Empirical Evidence of Bias in the Design of Experimental Stroke Studies
A Metaepidemiologic Approach

Nicolas A. Crossley, MSc; Emily Sena, BSc; Jos Goehler; Jannekke Horn, MD; Bart van der Worp, MD; Philip M.W. Bath, MD; Malcolm Macleod, PhD; Ulrich Dirnagl, MD

Background and Purpose—At least part of the failure in the transition from experimental to clinical studies in stroke has been attributed to the imprecision introduced by problems in the design of experimental stroke studies. Using a metaepidemiologic approach, we addressed the effect of randomization, blinding, and use of comorbid animals on the estimate of how effectively therapeutic interventions reduce infarct size.

Methods—Electronic and manual searches were performed to identify meta-analyses that described interventions in experimental stroke. For each meta-analysis thus identified, a reanalysis was conducted to estimate the impact of various quality items on the estimate of efficacy, and these estimates were combined in a meta–meta-analysis to obtain a summary measure of the impact of the various design characteristics.

Results—Thirteen meta-analyses that described outcomes in 15,635 animals were included. Studies that included unblinded induction of ischemia reported effect sizes 13.1% (95% CI, 26.4% to 0.2%) greater than studies that included blinding, and studies that included healthy animals instead of animals with comorbidities overstated the effect size by 11.5% (95% CI, 21.2% to 1.8%). No significant effect was found for randomization, blinded outcome assessment, or high aggregate CAMARADES quality score.

Conclusions—We provide empirical evidence of bias in the design of studies, with studies that included unblinded induction of ischemia or healthy animals overestimating the effectiveness of the intervention. This bias could account for the failure in the transition from bench to bedside of stroke therapies. (Stroke. 2008;39:000-000.)

Key Words: animal experimentation ■ cerebrovascular accident ■ meta-analysis

During the past several decades, the modeling of stroke in animals has led to great progress in our understanding of the pathophysiologic mechanisms by which focal cerebral ischemia kills brain cells. Despite the experimental identification of numerous therapeutic strategies for stroke therapy, there has been an overall failure to validate their efficacy in patients. The latest in a long list of randomized, clinical trials with negative results in acute stroke (SAINT II, which used a free-radical scavenger) showed no efficacy despite extensive and superficially convincing preclinical data. These failures demand a reexamination of the way experimental stroke studies are conducted and fuel a debate concerning the general predictive value of experimental modeling of this complex disorder. The failure of bench-to-bedside progression may be due to problems in the way the experiments are performed, or it may result from the use of unsuitable or intrinsically flawed models; in other words, the problem may lie either in the internal or the external validity of the experiments.

Bias is a key problem in internal validity, and 4 major types have been described: selection bias (creating groups with different confounders; solved by randomization); performance bias and detection bias (investigators respectively treating or assessing more positively those subjects on the treatment arm; controlled by blinding interventions and outcome assessments); and attrition bias (dropouts of subjects with a negative outcome not included in the final result; solved by an intention-to-treat analysis or reporting of dropouts). A metaepidemiologic approach has been used in human studies to find empirical evidence of bias in internal validity (reviewed in Juni et al), thereby exposing poor allocation concealment and poor blinding of outcome assessment as consistent sources of bias in human trials.

External validity is a matter of judgment that depends on the characteristics of the subjects included, the setting, the
treatment regimens, and the outcomes assessed. Perhaps 1 of the most consistently cited problems in experimental stroke research concerns the validity of extrapolating data from young, healthy animals to elderly patients with frequent comorbid conditions in human clinical trials.6,10

Although all of these biases have been addressed in individual meta-analyses that have examined 1 intervention, they have never been studied in their overall effect throughout different interventions. Here we use a metaepidemiologic approach to evaluate the evidence of bias in internal and external validity of study designs in experimental stroke.

