Methodological Quality of Animal Studies of Neuroprotective Agents Currently in Phase II/III Acute Ischemic Stroke Trials

Maria Philip, BS; Michael Benatar, MD; Marc Fisher, MD; Sean I. Savitz, MD

Background and Purpose—Numerous neuroprotective agents have proven effective in animal stroke studies, but every drug has failed to achieve its primary outcome when brought forward to clinical trials. We analyzed the quality and adequacy of animal studies supporting the efficacy of NXY-059 and other neuroprotective agents that are currently being investigated in phase II/III trials.

Methods—We conducted a systematic search of all neuroprotective drugs in Phase II or III trials and collected data from animal studies of focal cerebral ischemia testing agents systemically administered within 24 hours of occlusion. The methodological rigor of each individual study was evaluated using 5 criteria derived from the STAIR guidelines. The adequacy of the preclinical “package” for each drug was then evaluated by combining the results of all studies for each drug to determine which of a further 5 STAIR criteria were met before moving forward from animal to human studies.

Results—Our search yielded 13 agents of which 10 had published data in peer-reviewed journals. There is substantial within-drug variability in the quality of preclinical studies as well as substantial variation in the completeness of the collective preclinical literature for different drugs. There has been little or no improvement in the quality of animal studies since NXY-059, and current agents have not been subjected to a more complete preclinical evaluation.

Conclusion—There is significant heterogeneity in the quality of animal testing for neuroprotective agents in stroke. Drugs in the post-SAINT era have not been subjected to more thorough preclinical evaluation. (Stroke. 2009;40:577-581.)

Key Words: focal cerebral ischemia ■ neuroprotection ■ quality ■ animal model

Numerous neuroprotective agents have proven effective in animal stroke models, but all have failed to improve outcome in human phase III trials. The collective experience of numerous failed neuroprotective trials has called into question whether neuroprotection represents a viable therapeutic strategy for acute ischemic stroke.1 The latest trial, SAINT II, in which patients treated with NXY-059 achieved nearly identical outcomes compared with patients who had received placebo,2 has led some to believe that neuroprotection should be abandoned as a treatment for acute stroke.3 There are several potential reasons for the failure of neuroprotective agents to improve outcome in phase III clinical trials. Two important reasons are the high risk of bias and the failure to adequately control for physiological variables. The concern is that these methodological limitations yield misleading information regarding the potential efficacy of therapeutic agents.4 To explore this possibility, we analyzed the methodological quality of individual studies as well as the adequacy of the overall preclinical package of published animal studies supporting the efficacy of NXY-059 and other neuroprotective agents currently being investigated in phase II/III trials.

Because the STAIR publication5 considers study methodology with respect to both the risk of bias and the extent to which relevant physiological variables are controlled, we relied on these recommendations5 to investigate whether NXY-059 or those agents that have since come forward to clinical trial have met rigorous standards for preclinical testing.

Methods

We conducted a systematic search of all neuroprotective drugs in Phase II/III using Medline, Clinicaltrials.gov, other internet engines, and published review articles. The key word searches included “neuroprotective agents,” “brain ischemia,” “stroke,” and “animal models.” We also conducted key word searches using the names of all individual compounds identified in the initial search and selected for inclusion all studies that had tested these agents in focal cerebral ischemia. We reviewed the reference lists of all relevant studies that were identified and all review articles on neuroprotective drugs developed for ischemic stroke. Articles selected for inclusion in this study must have collected data from animal focal cerebral ischemia studies that tested agents systemically administered within 24 hours of arterial occlusion. Neonatal models were excluded. In addition, studies that focused on enhancing recovery from stroke were not included in this analysis. Models involving endothelin-1 were also...
excluded as it causes vasospasm and is unreliable in causing consistent damage. Awake animals were included in the analysis.

