

## Review Article

# Bench to bedside: the quest for quality in experimental stroke research

Ulrich Dirnagl

Department of Experimental Neurology, Center for Stroke Research, Humboldt-Universität Berlin, Universitätsklinikum Charité, Berlin, Germany

**Over the past decades, great progress has been made in clinical as well as experimental stroke research. Disappointingly, however, hundreds of clinical trials testing neuroprotective agents have failed despite efficacy in experimental models. Recently, several systematic reviews have exposed a number of important deficits in the quality of preclinical stroke research. Many of the issues raised in these reviews are not specific to experimental stroke research, but apply to studies of animal models of disease in general. It is the aim of this article to review some quality-related sources of bias with a particular focus on experimental stroke research. Weaknesses discussed include, among others, low statistical power and hence reproducibility, defects in statistical analysis, lack of blinding and randomization, lack of quality-control mechanisms, deficiencies in reporting, and negative publication bias. Although quantitative evidence for quality problems at present is restricted to preclinical stroke research, to spur discussion and in the hope that they will be exposed to meta-analysis in the near future, I have also included some quality-related sources of bias, which have not been systematically studied. Importantly, these may be also relevant to mechanism-driven basic stroke research. I propose that by a number of rather simple measures reproducibility of experimental results, as well as the step from bench to bedside in stroke research may be made more successful. However, the ultimate proof for this has to await successful phase III stroke trials, which were built on basic research conforming to the criteria as put forward in this article.**

*Journal of Cerebral Blood Flow & Metabolism* (2006) 26, 1465–1478. doi:10.1038/sj.jcbfm.9600298; published online 8 March 2006

**Keywords:** negative publication bias; power analysis; quality control; reporting; standard operating procedures; study design; systematic review

## Introduction

Over the past decades, great progress has been made in the prevention, diagnosis, and treatment of stroke (Cheng *et al*, 2004; Ward and Cohen, 2004; del Zoppo, 2004). At the same time, by modeling of stroke in animal experiments, our understanding of the pathophysiologic mechanisms by which focal cerebral ischemia kills brain cells has greatly improved (Lo *et al*, 2003; Dirnagl *et al*, 1999). In addition, experimental stroke studies have identified numerous therapeutic targets for stroke therapy (Dirnagl *et al*, 2003; Meisel *et al*, 2005; Endres, 2005;

Lo *et al*, 2005; Lindvall and Kokaia, 2004). However, it is an overall failure to validate the efficacy of such treatment strategies, which is fuelling a debate concerning the general predictive value of experimental modeling of this complex disorder (Hoyte *et al*, 2004; Fisher and Tatlisumak, 2005; Kaste, 2005). Much of the discussion on why bench results get 'lost in translation' to the bedside has focused on issues such as species differences, side effects, heterogeneity of strokes in humans, etc. (see below). Nevertheless, quality issues in experimental research that may impact on bench to bedside translation have recently been raised by a number of authors (Heard *et al*, 2005; Bebarta *et al*, 2003; Pound *et al*, 2004; Samsa and Matchar, 2001). More specifically, systematic reviews of experimental stroke research have exposed a number of important deficits in the quality of preclinical experimental stroke research, and a negative correlation between measured effect sizes and quality scores was demonstrated (see below, and Table 1). The discussion has been further sparked by research, which provocatively predicts that most published research

---

Correspondence: Dr U Dirnagl, Department of Experimental Neurology, Center for Stroke Research, Humboldt-Universität Berlin, Universitätsklinikum Charité, Schumannstrasse 20/21, Berlin 10098, Germany.  
E-mail: ulrich.dirnagl@charite.de

This work was funded by the Herman and Lilly Schilling Foundation.

Competing interests: None.

Received 20 November 2005; revised and accepted 26 January 2006; published online 8 March 2006

**Table 1** Overview of available quality score data from systematic reviews

Systematic review	Theoretical range of quality scores <sup>a</sup>	Quality score (median)	Quality score (range)	Number of studies reviewed (total = 357)	Effect of quality score on effect size
C Briscoe, E Sena, PAG Sandercock and MR Macleod, personal communication, 2005	0–10	4	0–8	115	‡
ES Sena, PCR Wheble, PAG Sandercock and MR Macleod, personal communication, 2005	0–10	5	2–6	18	NS
PCR Wheble, ES Sena, PAG Sandercock and MR Macleod, personal communication, 2005	0–10	4	2–6	5	†,‡
Willmot <i>et al</i> (2005a)	0–8	3	1–6	73	†
Macleod <i>et al</i> (2005b)	0–10	4	0–6	14	NS
Macleod <i>et al</i> (2004)	0–10	3,5	0–7	14	‘Studies scoring highly tended to give a more precise estimate of the global estimate of the effect size’
Macleod <i>et al</i> (2005c)	0–10	4	0–7	29	†,‡
Horn <i>et al</i> (2001)	0–8	2.85	1–5	19	Not tested
Macleod <i>et al</i> (2005b)	0–8	2	1–7	25	Not tested
van der Worp <i>et al</i> (2005)	0–10	3	1–7	45	Not tested

<sup>a</sup>Horn *et al* (2001), based on the Stroke Therapy Academic Industry Roundtable (2001) criteria, suggested a scale of 0–8, which in some articles has been modified to a scale of 0–10; it is therefore impossible to combine the data.

