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

Nicolas A. Crossley, MSc; Emily Sena, BSc; Jos Goehler; Janneke 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;39:929-934.)

Key Words: animal experimentation ■ cerebrovascular accident ■ meta-analysis
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 through-
out 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: “cerebrovascu-
lar accident,” “meta-analysis,” “animal experimentation,” and “mod-
els, animal” and the following text words: “stroke,” “cerebrovascu-
lar,” “meta-analysis,” “meta-analysis,” “systematic review,” and
“animals.” No language constraints were applied. Citations of all
selected studies were searched for additional meta-analyses. Rele-
vant 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 CAMARADES
study quality score (Macleod et al12; see supplemental Table I for
definitions, available at http://stroke.ahajournals.org/), a widely used
score based 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-analy-

sis 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 CAMA-
RADES 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. Meta-Analyses Included

<table>
<thead>
<tr>
<th>Meta-Analysis</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nava-Ocampo et al, 2000</td>
<td>Glutamate release blockers</td>
</tr>
<tr>
<td>Horm et al, 2001</td>
<td>Nimodipine</td>
</tr>
<tr>
<td>Macleod et al, 2004</td>
<td>Nicotinamide</td>
</tr>
<tr>
<td>Macleod et al, 2005</td>
<td>FK506</td>
</tr>
<tr>
<td>Macleod et al, 2005</td>
<td>Melatonin</td>
</tr>
<tr>
<td>Willmot et al, 2005</td>
<td>Nitric oxide synthase inhibitors</td>
</tr>
<tr>
<td>Willmot et al, 2005</td>
<td>Nitric oxide donors and L-arginine</td>
</tr>
<tr>
<td>Gibson et al, 2006</td>
<td>Estrogens</td>
</tr>
<tr>
<td>Perel et al, 2007</td>
<td>Tissue-type plasminogen activator</td>
</tr>
<tr>
<td>Sena et al, 2007</td>
<td>Tirilazad</td>
</tr>
<tr>
<td>Wheble et al26</td>
<td>Piracetam and Piracetam-like agent</td>
</tr>
<tr>
<td>Sena and Macleod (unpublished)</td>
<td>Interleukin-1 receptor antagonist</td>
</tr>
<tr>
<td>van der Worp27</td>
<td>Hypothermia</td>
</tr>
</tbody>
</table>
described as being randomized. 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. 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 (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. Figure 5 shows that no significant influence of “quality” was found (NMD, −3.4%; 95% CI, −8.5% to 1.7%).

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). 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...
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 \( \sim 10\% \) 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,\(^9\)\(^8\),\(^22\) no clear relation,\(^17\),\(^19\) and even a decrease in effect with poor quality.\(^20\) 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\(^29\) 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.\(^2\),\(^3\),\(^31\) 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 quantitatively 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,

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of animals</th>
<th>Change in infarct size</th>
</tr>
</thead>
<tbody>
<tr>
<td>E10g</td>
<td>1452</td>
<td>-16.83 (-34.84, 0.38)</td>
</tr>
<tr>
<td>F10g</td>
<td>1596</td>
<td>0.93 (-19.21, 20.07)</td>
</tr>
<tr>
<td>Hysteresm</td>
<td>3256</td>
<td>-5.11 (-14.11, 1.86)</td>
</tr>
<tr>
<td>I1-n</td>
<td>704</td>
<td>-1.05 (-23.08, 22.98)</td>
</tr>
<tr>
<td>Meteanin</td>
<td>157</td>
<td>-5.99 (-25.59, 2.51)</td>
</tr>
<tr>
<td>Nicotamade</td>
<td>719</td>
<td>3.36 (-6.12, 15.72)</td>
</tr>
<tr>
<td>Nidojoline</td>
<td>82</td>
<td>24.41 (-6.05, 49.51)</td>
</tr>
<tr>
<td>NO Conores</td>
<td>483</td>
<td>-7.83 (-23.98, 8.32)</td>
</tr>
<tr>
<td>NCS Inhibitors</td>
<td>1598</td>
<td>4.04 (4.65, 13.03)</td>
</tr>
<tr>
<td>Pracatin</td>
<td>197</td>
<td>-25.28 (-47.12, 5.54)</td>
</tr>
<tr>
<td>Trilazad</td>
<td>544</td>
<td>-8.27 (-27.27, 14.73)</td>
</tr>
<tr>
<td>IFA</td>
<td>3332</td>
<td>-0.65 (-8.09, 7.79)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>14886</td>
<td>-3.39 (-8.50, 1.72)</td>
</tr>
</tbody>
</table>

Table 1. Results of the meta-analysis.

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


Empirical Evidence of Bias in the Design of Experimental Stroke Studies: A Metaepidemiologic Approach
Nicolas A. Crossley, Emily Sena, Jos Goehler, Jannekke Horn, Bart van der Worp, Philip M.W. Bath, Malcolm Macleod and Ulrich Dirnagl

*Stroke.* 2008;39:929-934; originally published online January 31, 2008; doi: 10.1161/STROKEAHA.107.498725

*Stroke* 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/39/3/929

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org/subscriptions/