Our first aim was to examine the methodological quality of individual studies using 5 criteria derived from the STAIR guidelines for preclinical evaluation of stroke therapeutics. We used the term quality to reflect the extent to which the animal testing was rigorous, which was based on the following criteria: (1) whether or not physiological variables (most importantly, cerebral blood flow) were measured to document vascular occlusion as an index of the reliability of the ischemia model; (2) whether or not animals were randomly assigned to study treatments; (3) whether or not investigators were blinded to treatment administration; (4) whether or not investigators were blinded to treatments during outcome assessment; and (5) whether or not temperature was controlled during the ischemic period. Each study was assigned a score from 0 to 5 based on the number of methodological criteria met. Median scores (and interquartile ranges) were then calculated for those drugs which had been tested in multiple studies.

We then examined the collective literature for each individual drug to determine the range of experiments that were performed. Sufficiency of the preclinical literature for each drug was evaluated using a set of criteria also derived from the STAIR guidelines. These criteria required (1) that the drug should have been tested in both transient and permanent occlusion models, (2) reproducibility, with efficacy demonstrable in at least 2 independent laboratories, (3) evidence for efficacy based both on histological and behavioral outcome measures, (4) characterization of a therapeutic time window relative to the time of onset of the ischemic injury during which the drug is effective, and (5) evidence for efficacy in at least 2 species, 1 of which is a cat or primate. STAIR 1\textsuperscript{1} also recommended that data should be published in a peer-reviewed journal, but for the purposes of this article, only peer-reviewed data were evaluated and therefore this guideline was not included as a criteria.

For both sets of criteria, a methodological feature was regarded as absent if it was not reported. We chose the STAIR guidelines because they represent the first coherent attempt to provide guidance on the complete preclinical package of stroke therapeutics that should be performed before advancing to clinical trials.

Results

Summary Data
Our search yielded 54 studies involving 10 different therapeutic agents.\textsuperscript{6–59} Studies examining 3 additional agents, DP-b99, SUN N4057, and S18986, were excluded because the results have not been published in peer-reviewed journals. The heterogeneity of the data are illustrated by the fact that these 54 studies used 8 different models for inducing focal ischemia in 15 different species; 6 different histological stains were used to measure infarct size, and 19 different behavioral tests were used to evaluate outcome in nonprimate animals; 17 percent of the studies administered drugs before ischemia.

Methodological Quality
There is substantial within-drug and between-drug variability in the methodological quality of the published studies (Figure). The median quality score for an individual drug ranged from 1 to 4. IFN-β and erythropoietin had the lowest scores of 1, NXY-059 had a score of 2, whereas G colony-stimulating factor (CSF) had the highest score of 4, of a possible maximum score of 5. Less than 30% of all studies reported monitoring CBF, 10% of studies reported whether the investigator was blinded during treatment administration, but temperature was controlled during the experimental period in 87% of studies (Table 1).

Sufficiency of Preclinical Data
There is insufficient preclinical data for all of the drugs currently in phase II/III trials (Table 2). All agents since NXY-059 have not fulfilled at least 1 of the 5 STAIR criteria. Two drugs have only met 1 of these criteria. A single agent, traxoprodil, has been studied in a species higher than rodents (eg, cat or primate) suggested by STAIR for novel first-in-class drugs.

Discussion
Historically, a large number of purported stroke therapeutics have undergone animal testing before advancing to clinical trials. However, drugs that have proven effective in animals have not proven effective in humans. One of several possible explanations for the failure is that animal studies have been of insufficient methodological quality with the result that limited conclusions can be drawn from the data. We have systematically evaluated the methodological quality of animal studies of NXY-059 and other neuroprotective agents currently under clinical investigation and found them to be almost uniformly wanting.