†Quality score accounts for a significant amount of group heterogeneity of effect size of treatment; ‡Significant negative correlation between quality score and effect size of treatment (i.e., lower quality scores associated with an increased estimate of the effect size); NS, not significant.

findings are ‘false’ (Ioannidis, 2005), and that quality issues are important reasons for this shocking finding. In this article, I would like to explore whether translational experimental stroke research is affected by quality problems. The purpose of my critical approach is not to denounce experimental stroke research, but to initiate a discussion geared at improving the translation of basic research into successful clinical stroke therapy. It is important to note that animal experiments in other areas of medicine, such as neurosurgery (Heard *et al*, 2005), cardiology (Lee *et al*, 2003), cancer (Clarke, 1997), or intensive care (Bebarta *et al*, 2003; Roberts *et al*, 2002), are affected by similar quality issues (Pound *et al*, 2004). In addition, I would like to stress that quality issues are but one factor that may impact on the success of bench to bedside translation in stroke research: validity of models, species issues, clinical trial design, among many others, may be equally or even more relevant determinants.

In this article, I will concentrate on ‘experimental stroke research’, by which I refer to animal experiments, which test the effect of a physical, pharmacological, or genetic manipulation on outcome after an experimentally induced stroke. Experimental stroke research spans from explorative, basic science into disease mechanisms, to preclinical studies geared at informing companies and clinicians as to whether a compound, diagnostic or therapeutic principle, should be investigated in phase I/II clinical trial. As basic and preclinical stroke

**Table 2** Summary of potential weaknesses in experimental stroke studies and relative relevance to mechanism-driven basic stroke research versus preclinical stroke research

Weakness	Relative relevance to	
	Basic stroke research	Preclinical stroke research
Randomization and blinding missing	++	+++
Inclusion and exclusion criteria not properly set	+	+++
Lack of control of relevant physiological parameters	+++	+++
Statistical power too low	+++	+++
Improper statistics	++	+++
Deficient reporting	++	+++
Negative publication bias	+	+++
Correlation did not establish causality	+++	+
Lack of quality control	+	+++
Bias introduced by interest in positive study outcome	++	+++

+, Moderate; ++, high; +++, very high impact.

research have different aims and designs, some quality matters discussed here may have different impact on both types of studies. Where necessary, I will therefore point out the specific relevance of quality issues to basic and preclinical research (Tables 2 and 3). Issues like choice of species, modeling of age, and complicating diseases, etc., are almost

**Table 3** Specific recommendations for conducting research in mechanism-driven basic stroke studies or preclinical stroke studies (✓ denotes high relevance)

Recommendation	Type of stroke research	
	Basic	Preclinical
Randomize	✓	✓
Perform surgery as well as evaluation blind to exp. groups	✓	✓
<i>A priori</i> set inclusion and exclusion criteria of experimental animals		✓
Analyze as randomized (intention-to-treat analysis)		✓
Measure, control, and report relevant physiological parameters	✓	✓
<i>A priori</i> perform power analysis	✓	✓
<i>A priori</i> specify $\alpha$ , $\beta$ , and the planned $n$	✓	✓
Report and analyze ordinal data with median, range, and $\chi^2$ test	✓	✓
Use CIs, do not use SEMs	✓	✓
In case of doubt, seek professional advice on statistics	✓	✓
Replicate key findings	✓	✓
Report mortalities	✓	✓
Report exact details on animal (sub)strains (including genetically modified animals)	✓	✓
Report sponsorship, conflict of interest	✓	✓
Define SOPs (including quality checks)		✓
Consider using animals with comorbidities and aged animals		✓

exclusively relevant to preclinical stroke research (see below) and have been discussed a number of times previously (most prominently: STAIR, Stroke Therapy Academic Industry Roundtable, 1999), and they will therefore only be mentioned in passing.

### Decades of Progress in the Diagnosis and Treatment of Acute Stroke

Although effective pharmacological neuroprotection (or rather ‘brain protection’) is still unavailable for the treatment of patients with acute stroke, stroke diagnosis and therapy have made major progress over the past decades. Importantly, the demonstrated efficacy of recanalization within 3 h of the focally ischemic brain region via thrombolysis by recombinant tissue plasminogen activator was a major breakthrough in clinical medicine (Hacke *et al*, 1995; The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). In parallel to the establishment of this first pharmacological treatment of acute stroke, the ‘time is brain’ concept has emerged, and has led to significant improvements in the infrastructure of acute stroke care. In particular, stroke is now regarded as a treatable medical emergency by the general public, the emergency medical system, as well as physicians, with a resultant decrease in the time lag between symptom onset and initiation of treatment (Evenson *et al*, 2001; Hill *et al*, 2000). As a result of this concept of acute stroke as a treatable disease with a very short therapeutic window, a network of dedicated stroke units has been established in many countries. A clear therapeutic benefit of stroke units has been demonstrated (Langhorne

and Dennis, 2004; Alberts *et al*, 2005; Stroke Unit Trialists’ Collaboration, 1997; Indredavik *et al*, 1999; Drummond *et al*, 2005). Advanced diagnostic imaging of stroke is now widely available, and includes early measures of damage (DWI) as well as blood perfusion (magnetic resonance (MR)-angiography, MR perfusion, computed tomography perfusion; Warach, 2003; Hjort *et al*, 2005). In addition, stroke trial design has come of age. In particular, the study designs of randomized clinical trials (RCTs) have been refined with regards to patient selection and trial size, outcome measures and assessment, follow-up, as well as statistical analysis (Lees *et al*, 2003; Krams *et al*, 2005). All these improvements taken together have created an environment that is highly amenable for the successful establishment of pharmacological brain protection in acute stroke.

