Heterogeneity of Stroke Pathophysiology and Neuroprotective Clinical Trial Design

Keith W. Muir, MD, MRCP

Background and Purpose—Tissue substrates for action of neuroprotective agents may be absent in a significant proportion of strokes. Pathophysiological heterogeneity is a possible contributor to negative neuroprotective trials.

Methods—Stroke subtypes and their individual outcomes in neuroprotective trial control populations were used to derive models incorporating accuracy of clinical classification and probability of an ischemic penumbra. With the use of treatment effect sizes from successful trials (predominantly of reperfusion therapies), sample sizes for neuroprotective trials were calculated. The potential influence of altered recruitment strategies was explored.

Results—The proportion of informative patients in 2 large neuroprotective trials was probably only 27% to 30%. Optimistically, this proportion may be 50%; pessimistically, it may be only 17%. These figures necessitate a sample size of 3700 to 4500 subjects per group; at best, 1800 to 2200 are needed per group with optimistic assumptions about treatment effect. Strategies to enhance the proportion with tissue substrate for neuroprotection could reduce sample size to 500 per group and simultaneously reduce the total number of patients screened compared with inclusive trials.

Conclusions—Population heterogeneity alone may be sufficient to explain negative neuroprotective trials because even in the largest trials to date sample size is inadequate to detect effect size equivalent to those with thrombolysis, and it is possible that they have been severely underpowered. Reliable trials with inclusive entry criteria may be too large to be commercially feasible for novel compounds. Both sample size and total number of patients needing to be screened should be reduced by restricting entry to patients more likely to have a tissue target. (Stroke. 2002;33:1545-1550.)

Key Words: controlled clinical trials ■ neuroprotection

Clinical trials of neuroprotective therapies have to date been uniformly negative. The failure of trials to confirm clinical benefits implied by infarct volume reduction in animal models may result from pharmacological factors (irrelevance of the pharmacological target to human ischemia), clinical pharmacological factors (insufficient dose, inadequate treatment duration, or unfavorable pharmacokinetics), or inadequacies of trial design. Deficiencies of pharmacology and clinical pharmacology in neuroprotective drug development have undoubtedly contributed to the lack of trial success, and possible remedies have been considered.¹ Potential problems with the statistical power of neuroprotective trials have been identified,²,³ but the impact of stroke heterogeneity on trial design has not been assessed.

Pathophysiological heterogeneity is of particular relevance to neuroprotection. There are 2 principal groups of patients in whom neuroprotective therapies are unlikely to be effective: (1) those lacking a biological substrate relevant to the mode of action of the drug—eg, N-methyl-D-aspartate antagonists influence neuronal cell body survival but probably have no effect on white-matter injury,⁴ and excess glutamate release may play a role in cortical infarction but not lacunar stroke⁵—and (2) those lacking an ischemic penumbra, which seems likely to be of restricted volume in most patients but with wide interindividual variation⁶ and is probably absent in some stroke types such as intracerebral hemorrhage⁷ and lacunes. It is possible that a third group, those who do not reperfuse, also lacks a biological substrate for neuroprotection. By concentrating on early histological outcome, animal studies have left uncertainty over whether many neuroprotective drugs effect lasting reduction in infarct volume, especially in the absence of reperfusion, or simply prolong viability until delayed reperfusion occurs.

This article explores the potential influence of the inclusion of patients lacking a tissue target for neuroprotective drugs in clinical trials, with implications for trial design.

Methods

Models exploring the influence of stroke heterogeneity on sample size requirements and the magnitude of treatment effects were developed through the use of data derived from published trials. The possible impact of changed recruitment strategies was also explored. Assumptions are detailed below.

Tissue Targets

Imaging data do not support the existence of a penumbra around a primary intracerebral hematoma (PICH).⁷ There is also no biological...
Clinical Misclassification

Imaging studies emphasize the limited sensitivity of clinical diagnosis in the acute phase of stroke, particularly in distinguishing lacunar from partial middle cerebral artery (MCA) syndromes. Models therefore assume that a proportion of clinically diagnosed lacunar syndromes will in fact result from partial MCA occlusion and vice versa. Similar misclassification of large MCA syndromes that are partially occluded may spontaneously recanalize. Strategy 4 may restrict recruitment to 5% of patients, but up to 30% of patients may be misclassified and a further 23% of patients from 6-hour window trials (Figure 1). Strategy 2 restricts recruitment to about one third of current trial populations, 20% of patients from 6-hour window trials, and 10% to 12% of patients with confirmed proximal MCA occlusion were derived from the PROACT II trial.