The data from this study indicate that NXY-059 achieved one of the lowest scores on our quality metric compared with other agents currently undergoing clinical testing. Only 9% of the studies reported monitoring CBF and only 45% of the studies reported blinding during outcome assessment. O’Collins et al\textsuperscript{66} using their own set of guidelines concluded that NXY-059 nearly achieved all their quality criteria (9/10) but subsequently one of the authors provided a reanalysis in a subsequent publication (reporting that the drug achieved only 4.5/10 criteria).\textsuperscript{61} The current study provides additional evidence that although NXY-059 met the preclinical criteria on models, species, outcomes, reproducibility, and therapeutic window (Table 2), the individual studies were of limited methodological quality (Figure; Table 1).

Have the neuroprotective agents currently in clinical studies since NXY-059 undergone more rigorous preclinical evaluation? Although the methodological quality of individual studies for different drugs was quite variable, median quality scores were generally low. None of the drugs since NXY-059 have fully met the individual preclinical criteria.
The overall quality scores were not appreciably higher than NXY-059 with the possible exception of G-CSF. This would suggest that the recommendations of STAIR have had little impact on which drugs are being selected for testing in clinical trials.

Among the quality criteria not being met, the lack of CBF monitoring calls into question the reliability of the model to induce ischemia and reproducibly cause infarction. Sustained CBF reduction has been shown to reliably predict infarction. There was also a low incidence of reporting whether animals were randomized to treatment groups and whether investigators were blinded during outcome assessment. Blinding is important especially when the adjudication of the outcome measure is somewhat subjective in which case the lack of blinding is an important source of potential bias. In addition, very few studies reported blinding of investigators to study drug versus placebo during treatment administration, a criterion essential to phase III double blind human trials. Finally, only a single drug since the NXY-059 studies has been tested in a higher species.

There are many reasons why studies may not be fully meeting STAIR guidelines. Insufficient funds may be available to complete the preclinical package. The costs of primate testing, for example, have become prohibitive. Not all investigators agree that it is necessary to fulfill the STAIR recommendations before advancing to clinical trials. STAIR is certainly not a recipe or guarantee that an agent will be effective in clinical trial. It is unknown which, if any, of these criteria are critical for the success of a stroke therapeutic.

Our analysis has important limitations. First, the search strategy may have missed studies and did not include unpublished work. As a result, almost all of the studies we analyzed reported positive results. Second, our assessment of methodological quality focused on the risk of bias and whether or not individual animal studies adequately considered important physiological variables. This assessment was based on the STAIR recommendations, which are not uniformly accepted as the gold standard by investigators in the field. However, it represents the only published consensus document that has yet to be developed. Third, we recognize that the weighted

**Table 1. Methodological Quality of Individual Studies**

<table>
<thead>
<tr>
<th>Methodological Features Contributing to 5-Point Quality Metric</th>
<th>NXY059</th>
<th>Caffeinol</th>
<th>Albumin</th>
<th>EPO</th>
<th>GCSF</th>
<th>Magnesium</th>
<th>ONO2506</th>
<th>IFN-B</th>
<th>Traxoprodil</th>
<th>Minocycline</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Studies</td>
<td>11</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>Randomized, %</td>
<td>45</td>
<td>50</td>
<td>25</td>
<td>0</td>
<td>100</td>
<td>82</td>
<td>100</td>
<td>33</td>
<td>33</td>
<td>14</td>
<td>54</td>
</tr>
<tr>
<td>Blinded Assessment of Outcome %</td>
<td>45</td>
<td>75</td>
<td>100</td>
<td>0</td>
<td>88</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>53</td>
</tr>
<tr>
<td>Blinded Administration of Drug %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>5</td>
<td>0</td>
<td>33</td>
<td>50</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Reported CBF,%</td>
<td>9</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>26</td>
</tr>
<tr>
<td>Temperature, %</td>
<td>64</td>
<td>75</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>91</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>87</td>
</tr>
<tr>
<td>Pre Tx, %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>73</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>17</td>
</tr>
</tbody>
</table>

Data show the percent of studies (for each neuroprotective drug) meeting each quality measure.