### Decades of Progress in Basic Stroke Pathophysiology

These impressive achievements in clinical stroke research and therapy were accompanied by two decades of tremendous progress in basic stroke research. Robust animal models for ischemic stroke are now available. They may not faithfully reflect the pathophysiologic process, which leads to a stroke, but most researchers agree that they are suited to study the pathobiology that leads to tissue damage once a major brain vessel is occluded (Mergenthaler *et al*, 2004; Hossmann, 2004; Liang *et al*, 2004; Richard *et al*, 2003; Fukuda and del Zoppo, 2003; Traystman, 2003; Endres and Dirnagl, 2002; Leker and Constantini, 2002). Research on animal models of stroke, but also on cultured brain

cells and tissues, has led to a basic understanding of how focal substrate deprivation kills brain cells (Dirnagl *et al*, 1999). In addition, we have learned that the brain mounts an endogenous protective response to reduce the deleterious effects of substrate deprivation (Dirnagl *et al*, 2003). Understanding such endogenous neuroprotective mechanisms has led to novel strategies of therapeutic brain protection. Although numerous questions with regard to this complex pathophysiology remain unanswered, and a plethora of cellular and molecular mechanisms remain to be explored, many promising therapeutic targets have been found. Experimentally induced acute stroke of the rodent is a treatable disease, as typical outcomes in experimental stroke studies have shown reductions in infarct size and improvements in neurologic outcome of 50% or more (see below). Although traditionally experimental stroke research has focused on early mechanisms of damage, only recently a surprising capacity of the central nervous system

(CNS) for regeneration and plasticity has been identified, raising the hopes of developing therapeutic approaches based on these findings (Chen *et al*, 2002; Lindvall and Kokaia, 2004; Savitz *et al*, 2004, 2002; Roitberg, 2004; Chopp and Li, 2002). Further progress in the field has come from a recent tremendous expansion of the methodological portfolio of experimental stroke research, ranging from genome-wide transcriptional and protein expression screens (Carmichael, 2003; Hubner *et al*, 2005), to the targeted deletion or overexpression of candidate genes (Liang *et al*, 2004), to sophisticated high-resolution *in vivo* imaging (Dijkhuizen and Nicolay, 2003).

## Lost in Translation?

Surprisingly, these exciting developments in basic and clinical stroke research occurred in parallel, with too little direct translation between bench and

**Table 4** Useful resources on the web

<i>Title</i>	<i>URL</i>	<i>Comment</i>
Cochrane Collaboration	<a href="http://www.cochrane.org">www.cochrane.org</a>	Premier source on systematic review and meta-analysis
Collaborative Approach to Meta-analysis and Review of Animal Data from Experimental Stroke (CAMARADES)	<a href="http://www.camarades.info">www.camarades.info</a>	Framework for groups involved in the systematic review and meta-analysis of data from animal studies in experimental stroke
Consolidated Standards of Reporting Trials (CONSORT)	<a href="http://www.consort-statement.org">www.consort-statement.org</a>	Checklist and flow diagram to help improve the quality of reports of randomized controlled trials
Current controlled trials	<a href="http://www.controlled-trials.com">www.controlled-trials.com</a>	Allows users to search, register, and share information about randomized controlled trials
GPower	<a href="http://www.psych.uni-duesseldorf.de/aap/projects/gpower">www.psych.uni-duesseldorf.de/aap/projects/gpower</a>	Simple but powerful program for sample size calculations, power analysis, etc. Freeware
<i>Journal of Negative Results in Biomedicine</i>	<a href="http://www.jnrmb.com/home">www.jnrmb.com/home</a>	Peer-reviewed, open-access, PubMed-indexed publishing of negative results
Primer on Bayesian statistics	<a href="http://www.bayesian-initiative.com">www.bayesian-initiative.com</a>	A well-written and easy-to-understand primer for those who want to understand the basics of Bayesian statistics and how Bayesian methods may be applied. Balanced comparison between frequentist and Bayesian approaches
Statistics at square one	<a href="http://bmj.bmjournals.com/collections/statsbk/index.shtml">bmj.bmjournals.com/collections/statsbk/index.shtml</a>	The classic text by TDV Swinscow
Uniform requirements for manuscripts submitted to <i>Biomedical Journals</i>	<a href="http://www.icmje.org">www.icmje.org</a>	Helps authors and editors in their mutual task of creating and distributing accurate, clear, and easily accessible reports of biomedical studies
Washington University Internet Stroke Center	<a href="http://www.strokecenter.org">www.strokecenter.org</a>	Comprehensive resource on stroke, includes stroke trial register (completed and ongoing)

bedside. Pharmacological brain protection, which has been a major focus of and a reality in experimental models of stroke for 20 years, is not available to stroke patients, despite hundreds of phase II and III clinical studies and billions of dollars spent in these trials (for an overview, see Washington University Internet Stroke Center, URL in Table 4). It is important to stress that there is no principal barrier for bench to bedside translation in the stroke field. Two key concepts, which were developed in *in vivo* models of cerebral ischemia, successfully found their way into clinical practice: the concept of the ischemic penumbra (Astrup *et al*, 1977; Symon *et al*, 1975) and fibrinolysis in acute ischemic stroke (Zivin *et al*, 1985) were developed in animal models of cerebral ischemia, and are now the cornerstones of diagnosis and treatment of stroke (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995; Fisher, 2004; Kidwell *et al*, 2003).

