the results need to be reported in the context of the included covariates. If trialists prefer to power their study for an adjusted analysis, then the sample size will need to be estimated by using a simulation study such as this one. If pilot data are available, they could be used to estimate the reduction in sample size with adjustment for particular covariates.

Acknowledgments

We thank Stuart Pocock (London, UK) and James Carpenter (London, UK) for their statistical advice and Scott Hamilton (US) for commenting on the draft of the paper. We thank the patients who took part in these studies and the trialists who shared their data. The study was conceived, initiated, managed, analyzed, and interpreted independently of any pharmaceutical company. Each collaborator commented on the manuscript.

The following collaborators contributed individual patient data from their trial and commented on the manuscript: Abciximab: H. Adams (US) and E. Barnathan (US), W. Hacke (Germany); ASSIST 07 and 10: S. Davis (Australia); Citicoline 1, 7, 10, 18: A. Davalos (Spain); DCLHb: P. Koudstaal and R. Saxena (The Netherlands); DESTINY: E. Juettler and W. Hacke (Germany); Dover Stroke Unit: P. Langhome (UK); Ebselen: T. Yamaguchi (Japan); FOOD 3: M. Dennis (UK); INWEST: N.-G. Wahlgren and N. Ahmed (Sweden); IST: P. Sandercock and S. Lewis (UK); MAST-I: L. Candelise (Italy); J. Wardlaw (UK); Minocycline: Y. Lampl and M. Boaz (Israel); RANNTAS I and II, STIPAS, TESS I and II: P. Bath (UK); Statin withdrawal: J. Castillo (Spain).

Writing Committee: Laura J. Gray (lead statistician, Nottingham, UK); Philip M. W. Bath (Chief Investigator, Nottingham, UK); and Timothy Collier (Statistical Advisor, London, UK).

Sources of Funding

L.J.G. is funded, in part, by the UK Medical Research Council, P.B. is Stroke Association Professor of Stroke Medicine.

Disclosures

None.

References

Should Stroke Trials Adjust Functional Outcome for Baseline Prognostic Factors? 

The Optimising the Analysis of Stroke Trials (OAST) Collaboration

*Stroke*. 2009;40:888-894; originally published online January 22, 2009; 
doi: 10.1161/STROKEAHA.108.519207

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 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/40/3/888](http://stroke.ahajournals.org/content/40/3/888)