Alternative Recruitment Strategies

The effect of 4 different strategies was explored: (1) eliminating PICH by pretreatment CT, (2) restricting recruitment to complete MCA syndromes (TACS), (3) restricting recruitment to patients with DWI-PWI mismatch, and (4) restricting recruitment to patients with MCA occlusion confirmed on imaging. Strategy 1 removes 10% to 12% of patients from 6-hour window trials (Figure 1). Strategy 2 restricts recruitment to about one third of current trial populations, but up to 30% of patients may be misclassified and a further 23% may spontaneously recanalize. Strategy 3 restricts recruitment to 20% to 56% of patients with anterior circulation stroke confirmed to involve the MCA territory, probably 70% of those clinically diagnosed as such. Strategy 4 may restrict recruitment to 5% of patients if conventional angiography forms the basis for diagnosis but may include 40% to 60% of anterior circulation strokes (about two thirds of current trial populations) if alternative techniques such as MR angiography, CT angiography, and transcranial Doppler ultrasound are included. MCA proximal occlusion correlates closely with the presence of PWI-DWI mismatch and lesion volume expansion. Alternative recruitment strategies were explored

**TABLE 1. Models**

<table>
<thead>
<tr>
<th></th>
<th>Misclassified, %</th>
<th>Penumbra, %</th>
<th>Proportion of Syndrome With Tissue Target, %</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacunar syndromes</td>
<td>30</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>PICH</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partial MCA syndromes</td>
<td>30</td>
<td>70</td>
<td>49</td>
</tr>
<tr>
<td>Total MCA syndromes</td>
<td>10</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td><strong>Realistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacunar syndromes</td>
<td>20</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>PICH</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partial MCA syndromes</td>
<td>40</td>
<td>30</td>
<td>18</td>
</tr>
<tr>
<td>Total MCA syndromes</td>
<td>20</td>
<td>70</td>
<td>62</td>
</tr>
<tr>
<td><strong>Pessimistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacunar syndromes</td>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>PICH</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partial MCA syndromes</td>
<td>50</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Total MCA syndromes</td>
<td>30</td>
<td>50</td>
<td>41</td>
</tr>
</tbody>
</table>

Misclassification assumes that lacunar syndromes have partial MCA syndromes and vice versa and that total MCA syndromes are always misclassified as partial MCA syndromes.

**TABLE 2. Event Rates and Treatment Effect Sizes in Recent Clinical Trials**

<table>
<thead>
<tr>
<th></th>
<th>NINDS, † %</th>
<th>PROACT II, %</th>
<th>STAT, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treated</td>
<td>Control</td>
</tr>
<tr>
<td>Event rate BI &lt;55-65</td>
<td>39*</td>
<td>34*</td>
<td>53</td>
</tr>
<tr>
<td>RRR (95% CI)</td>
<td>13 (8–20)</td>
<td>13 (7–25)</td>
<td>15 (10–21)</td>
</tr>
<tr>
<td>Event rate BI &lt;95</td>
<td>62</td>
<td>50</td>
<td>68</td>
</tr>
<tr>
<td>RRR (95% CI)</td>
<td>22 (16–30)</td>
<td>13 (7–26)</td>
<td>9 (6–13)</td>
</tr>
</tbody>
</table>

* BI <55; † BI <65; ‡ NINDS part 2 data only have been used for BI <55, both parts 1 and 2 for BI <95.
through the use of the realistic model estimates for tissue targets and event rates modified to correspond to the scenarios described. The sample size for different RRRs was calculated for each, as was the proportion of screened patients likely to be eligible for each scenario using the study flow data from PROACT II (Figure 2).

Two large neuroprotective clinical trial populations were considered in each of the model situations: those in CLASS and in the Glycine Antagonist in Neuroprotection–International (GAIN-I) trial. The GAIN trial reported proportions of patients with intracerebral hemorrhage (18.6%) and lacunar stroke (18%) but did not specify cortical stroke subtypes by OCSP; therefore, 30% have been assumed to be PACS and 33% to be TACS.

### Results

The CLASS and GAIN-I trial populations are described in Table 3. In each trial, the proportion of potentially uninformative patients (those with no tissue target) based on the model assumptions is given. For optimistic, realistic, and pessimistic models, the proportions were 0.46, 0.36, and 0.21, respectively, in CLASS and 0.41, 0.32, and 0.19, respectively, in GAIN-I. The sample size estimates for different effect sizes are given in Table 4. The total number of patients enrolled in CLASS was 1360; the 2 GAIN trials enrolled 3391. The models indicate that GAIN at best may have been only 77% of the necessary size and CLASS at best 38%; at worst, these trials may have been only 4% and 2% of necessary size, respectively, and for the realistic model and RRR of 15%, still only 22% and 11%, respectively.

The influence of the proportion of informative patients in a trial on sample size is shown in Figure 3, with calculations for different effect sizes.