Many highly cited articles (Dirnagl *et al*, 1999; Wahlgren and Ahmed, 2004; DeGraba and Pettigrew, 2000; del Zoppo, 1998; Danton and Dietrich, 2004; Curry, 2003; Green, 2002) and expert symposia, most importantly the Stroke Therapy Academic Industry Roundtable (STAIR, Stroke Therapy Academic Industry Roundtable, 1999), have tried to explain this striking lack of success of bench to bedside translation in the stroke field. High up on the list of potential reasons explaining this apparent 'loss in translation' are species differences, inappropriate time windows of treatment, effective drug levels not achievable in humans because of toxicity, use of young animals without comorbidity, failure to model white matter damage and protect axons, incongruent end points (infarct size in animal experiments versus neurologic outcome in clinical trials), heterogeneity of stroke subtypes in patients, unrealistic expectation of effect size and false negatives because of a lack of statistical power in clinical trials, among others.

Surprisingly, a potential contributor to the apparent gap between bench and bedside results has only received very little attention: that quality problems in some of the basic research or preclinical studies on which clinical trials were based may have led to false-positive results, inflated effect sizes, and marginal reproducibility. Several authors have mentioned this possibility (Heard *et al*, 2005; Pound *et al*, 2004), and recent systematic reviews (Table 1) have found quantitative evidence for it in preclinical stroke research. However, the subject was never addressed comprehensively.

## A Provocative Hypothesis

I propose that a relevant fraction of experimental stroke studies are affected by quality problems, and that this, among several other factors, may negatively impact successful bench to bedside

translation. I further propose that by a joint effort of scientists and journal editors, this problem can be corrected. In the following, I will discuss the most common and important quality problems in this field, illustrated by examples on how they impact on study results (overview in Table 2). My hypothesis is mainly based on systematic reviews of preclinical stroke research in which a 'quality score' was applied (C Briscoe, E Sena, PAG Sandercock and MR Macleod, personal communication, 2005; ES Sena, PCR Wheble, PAG Sandercock and MR Macleod, personal communication, 2005; PCR Wheble, ES Sena, PAG Sandercock and MR Macleod, personal communication, 2005; Willmot *et al*, 2005a,b; Macleod *et al*, 2004, 2005b,c,d; Horn *et al*, 2001). As the existing systematic reviews have a limited coverage of published experimental stroke studies, and as such quality scores represent a rather coarse measure, which does not include a number of potentially relevant aspects of 'quality', and are therefore subject to criticism (see below), I will also raise some issues which have not been systematically reviewed so far. In these cases, quantitative evidence is lacking that they indeed significantly affect the effect sizes or general outcomes of experimental stroke studies. The purpose of this approach is to generate a framework for the improvement of quality scales, and to promote systematic studies into the effects of these potential determinants of quality and hence outcome in preclinical as well as basic stroke research.

## Quantification of Methodological Quality in Preclinical Stroke Studies

Following the quality criteria of the STAIR recommendations (Stroke Therapy Academic Industry Roundtable, 1999), Horn *et al* (2001), were the first to analyze methodological quality of preclinical experimental stroke studies in a systematic review. In their rating scale (possible total score of 8), points are attributed for (1) the dose/response relationship that was investigated, (2) randomization of the experiment, (3) optimal time window of the treatment investigated, (4) monitoring of physiologic parameters, (5) masked outcome assessment, (6) assessment of at least two outcomes (infarct size and functional outcome), (7) outcome assessment in the acute phase (1 to 3 days), and (8) outcome assessment in the chronic phase (7 to 30 days). Macleod *et al* (2004) have further refined this scale by giving one point (possible total score of 10) each for (1) peer-reviewed publication, (2) statement of control of temperature, (3) random allocation to treatment or control; (4) masked induction of ischemia; (5) masked assessment of outcome, (6) use of anesthetic without significant intrinsic neuroprotective activity, (7) appropriate animal model (aged, diabetic, or hypertensive), (8) sample

size calculation, (9) compliance with animal welfare regulations, and (10) statement of potential conflict of interests. Further modifications of these scores exist (van der Worp *et al*, 2005).

These authors should be commended for approaching the issue of methodological quality in a quantitative manner. However, the ordinal data from such scales have to be interpreted with great caution. For example, does random allocation to experimental groups have the same implications with regard to methodological quality as compliance with animal welfare? Further, only binary decisions are possible in each category: the use of an aged hypertensive animal thus gives the same score as the use of an aged one. Power calculation *per se* is rewarded with one point, regardless of its result, etc. As will be shown below, a number of additional indicators for methodological quality can be identified (e.g., flaws in statistics, reporting, study interpretation, etc.), which are not scored in the presently used systems.

Notwithstanding the limitations of quality scores, the fact that the median quality scores of 357 preclinical stroke studies systematically reviewed in 10 articles were 50% or less of the maximum score is *per se* alarming. Horn *et al* (2001) qualified studies scoring below 50% as 'of poor methodological quality'. Importantly, many of these meta-analyses were able to show that the quality score accounted for a significant amount of group heterogeneity of effect size of treatment, and/or that there was a statistically significant negative correlation between quality score and effect size of treatment. In other words, lower quality scores were associated with an increased estimate of the effect size.

It should be noted here that the scores discussed above were developed to assess methodological quality of preclinical studies. For example, it is certainly not a sign of low quality if a basic research study identifies a mechanism or a novel therapeutic target, but does so only in adolescent animals, and not in aged ones. However, some of the 'quality' ratings may also be applied to basic research (e.g., randomization, monitoring of physiologic parameters, statistics, etc.). I propose that specific scales should be developed for the systematic review of mechanism-driven explanatory studies.

As quantitative indicators from systematic reviews point to defects in methodological quality in preclinical stroke research, I would like to now explore this issue in greater detail by scrutinizing the relevant methodological aspects that collectively determine the 'quality' of a study. I will expand my discussion to include aspects relevant for basic stroke research.