Methods

Inclusion Criteria and Outcome Measures
We included published or unpublished meta-analyses that reported the efficacy of potential neuroprotectant therapies in experimental focal cerebral ischemia wherein outcome was reported as a change in infarct volume. We included data for all species and for all methods of inducing focal cerebral ischemia. When data were unavailable for extraction, authors of the original meta-analyses were contacted, or data were extracted from the original studies.

Search
We searched for all pertinent meta-analyses on experimental stroke with use of a computer-based search of MEDLINE (1966 to February 2007) with the following subject headings: “cerebrovascular accident,” “meta-analysis,” “animal experimentation,” and “models, animal” and the following text words: “stroke,” “cerebrovascular,” “meta-analysis,” “meta-analysis,” “systematic review,” and “animals.” No language constraints were applied. Citations of all selected studies were searched for additional meta-analyses. Relevant published and unpublished studies were also identified from the CAMARADES web page,11 from review articles,4 and by contacting authors.

Data Extraction and Statistical Analysis
The study intended to evaluate the influence of quality variables in experimental stroke, irrespective of the therapy used. The quality characteristics studied included aspects that referred to the internal validity of the studies, such as randomized generation of the sequence of allocation, blinded induction of ischemia (allocation concealment), and blinded assessment of outcome. Studies had to explicitly report being randomized, using blinded induction, or using blinded assessment; if not, they were considered as not having the quality studied. This was obtained from the original meta-analysis, or if not reported, from the individual study. We also looked at certain aspects of external validity of the studies, particularly the effect of the use of animals with comorbidities, ie, old age, hypertension, hyperglycemia, or diabetes. We explored the overall effect of study quality by dichotomizing studies into those scoring 4 or less (low quality) or more than 4 (high quality) on the STAIR criteria.13 Attrition bias indicators (such as the reporting of excluded animals due to death during intervention or to prespecified inclusion or exclusion criteria) were not extracted, because a preliminary overview of included studies showed that such indicators were not reported with sufficient frequency to allow such an analysis.

For each study, infarct size in treatment and control groups was calculated and the normalized mean treatment effect (percentage reduction in infarct volume in the treatment group; NMD) and standard deviations were calculated. We chose this approach, in preference to standardized mean difference (SMD) meta-analysis, because NMD analysis appears to perform better when the size of individual experiments is small, presumably because the observed variance used for weighting is a less precise estimate of the population variance than is the case for larger studies.14 Furthermore, there is a potential confounding effect of weighting studies according to variance. Given first, that the observed variance represents a combination of measurement error in addition to inherent or biologic variance, and second, that low study quality is likely to be associated with high measurement error, a weighting system that included measurement error might minimize the impact of low-quality studies and therefore obscure the effects of interest. To explore whether this was indeed the case, all analyses were repeated with Hedges’ g SMD.15

A 2-level analysis was performed by a “meta–meta-analytic” approach with a random-effects model to allow for within– and between–meta-analysis heterogeneity. In brief, for each meta-analysis identified and for each variable, included studies were divided into 2 groups according to the relevant quality item (eg, blinding). Meta-analyses in which all studies were in 1 arm of the analysis (eg, all nonblinded) were not included in this part of the analysis. Infarct sizes for each study were extracted and then pooled independently for each category described by NMD analysis and a random-effects model. Two effect sizes, each with its variance, were calculated for each meta-analysis, 1 corresponding to efficacy pooled from those studies, which had the characteristic of interest (eg, blinded), and the other for those studies that did not (eg, nonblinded).

The second-order analysis involved pooling the results of the previous analysis to describe the effect of the methodologic quality item in general rather than in the context of a specific therapy. The use of a random-effects model at this stage allows for between–meta-analysis heterogeneity and does not rely on the assumption of a constant effect of the variable studied in the different therapies. Heterogeneity was tested by the χ2 test. Analyses were performed with Cochrane’s RevMan software for meta-analysis.16 A probability value <0.05 was considered statistically significant.