Alternative entry criteria limit eligibility and restrict recruitment to varying proportions of those screened (Table 5). Judging from PROACT II screening figures, up to 30% of patients may be eligible for a 6-hour window neuroprotective trial, but restriction to patients with DWI-PWI mismatch on MRI or MCA occlusion reduces this to 5% to 7% (Figure 2). The 30% figure represents a very optimistic result in the experience of most stroke centers. The number of patients needing to be screened for each patient eligible is therefore 3 for current inclusive entry criteria and increases to 15 to 18 for screening on the basis of imaging correlates of a penumbra. However, because the proportion of informative patients is higher for imaging-based trials, fewer patients must be screened for this type of trial than for conventional trials: ≈12 600 using conventional inclusion criteria and 20% RRR compared with 5100 for DWI-PWI mismatch.

When the proportions of patients with BI <95 were used to define poor outcome (using event rates published for CLASS PICH and TACS patients and extrapolations for LACS [35% poor outcome] and PACS [60%]), there was a reduction in sample size requirements of ≈40%. Further exploration of the possible influence of different end-point dichotomies is shown in Figure 4. Because outcome event rates are similar for different end points (For example, in the placebo group in GAIN-I, 34% had BI >90, 28% had modified Rankin Scale score 2, and 26% had National Institutes of Health Stroke Scale score <2), there were no substantial changes in the results for different scales.

### Discussion

Inclusion of patients not affected by a treatment may have important implications for the outcome of clinical trials, influencing estimates of benefit and potentially rendering trials negative. It has been argued that stroke trials to date have suffered from overoptimism, with assumptions of large effect sizes leading to an all-inclusive approach in phase III trials that ignores fundamental differences in stroke mechanism. The PROACT II trial results supported the view that smaller trials in clearly defined and homogeneous stroke pathologies can yield answers that may be obscured by dilution in larger groups of patients who are not as well characterized.
Most neuroprotective trials to date have been powered to
detect effects on the order of an absolute risk reduction of
10%. This strategy reflected a combination of anticipation
that large reductions in histological infarct volume in animal
models would translate into large functional effects in hu-
mans, extrapolation from small phase II trials that risk biasing
effect size estimates upwards,2 and limited recruitment rates
resulting from the inherent structural inadequacies of stroke
care organization, as well as commercial considerations. The
only data that permit estimation of the likely effect size of an
efficacious treatment are derived from reperfusion therapies:
the NINDS rtPA trial, PROACT II, and STAT. Event rates in
the control groups of these trials are comparable to those in large
neuroprotective studies, and the entry criteria are comparable,
except for excluding hemorrhage by CT. Realistic estimates of
potential treatment effect indicate an RRR of 10% to 20% for
detrimental outcome (death and dependence), with an upper
95% CI limit of 24% for the 2 thrombolytic trials.

Assuming neuroprotection to have a more restricted target
population than reperfusion therapies for the reasons outlined
above, the sample size estimates derived from the models in
this study indicate a requirement for \( \approx 4000 \) patients per
treatment arm to detect even large treatment effects (RRR of
20%) using current trial entry criteria and a typical trial end
point. Even optimistic assumptions about treatment effect,
accuracy of initial diagnosis, and penumbra yield a sample
size of 1800 to 2200 per group. To detect a more modest
treatment effect (RRR of 10%), the minimum sample size
exceeds 5000 per group; more patients may need to be
recruited than have taken part in all trials ever undertaken of
calcium antagonists (\( \approx 7500 \)) or glutamate antagonists
(\( \approx 11000 \)). At best, even the GAIN trial program may have
recruited only three quarters of the necessary subjects. Neg-
ative results from trials to date may thus potentially be
explained entirely on the basis of stroke heterogeneity.

There may be important statistical advantages in using a
definition that increases the proportion of patients categorized
as having poor outcome, with reduction of sample size of
\( \approx 40\% \) if BI < 95 is used as an end point (Figure 4). Although
this implies that GAIN-I may have had adequate sample size
to detect a large treatment effect (RRR of 20%) in the realistic
model, it would still be only 75% of the required size for a
more modest 15% RRR, and CLASS remains too small to
detect anything other than a large effect with optimistic
assumptions. It remains unclear whether full or nearly full
recovery is an appropriate end point for a neuroprotective
treatment as opposed to reperfusion therapy. Although this
end point would be valid if the assumption of proportional
odds reduction is correct (ie, patients move across all cate-
gories of outcome in the same proportion in response to
treatment), this remains unproven for neuroprotective agents,
and there are biological reasons for uncertainty.