## Study Design

Wherever systematically studied, it has been quantitatively shown that experimental studies that

violate standards of study design (Festing, 2003, 2006; Johnson and Besselsen, 2002), such as randomization and blinding (see below), are more likely to report differences between study groups (Bebarta *et al*, 2003; Lee *et al*, 2003). It appears to be common practice to start testing a compound without *a priori* setting an effect size that will be considered biologically meaningful, and without accordingly calculating the necessary numbers of animals (see below). It is very rare that detailed protocols are specified. It appears that often without consideration of statistical power between 6 and 12 animals per group are investigated, and a *P*-value (usually 0.05) is used to determine the biologic relevance of the findings. In the following, I will discuss several study design issues in more detail.

## Randomization

In a systematic review by Bart van der Worp *et al* (2005), only 42% of 45 analyzed experimental stroke studies reported randomized treatment allocation. It appears trivial that random allocation to the various experimental groups is a mandatory exercise. Randomization is necessary to exclude systematic errors of sampling. These may arise, for example, if one batch of animals is affected by an undetected infection, or if the experimenter on a 'bad day' only operates control animals, or if the surgeon is trained while operating the control animals, and investigates the treatment group once the technique is mastered. For all these reasons, it is unacceptable to use historical controls in stroke research.

## Blinding

As with randomization, blinding is reported only in a fraction of experimental stroke studies: in one systematic review (van der Worp *et al*, 2005), only 22% of 45 studies analyzed administered study medication in a masked manner. We have to accept that investigators are biased towards the outcomes of their experiments, which necessitates an outcome assessment, which is blind to the experimental groups. In addition, surgery should be performed without prior knowledge to which experimental group the animal belongs (e.g., an mutant mouse), or will be randomized (in a drug trial). For obvious reasons, double-blind protocols, in which the subject under study is unaware to which group it belongs, are neither feasible nor necessary in animal experiments. Interestingly, however, double-blind animal stroke studies have been published (Germano *et al*, 1987; Xu *et al*, 2005), raising concerns about the review process (see below).

## Inclusion and Exclusion Criteria

Only a negligible minority of experimental stroke studies report *a priori* criteria for inclusion or exclusion of animals. It is reasonable to assume that even with highly skilled surgeons, practically all middle cerebral artery occlusion (MCAO) models have at least some mortality (approx. 5%). It is therefore distressing that many experimental stroke studies do not report mortalities (27% of the analyzed studies in van der Worp *et al* (2005)). Why is it important to *a priori* set and report such criteria, as well as publish how the experimental cohorts were stratified accordingly? A strong bias can be introduced by exclusion of certain animals after randomization, such as those that did show no or 'unexpectedly small' infarcts, or had 'unphysiologic' blood gases or systemic arterial pressures. Even such a 'clear' exclusion criterion as death of the animal before reaching end-point evaluation may pose problems: a putative neuroprotectant X may be toxic (e.g., respiratory depressant), but without any direct effect on infarct size. Because of the differences in vascular collaterals, among other factors, MCAO produces quite heterogeneous infarcts and neurologic outcomes in individual rodents. A scenario can be envisioned, in which compound X kills only animals that are most severely affected by MCAO (as they are already struggling to survive), whereas animals with minor damage are able to survive and will be evaluated for outcome. Thus, the surviving treated animals have been selected for smaller infarcts, and the toxic compound X appears as effective as a neuroprotectant!

## Control of Relevant Physiologic Parameters

In 1990, Buchan and Pulsinelli explained the neuroprotective effect of the N-methyl-D-aspartate receptor blocker MK-801 in a model of global cerebral ischemia with its effect on postischemic body (and thus brain) temperature (Buchan and Pulsinelli, 1990). This is probably the most prominent example in which the effects of an experimental manipulation (e.g., a drug, or a genetic manipulation) were explained by its effect on a systemic variable (such as blood pressure, or temperature), or on cerebral blood flow, and not by a direct neuroprotective effect. In the last decade of the 20th century, therefore, great emphasis was put by researchers, reviewers, and editors on demonstrating the (lack of) effect of the mechanistic manipulation under study on such variables. In parallel to a proliferation of studies in (genetically modified) mice, in which physiologic monitoring is much more demanding and less reliable than in larger rodents, the emphasis on proper monitoring and control of physiologic parameters (including

cerebral blood flow) has clearly diminished. There is no formal proof that this has significantly affected experimental stroke research, but this trend should be closely monitored, and its effects investigated.

## Statistics

Scholarly journals like *Stroke* or the *Journal of Cerebral Blood Flow* no longer feature statistical consultants on their editorial boards, and statisticians no longer partake in the scientific discourse (Ford, 1983). In addition, instructions for authors do not mention statistics, let alone specific statistical requirements or standards. In contrast, some experimental (e.g., American Physiological Society, 2005) and eminent clinical journals like *The New England Journal of Medicine*, *The British Medical Journal*, or *The Lancet* publish extensive and specific guidelines for acceptable statistical procedures, as proposed by the 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication' (URL see Table 4). This apparent lack of attention to statistics is distressing, as it is not uncommon to find violations of statistical standards in published papers in experimental stroke research.

As statistical power, or the lack thereof, is one of the most common statistical problems encountered in experimental stroke studies, I will focus on this issue first. When experimental stroke studies report negative results, it is very rare to find mention of *a priori* power analysis or the effect size that could have been detected given the variance of the data and the preset levels for  $\alpha$  (risk of committing a type I error, or false positive) and  $\beta$  (risk of committing a type II error, or false negative, (Sterne and Davey, 2001; Mulaik *et al*, 1997) (Table 1). van der Worp *et al* (2005), in a systematic review of 45 experimental stroke studies, identified only one study that performed power analysis.