*Percent of the total studies that administered drug before onset of ischemia.

The overall quality scores were not appreciably higher than NXY-059 with the possible exception of G-CSF. This would suggest that the recommendations of STAIR have had little impact on which drugs are being selected for testing in clinical trials.

Among the quality criteria not being met, the lack of CBF monitoring calls into question the reliability of the model to induce ischemia and reproducibly cause infarction. Sustained CBF reduction has been shown to reliably predict infarction. There was also a low incidence of reporting whether animals were randomized to treatment groups and whether investigators were blinded during outcome assessment. Blinding is important especially when the adjudication of the outcome measure is somewhat subjective in which case the lack of blinding is an important source of potential bias. In addition, very few studies reported blinding of investigators to study drug versus placebo during treatment administration, a criterion essential to phase III double blind human trials. Finally, only a single drug since the NXY-059 studies has been tested in a higher species.

There are many reasons why studies may not be fully meeting STAIR guidelines. Insufficient funds may be available to complete the preclinical package. The costs of primate testing, for example, have become prohibitive. Not all investigators agree that it is necessary to fulfill the STAIR recommendations before advancing to clinical trials. STAIR is certainly not a recipe or guarantee that an agent will be effective in clinical trial. It is unknown which, if any, of these criteria are critical for the success of a stroke therapeutic.

Our analysis has important limitations. First, the search strategy may have missed studies and did not include unpublished work. As a result, almost all of the studies we analyzed reported positive results. Second, our assessment of methodological quality focused on the risk of bias and whether or not individual animal studies adequately considered important physiological variables. This assessment was based on the STAIR recommendations, which are not uniformly accepted as the gold standard by investigators in the field. However, it represents the only published consensus document that has yet to be developed. Third, we recognize that the weighted

**Table 2. Adequacy of the Preclinical Literature**

<table>
<thead>
<tr>
<th>Two Models$</th>
<th>At Least Two Laboratories</th>
<th>Two Outcome Measures</th>
<th>Two Species*</th>
<th>Therapeutic Time Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>NXY-059</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Caffeinol</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Albumin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EPO</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GCSF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Magnesium</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IFN-B</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ONO2506</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Traxoprodil</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Minocycline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

$Transient and permanent models of focal cerebral ischemia. Embolic models were not included in the fulfillment of this criterion.

*Two species where first is rodent and second is higher order such as cat or primate. Rabbits were not considered higher order to meet this criterion as specified in STAIR.
5-point quality score has not been validated and a checklist of all studies meeting the individual criteria may have more utility. Fourth, it is important to acknowledge that insufficient quality of animal testing is only one of several potential reasons to account for the clinical trial failures and that meeting quality criteria is not sufficient to predict efficacy. Finally, there may be different interpretations on the fulfillment of individual quality measures. For example, there may be divergent opinions about how and when animals should be randomized and what constitutes blinding of the animal surgeon during treatment allocation.

In conclusion, the quality and adequacy of animal testing for neuroprotective agents for stroke could be significantly improved. It is therefore premature to dismiss the importance of preclinical animal models until these studies are undertaken with greater methodological rigor. The results from this study raise concern whether agents currently in clinical trial on G-CSF.

Dr Fisher is a consultant to the company conducting a clinical trial of nxy-059.

Dr Savitz is funded by an American Heart Association grant #0475008N.

Sources of Funding

Dr Savitz is funded by an American Heart Association grant #0475008N.

Disclosures

Dr Fisher is a consultant to the company conducting a clinical trial on G-CSF.

References


Methodological Quality of Animal Studies of Neuroprotective Agents Currently in Phase II/III Acute Ischemic Stroke Trials
Maria Philip, Michael Benatar, Marc Fisher and Sean I. Savitz

_Sroke._ 2009;40:577-581; originally published online December 12, 2008; doi: 10.1161/STROKEAHA.108.524330

_Sroke_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 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/40/2/577