The proportion of informative patients should be enhanced
by restricting trial time windows further (perhaps to 3 hours),
excluding PICH with CT, instituting mandatory clinical
criteria such as minimum severity on stroke scales or the
presence of cortical features, and using MRI signatures of a
penumbra. All of these strategies restrict recruitment rates,
but sample size may to be reduced to 400 per treatment arm
to show benefit using conventional outcome measures. De-
TABLE 5. Proportion of Patients Eligible for Different Entry Criteria Models, Sample Size, and Number of Patients Needed to Screen for Each Informative Patient

<table>
<thead>
<tr>
<th>Entry Criteria</th>
<th>Proportion Eligible, %</th>
<th>Sample Size, n per group</th>
<th>Needed to Screen, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS</td>
<td>29</td>
<td>13 383</td>
<td>3680</td>
</tr>
<tr>
<td>CT</td>
<td>27</td>
<td>11 571</td>
<td>3181</td>
</tr>
<tr>
<td>TACS</td>
<td>11</td>
<td>2375</td>
<td>600</td>
</tr>
<tr>
<td>DWI-PWI mismatch</td>
<td>7</td>
<td>1390</td>
<td>348</td>
</tr>
<tr>
<td>MCA occlusion</td>
<td>5</td>
<td>1862</td>
<td>464</td>
</tr>
</tbody>
</table>

Despite needing to screen ≈18 patients per eligible subject using MRI criteria (6 times as many as with conventional criteria), the total number needing to be screened for a trial should in fact be far fewer than for inclusive entry criteria. Costs may be increased through the need to screen more ineligible patients with expensive imaging and the likely need to expand the number of participating centers and lengthen study duration. Although there are advantages in conducting very large, inclusive trials with appropriate subgroup analyses rather than very restricted trials, a more restrictive trial model may be more cost effective in early drug development.

There are inevitable caveats about the results of this modeling exercise. First, the assumption of a lack of effect in hemorrhage and lacunar syndromes is based on limited data and may be overly pessimistic. Reperfusion appears beneficial in patients classified as having small-vessel disease despite similar mechanistic arguments against it, and ultimately there remains uncertainty about the pathological basis for many lacunar infarcts. All MRI data are based on very small numbers of patients and almost certainly reflect a biased sample. Second, the numbers of patients with MRI evidence of penumbra may be less than assumed. A further concern is the restricted volume of the penumbra in MRI and PET studies25,26; if anatomically very limited, then even major salvage of target tissue may be without clinically relevant benefit. Third, the clinical classification and scales used are subject to error, and the BI in particular has limited discriminatory utility as an outcome measure. The OCSP classification, although not designed for acute use, categorizes patients in a biologically relevant manner by distinguishing large from small cortical strokes and lacunes, something not possible from total stroke scale scores, for instance. There are no published data with alternative categories, but the same modeling exercise could, for example, be applied on the basis of imaging findings. With respect to outcome, almost all trials have used a responder analysis based on the proportions in favorable versus unfavorable outcome categories, and the results from this model using the BI can be extrapolated to any other dichotomous outcome measure, including handicap scales such as the modified Rankin Scale or imaging-based biomarkers (eg, proportion with expansion of MRI lesion volume27) with little change in the implications for sample size, although there are some advantages in choosing a smaller rather than larger outcome category, as shown in Figure 4. Finally, event rates derived from reperfusion trials may be too conservative because the net RRR reflects an aggregate of very favorable outcomes in patients who reperfuse early and less favorable outcomes in those who do not or those who encounter hemorrhagic complications. Reperfusion trials may in fact have a proportion of uninformative patients similar to that of neuroprotective trials. However, most patients in the trials used to estimate effect size were randomized within 3 hours, and effect sizes are likely to be greater than achievable with neuroprotectives, typically delivered 4 to 6 hours after onset. There was no significant heterogeneity between stroke subtypes with respect to benefit. Most importantly, had the true effect size of any of half a dozen neuroprotective agents been as great as or greater than that evident in reperfusion trials, it should have been seen in the trials, many of which had sample sizes many times greater than NINDS. In the absence of any data from neuroprotective trials themselves, effect size estimates represent the best information from which to extrapolate.

In conclusion, heterogeneity of stroke populations recruited to typical neuroprotective trials may reduce substantially the likelihood of showing efficacy, even if a neuroprotective agent were to have an effect size equivalent to thrombolitics. Logical and feasible strategies to limit the variability in the patient populations recruited may enhance the ability of future trials to demonstrate efficacy.

Acknowledgment

Dr Muir is supported by Chest, Heart and Stroke, Scotland.

References


Heterogeneity of Stroke Pathophysiology and Neuroprotective Clinical Trial Design
Keith W. Muir

Stroke. 2002;33:1545-1550
doi: 10.1161/01.STR.000018648.86293.AB
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 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/33/6/1545