This is particularly distressing, as it appears to be unknown to most researchers that statistical power is not only an issue when the null hypothesis was not rejected (i.e., no difference was found between experimental groups at a given  $\alpha$ ). When the null hypothesis is in fact false (the drug really works, the knockout mouse has a phenotype, etc.), the overall error rate is not the  $\alpha$  level, but the type II error rate  $\beta$  (Schmidt and Hunter, 1997). It is impossible to falsely conclude that the null hypothesis is false, when in fact it is false! This apparently trivial insight may have important consequences for the interpretation of experimental data, and ultimately for its translation to the patient: If  $\beta$  is high (i.e., statistical power is low:  $\text{power} = 1 - \beta$ ), the probability to be able to reproduce data becomes small. For example, at a power of 0.5, which is not uncommon in experimental stroke research (see below), the probability to be able to replicate the findings of a

study is 50%, equivalent to throwing a coin (Mulaik *et al*, 1997)!

Is this relevant to experimental stroke studies? Many 'statistically significant' experimental stroke studies find reductions in infarct size of around 40%, have standard deviations of about 40% of the mean, group sizes of approximately 10 (van der Worp *et al*, 2005), and accept  $P < 0.05$  as statistically significant. In the above scenario, statistical power is 0.56! As no systematic review of power in experimental stroke studies exists, and to test whether this is a realistic estimate, I randomly sampled 30 preclinical as well as basic experimental stroke studies published in 2004 or 2005 in *Stroke* and the *Journal of Cerebral Blood Flow and Metabolism* and calculated *post hoc* their statistical power in analyzing infarct volumes. The mean  $\beta$  of this sample was 0.58 (95% confidence interval (CI) 0.49 to 0.63).

Unfortunately, lack of power is not the only statistical flaw found in some experimental stroke studies. A list of such defects includes: lack of correction for multiple comparisons; reporting of ordinal variables (e.g., from histologic or neuro-behavioral scores) with standard error of the mean (s.e.m.) or standard deviations (s.d.) (instead of median and range), and statistical comparisons of these values with a *t*-test (instead of a  $\chi^2$  test). The majority of experimental stroke studies describe numerical data as means  $\pm$  s.e.m.'s instead of s.d.'s. This promotes 'nice' graphical displays but obscures the (large) variance of the data. The s.e.m. is a measure of precision of an estimate of a population parameter. In other words, it tells us how accurately we can estimate the mean. However, what we usually want to be informed about is the variability of the observations, which is given by the s.d. Statisticians have fought for decades to ban the use of s.e.m.'s (and with less fervor, even of s.d.'s) for reporting numerical biomedical data, and instead advocated the use of CIs (Reichhardt and Gollob, 1997; Gardner and Altman, 1995).

Another prevalent statistical misconception is that *P*-values indicate biologic significance (Cohen, 1997; Sterne and Davey, 2001; Greenwald *et al*, 1996). This has contributed to the popularity of null-hypothesis testing in biomedical research, and may have sometimes clouded a more (neuro)biologically driven interpretation of the data.

## Reporting

That deficits in reporting can affect the outcome of a study has been shown by many studies (Sutton *et al*, 2000; Chan and Altman, 2005*a, b*; Chan *et al*, 2004; Kyzas *et al*, 2005). Interestingly, no clear standards exist for the reporting of experimental data. This is in sharp contrast to the publication of RCTs, where international guidelines exist (CONSORT; Moher *et al*, 2001): failure to adhere to these guidelines leads

to automatic rejection of a manuscript. Potentially as a consequence of missing standards, the following deficiencies can be found in some experimental stroke studies: failure to report mortality, no numerical values given (only graphs), (sub)strain information lacking, no information on sponsorship, conflict of interest, etc. In the early days of research with genetically modified animals, issues of genetic background (e.g., flanking gene problem or genetic drift; Wolfer *et al*, 2002) were widely discussed (Picciotto and Wickman, 1998; Majzoub and Muglia, 1996; Aguzzi *et al*, 1996), and details on strains, source, and generation of animals, etc., were reported in sufficient detail in most publications. Today, however, crucial information is missing in many articles, and it is not rare that even the source and background of the genetically modified animal is not reported.

## Negative Publication Bias

Studies that do not show an effect of a compound on outcome after a stroke, or find no stroke phenotype of a mouse with a null mutation, are considered 'negative'. This label already reveals the negative bias towards such study outcomes. This results in the practical problem that it is very hard to publish negative study results. The so-called 'negative publication bias' is well known, and leads to a skewed view on the effects of drugs and mechanisms alike: conflicting, that is, negative findings, may exist, but they are not accessible to the scientific community. Negative publication bias can therefore strongly affect the overall quality of research in a given field. An effect of publication bias in experimental stroke research has already been shown (Macleod *et al*, 2004). Even in clinical research, where the publication of negative results is easier and more common than in experimental research, negative publication bias exists, and its negative consequences have been intensely documented (Veitch, 2005; Hayashino *et al*, 2005; Demaria, 2004; Phillips, 2004).

## Study Interpretation: Correlation Does Not Establish Causality

Experimental stroke researchers apply mechanistic manipulations (pharmacology, genetically modified animals) to modify infarct volume. With a reduction in the amount of damaged brain tissue, secondary pathophysiologic events are also reduced. In this situation, it is hard to discriminate whether a mechanism that covaries with lesion size, is a causal agent of tissue damage, or an epiphenomenon. A good example for this problem (which is a pitfall in the interpretation of study data in any scientific field) is the investigation into the effects of mechanisms and compounds which, among other effects,

may reduce stroke-induced inflammation (Emerich *et al*, 2002).