Results

Electronic search identified 9 studies from MEDLINE,12,17–24 1 study from a reference,25 and 1 study from the CAMARADES web page.26 Two studies unpublished at the time of the search were provided by 1 of the authors (27 and Sena and Macleod, unpublished data, 2007); this gave a total of 13 meta-analyses that described outcome in 15 635 animals (the Table).

Randomization

Eleven meta-analyses involving 14 804 animals assessed the effect of randomization in experimental stroke. Two meta-analyses were excluded because none of their studies was

<table>
<thead>
<tr>
<th>Table. Meta-Analyses Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-Analyses</td>
</tr>
<tr>
<td>Nava-Ocampo et al, 2000(^{15})</td>
</tr>
<tr>
<td>Horn et al, 2001(^{17})</td>
</tr>
<tr>
<td>Macleod et al, 2004(^{12})</td>
</tr>
<tr>
<td>Macleod et al, 2005(^{18})</td>
</tr>
<tr>
<td>Macleod et al, 2005(^{19})</td>
</tr>
<tr>
<td>Willmot et al, 2005(^{20})</td>
</tr>
<tr>
<td>Willmot et al, 2005(^{21})</td>
</tr>
<tr>
<td>Gibson et al, 2006(^{22})</td>
</tr>
<tr>
<td>Perel et al, 2007(^{23})</td>
</tr>
<tr>
<td>Sena et al, 2007(^{24})</td>
</tr>
<tr>
<td>Wheble et al(^{26})</td>
</tr>
<tr>
<td>Sena and Macleod (unpublished)</td>
</tr>
<tr>
<td>van der Worp(^{27})</td>
</tr>
</tbody>
</table>
described as being randomized.\textsuperscript{17,25} No significant effect of randomization was found (NMD, −7.0%; 95% CI, −15.1% to 1.2%; Figure 1).

**Blinded Induction of Ischemia**

Seven meta-analyses involving 8921 animals assessed the effect of blinding of the induction of ischemia; 6 meta-analyses were excluded because all of their studies were described as nonblinded.\textsuperscript{12,18,20,22,25,26} Studies not reporting blinded induction of ischemia had a significant overestimation of the effect of the therapy being studied, overestimating the effect size by 13.3% (95% CI, 0.2% to 26.4%; Figure 2).

**Blinded Assessment of Outcome**

Thirteen meta-analyses involving 15 635 animals assessed the effect of blinding the assessment of outcome. No effect of blinding the assessment of infarct size was found (NMD, 2.1%; 95% CI, 8.3% to 4.0%; Figure 3).

**Comorbidity**

Ten meta-analyses were included for assessing the effect of using animals with comorbidities, with 13 639 animals. Three meta-analyses were not included because none of their studies included animals with comorbidities\textsuperscript{19,25} (and Sena and Macleod, unpublished data, 2007). As shown in Figure 4, studies that included healthy animals tended to overestimate the normalized infarct size by 11.5% (NMD; 95% CI, 1.9% to 21.2%).

**Effect of Study Quality**

Finally, we compared high-quality and low-quality studies, dichotomized according to a score based on STAIR criteria. Twelve meta-analyses with 14 886 animals were included, and 1 meta-analysis had to be excluded because it included no high-quality studies.\textsuperscript{25} Figure 5 shows that no significant influence of “quality” was found (NMD, −3.4%; 95% CI, −8.5% to 1.7%).

Heterogeneity was significant in all of the aforementioned analyses, except for the effect of quality. All analyses were performed again with the SMD approach, and the results were broadly similar. Specifically, the effect of comorbidity was no longer significant (SMD, 0.20; 95% CI, −0.10 to 0.51). Heterogeneity was lower in all SMD analyses than in the corresponding NMD analyses.