## How Can We Improve Quality In Experimental Research?

For the individual researcher planning and conducting experimental stroke studies, it is easy to improve quality. Many of the quality problems discussed above are easily overcome, particularly the one's listed under study design, reporting, and study interpretation (overview in Table 3). In theory, it is also easy to deal with the *statistical power* issue: a biologically relevant effect size should be set based on pathophysiologic and clinical reasoning. Then error levels for  $\alpha$  and  $\beta$  are chosen (e.g., 0.05 and 0.2). By estimating the variance of the data, the necessary number of animals for this design can then be calculated. To reiterate, this should be performed *a priori*, that is, before starting the study. As outcome data of experimental stroke studies often have a high variance, the numbers of animals needed to achieve acceptable  $\beta$ -values may necessitate group sizes larger than what is common practice. For the typical stroke study mentioned above (effect size and standard deviation around 40% of the mean), *a priori* sample size calculation for  $\alpha < 0.05$  and  $\beta < 0.2$  suggests a total  $n$  of 34 (17 control, 17 treatment). It should be stressed here, however, that power calculations should play no role once the data has been collected (Goodman and Berlin, 1994). Instead, 'null' results should be presented using CIs (see also below) around parameters such as the differences between means and effect sizes. This allows for conclusions about how far the population parameter could reasonably deviate from the value in the null hypothesis (Aberson, 2002).

In addition to using appropriately powered designs, replication of results should be considered (Dirnagl *et al*, 1990) before using data as a foundation for an extensive mechanistic investigation of the effect, which may not be reproducible, or the planning of a clinical study based on the experimental results.

Although beyond the scope of this article, I strongly believe that a more profound discussion of the use of statistics would benefit the field. This discussion could lead to a more critical and responsible use of null hypothesis testing and classical ('frequentist') statistics (Cohen, 1997; Greenwald *et al*, 1996), and potentially to the introduction of Bayesian approaches in the planning and evaluation of experimental stroke studies (O'Hagan and Luce, 2003). For a specific data set and specific hypothesis (not just the null hypothesis), Bayesian statistics allows the computation of the probability that the hypothesis is true, and it specifies the probability that the true value of the parameter is within any given interval. Bayesian

statistics provides an easy way to revise estimates in an orderly and defensible manner as one collects new data. This way of thinking and computing frees the investigator from some of the concerns about disclosing data before the planned end of the trial, and thereby provides a set of tools to dynamically optimize the number of animal needed with optional stopping rules (O'Hagan and Luce, 2003). As a result of these advantages, Bayesian study designs have already entered the stage in clinical stroke trials (Krams *et al*, 2005; Lees, 2001).

## The Case for Reporting Inclusion and Exclusion Criteria, as well as Mortality

I propose that experimental stroke studies evaluating treatment regimens should not only report criteria for inclusion and exclusion of animals but also for mortality. In particular, preclinical studies, which affect the decision to further develop a compound in clinical trials (at least supplementing conventional analysis), an *intention-to-treat analysis* should be considered, which is the gold standard in RCTs (Fergusson *et al*, 2002). In this type of analysis ('Analyze as randomized!'), animals that do not reach a predefined end point are scored with the worst possible outcome (e.g., an infarcted hemisphere).

## The Case for Systematic Reviews in Preclinical Stroke Research

Systematic reviews are the foundation of evidence-based medicine and therefore common practice in clinical research. Only recently some authors have started to conduct systematic reviews of experimental stroke studies. As already mentioned above, quality deficiencies in experimental stroke research were exposed in these reviews. Interestingly, some (if not all!) clinical trials that resulted in negative outcomes (e.g., nimodipine) were, contrary to common belief, not supported by sufficient evidence from animal experiments (Horn *et al*, 2001; Horn and Limburg, 2001; Jonas *et al*, 1997). Presently, systematic reviews of experimental stroke studies exist on nimodipine (Horn *et al*, 2001), FK506 (Macleod *et al*, 2005c), nicotinamide (Macleod *et al*, 2004), melatonin (Macleod *et al*, 2005b), nitric oxide donors (Willmot *et al*, 2005b), nitric oxide synthase inhibitors (Willmot *et al*, 2005a), glutamate release inhibitors (Nava-Ocampo *et al*, 2000), and diffusion-weighted imaging (Rivers and Wardlaw, 2005). Systematic reviews on thrombolysis (C Briscoe, E Sena, PAG Sandercock and MR Macleod, personal communication, 2005), piracetam (PCR Wheble, ES Sena, PAG Sandercock and MR Macleod, personal communication, 2005), tirilazad (ES Sena, PCR Wheble, PAG Sandercock

and MR Macleod, personal communication, 2005), are in submission.

Systematic reviews are an important tool to improve quality in experimental stroke research (Sandercock and Roberts, 2002; Macleod *et al*, 2005a, d; Roberts *et al*, 2002; van der Worp *et al*, 2005), as they not only, at least partially, overcome the problem of small sample sizes, but they can also deliver quantitative estimates for such confounders as publication bias (e.g., by funnel plot; Sutton *et al*, 2000) or effect of industry sponsorship on outcome. The CAMARADES (Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Stroke) collaboration, an initiative of Clinical Neurosciences, University of Edinburgh, UK and the National Stroke Research Institute, Melbourne, Australia, is spearheading the introduction of this approach into experimental stroke research. CAMARADES provides a supporting framework for groups involved in the systematic review and meta-analysis of data from animal studies in experimental stroke. Their website (URL see Table 4) features background information and practical tools to conduct such reviews.