**Discussion**

Herein we show for the first time the feasibility and utility of a metaepidemiologic approach in the assessment of the presence of bias in the design of experimental stroke studies. We provide empirical evidence on the effect of design characteristics in experimental stroke research that could partly account for the failure in the transition from bench to bedside. Previous individual meta-analyses have included stratified analyses by dividing their studies according to similar quality variables that were used in our approach and looking at their impact, but to our knowledge, this is the first time that this issue has been studied with respect to different interventions.

One of the strengths of our approach is the use of a random-effects model; this does not require the assumption of a constant effect of the variable studied in different interventions (between–meta-analysis heterogeneity).\textsuperscript{28} However, by allowing more variance in its calculations, this method compromises statistical power.

Randomization had no consistent effect in our analysis. When one considers that animals used in experimental studies represent a very homogeneous population in terms of same strain, sex, age, weight, etc, the possibility of selection bias seems small compared with the more heterogeneous human clinical situation. However, this is similar to findings from

---

**Table 1.**

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of animals</th>
<th>Change in infarct size (in percentage, 95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothermia</td>
<td>3256</td>
<td>-35.1 (-44.5, -25.7)</td>
</tr>
<tr>
<td>IL-1ra</td>
<td>784</td>
<td>-11.5 (-21.2, 0.2)</td>
</tr>
<tr>
<td>Melatonin</td>
<td>443</td>
<td>-21.6 (-39.3, 6.1)</td>
</tr>
<tr>
<td>Nimodipine</td>
<td>79</td>
<td>-15.5 (-27.6, -3.4)</td>
</tr>
<tr>
<td>NO Donors</td>
<td>480</td>
<td>-35.1 (-47.2, 0.4)</td>
</tr>
<tr>
<td>Trilazad</td>
<td>544</td>
<td>-7.6 (-12.4, 2.8)</td>
</tr>
<tr>
<td>IPA</td>
<td>3332</td>
<td>-2.27 (-4.26, 1.79)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>8921</td>
<td></td>
</tr>
</tbody>
</table>

Heterogeneity: $\chi^2$ = 20.25 (p = 0.003)
some human studies, wherein concealment of the sequence of randomization appeared to be more important. Inadequate concealment of the sequence of randomization in human studies may lead to selection bias, because investigators may consciously or unconsciously select patients to 1 or the other arm according to their characteristics and likely prognosis. This concern is rarely addressed in the experimental stroke literature. One of the factors most likely to influence outcome in experimental stroke is the care and enthusiasm of induction of ischemia and the precise duration of that ischemia; therefore, blinding the investigator to group assignment (allocation concealment, or blinded induction of ischemia) is crucial and meets the same purpose as concealment of allocation sequence in human studies. Indeed, we have shown a significant overstatement of efficacy when induction of ischemia was unblinded.

Blinding the assessment of outcome had no effect on efficacy when outcome was measured as infarct size. Perhaps as a result of the use of semiautomated measurement techniques, infarct size appeared to be a robust and relatively objective measure of outcome less prone to observer bias; the performance of other outcome measures, for instance, neurobehavioral scores, is not known.

Interestingly, studies that included animals with comorbidities and those that included healthy animals differed by \( \approx 10\% \) in their effect size, a difference similar to the one found in those that used blinded induction of ischemia compared with those that were unblinded. Because most interventions reported a decrease of 30% to 40% of the control infarct size, this result seems to be of great importance.

Finally, despite an effect of 2 components of the study score, we found no apparent difference between high- and low-quality studies. Previous individual meta-analyses on neuroprotectant interventions did look at the difference in infarct size according to a quality scale and found varying results, including a decreased effect with increasing quality, no clear relation, and even a decrease in effect with poor quality. All of these studies had less power than our approach, because they looked at only 1 intervention instead of analyzing and pooling several different therapies. One other study reported the effect of a similar quality scale in the effect size by pooling several interventions, and importantly, although using a different approach, O’Collins et al. found no difference between the effect sizes of studies defined as high compared with low quality. There are a number of possible explanations for this. First, the bias from different quality characteristics might operate in different directions (some increasing and others reducing the estimate of efficacy). However, in univariate analysis, there is no evidence for this (Sena and Macleod, unpublished observations). Second, the pooling of multiple characteristics in a global quality score might dilute the effects of important predictors of bias. The interpretation of quality data as existing on an ordinal scale with the same weight attributed to different aspects of methodologic quality is clearly a highly simplistic view of the complex entity of study quality. Finally, dichotomization of scales for statistical analysis introduces bias, and minor changes in cutoffs for dichotomization may strongly affect the result of the analysis. We propose that qualitative scales might be most useful as a “checklist” to qualitatively describe different studies, rather than as a quantitative marker of overall quality.
Our approach has a number of potential weaknesses. The statistical power of this approach is unknown, and we might falsely conclude that a potential source of bias is unimportant when in fact it is. To avoid the problem of multiple comparisons, a finite number of variables were included, but many important variables were excluded, particularly with respect to the external validity of the studies, such as drug dosing time or timing of the assessment of outcome. Specifically, these 2 variables were balanced in our studies and did not confound the results, but their impact remains to be assessed in future studies. Furthermore, this was essentially a univariate approach; interactions between different potential sources of bias or between potential sources of bias and other attributes of contributing studies (such as the drug or species used, drug dose, etc) were not captured in this analysis, and a multivariate approach would be required.

In line with similar meta-epidemiological studies of clinical meta-analyses, our current study looked at the quality reported by the studies and assumed that quality was inadequate unless information to the contrary was provided (the “guilty until proved innocent” approach), and therefore, some studies might have been deemed as not having quality when the problem was of underreporting. Although we cannot rule out the possibility that including “real” quality variables instead of reported quality variables would have affected our results, there is some evidence that the frequency of underreporting quality variables is low, and therefore also the results, there is some evidence that the frequency of underreporting quality variables is low, and therefore also the probability of having an impact in our analysis.

Infarct size is certainly a widespread outcome reported in experimental stroke, but valid doubts exist regarding how useful this outcome is for human studies. It would be interesting to study the impact of these quality variables on this type of outcome. Unfortunately, although we have seen in recent years an improvement in the reporting of compliance with legislative requirements promulgated by STAIR, there has been no substantial improvement in study quality. We hope that our revelation of the importance of these issues contributes to a change in their practice.

Conclusions
Blinding the induction of ischemia and the use of comorbid animals each significantly affected the estimate of how effective an intervention was in experimental stroke; Effect sizes in studies with or without either of these characteristics differed by \( \approx 10\% \) in their effect size. Given an effect size for most interventions of between 30% and 40%, this result is of substantial importance. Such design characteristics can introduce bias in experimental stroke studies that can at least partly account for the failure in the transition from bench to bedside.

Source of Funding
This study was supported by the Hermann and Lilly Schilling Foundation (UD).

Disclosures
None.

References


**Table. Quality Scale**

<table>
<thead>
<tr>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer-reviewed publication</td>
</tr>
<tr>
<td>Statement of control of temperature</td>
</tr>
<tr>
<td>Random allocation to treatment or control</td>
</tr>
<tr>
<td>Blinded induction of ischemia</td>
</tr>
<tr>
<td>Blinded assessment of outcome</td>
</tr>
<tr>
<td>Use of anesthetic without significant intrinsic neuroprotective activity</td>
</tr>
<tr>
<td>Appropriate animal model (aged, diabetic, or hypertensive)</td>
</tr>
<tr>
<td>Sample size calculation</td>
</tr>
<tr>
<td>Compliance with animal welfare regulations</td>
</tr>
<tr>
<td>Statement of potential conflicts of interest</td>
</tr>
</tbody>
</table>

The scale used was originally proposed by Macleod et al and based on the STAIR criteria, with a possible maximum score of 10. One point is given for each criterion met.