Whether systematic reviews should also play an important role in safeguarding high standards of experimentation in mechanism-driven experimental stroke research remains unclear at present. Recently, meta-analysis has been applied successfully to assess the information available, which correlates histologic features with lesion appearance in diffusion-weighted imaging (Rivers and Wardlaw, 2005). In other neuroscience fields, meta analysis has been already applied to exploratory experimental studies (e.g., Myhrer, 2003).

### **The Case for Standard Operating Procedures in Preclinical Stroke Research**

Monitoring, auditing, and standard operating procedures (SOPs) are key elements of quality control in RCTs. I propose that stroke laboratories should set up and publish (e.g., on institutional websites) their SOPs, and guarantee that their studies adhere to these standards. This is all the more important, as a relevant fraction of experiments, evaluations, etc. in experimental stroke studies are not performed by professionals, but rather by students in training unaware of these issues. International professional societies, such as the International Society for Cerebral Blood Flow and Metabolism (ISCBFM), may help to establish standards, propose schemes to safeguard adherence to them, and may even certify laboratories. In its most comprehensive form, certification would involve the participation in interlaboratory ('round robin') tests, in which a known neuroprotectant (and/or placebo) is sent to participating laboratories in a masked manner. For

certification, the results would have to lie within certain values. Most preclinical laboratories of the pharmaceutical industry use SOPs, and test novel compounds together with known neuroprotectants (very often MK-801) in a masked manner. Under the auspices of professional societies, stroke research laboratories could also audit each other ('peer audit'). In any case, I propose that scientists and professional societies should proactively discuss issues of research governance.

### **The Case for Publishing Negative Results and a Trial Register**

Negative results in experimental stroke research should be made accessible to other researchers after peer review. This could be done via dedicated journals (e.g., the open-access publication *Journal of Negative Results in Biomedicine*; URL see Table 4), or special repositories of websites of scholarly journals in the field. In addition, it might be worthwhile to consider a trial database for ongoing experimental studies (such as the International Standard Randomized Controlled Trial Number Register for Clinical Trials; URL see Table 4). Almost complete coverage of all ongoing trials and their negative or positive results once finished could be guaranteed if journals were to require registration before study onset when a study is submitted (as is the case for clinical studies in top journals). The running costs of such a database could be provided by registration fees, whereas the implementation of the database might be financed by grants from the National Institute of Neurological Disorders and Stroke (NINDS) or the European Union.

### **An Important Role for the Editors of Scholarly Journals**

Editors of major clinical journals have spearheaded the recent improvements in the quality of clinical trials. This success story could be repeated in the experimental research domain if editors would only accept articles that (1) adhere to certain standards for reporting and statistics, (2) are internationally registered, and (3) lay open sponsorship or conflicts of interest.

### **A Gold Standard for Experimental Stroke Research?**

Using multidisciplinary approaches, *in vivo* and *in vitro* experimental stroke researchers investigate basic mechanisms of stroke injury. However, experimental stroke studies may also test treatment or diagnostic strategies before a decision is made whether clinical trials should be undertaken. Thus,

there can be no gold standard concerning models, species, or outcome assessment strategies in experimental stroke research. As recommended by STAIR (Stroke Therapy Academic Industry Roundtable, 1999), in preclinical stroke research animals with comorbidities, mixed gender, advanced age, etc. should play an important role to model the complexities of risk factors, patient profiles, and clinical situation as much as possible. The 'quality' of such studies, for example, if measured by a score (see above) would critically depend on how successfully clinical relevance was implemented in the research. Clearly, this would not be viable in reductionist mechanistic stroke research, which depends on excluding confounders, and therefore has no reason to cope with the consequences of increased variability, high numbers of required animals, and increased cost of long-term studies. A similar reasoning applies to effect sizes. For example, whether a 10% or a 30% reduction in infarct size is used for *a priori* power calculations depends entirely on the hypothesis to be tested. I suggest that authors, in particular if they report 'negative' results, justify their choice of effect size threshold.

It was the aim of this article to discuss some potential weaknesses (and therefore sources of bias) in the design, evaluation, and reporting of published experimental studies in stroke research. Where available I have quoted quantitative data on the frequency and impact of such weaknesses. I have also included, and clearly labeled, quality-related sources of bias, which have not been systematically studied (e.g., lack of inclusion/exclusion criteria; low statistical power). I have done this to spur discussion and in the hope that they will be exposed to meta-analysis in the near future. I am aware of the fact that until then it has to remain open whether these factors truly impact on effect sizes and reproducibility. However, to fully harness the power of meta-analysis, we need to improve the standard of reporting of results in this field, which is the reason why I have particularly emphasized this issue. From an analysis of potential weaknesses, I have deduced a catalog of measures to improve or implement quality standards. The use of animals for studies with questionable predictive value is not justifiable. Equally important, patients may be put at risk when clinical trials are based on weak or flawed experimental evidence. I propose that by a number of rather simple measures, the step from bench to bedside in stroke research can be made more successful. However, the ultimate proof for this can only come from successful phase III stroke trials, which built on basic research conforming to the criteria as put forward in this article or by STAIR (Stroke Therapy Academic Industry Roundtable, 1999).

## Note added in proof

Since this review was accepted for publication, an important study (O'Collins *et al.*, 2006) was

published, which systematically evaluated experimental studies on 1,026 putative neuroprotectants and concludes that the drugs taken forward to clinical trial have not been distinguished by superior efficacy. The study urges for greater rigor in the conduct and reporting of animal data in experimental stroke research.

## References

- Aberson C (2002) Interpreting null results—improving presentation and conclusions with confidence intervals. *J Articles Support Null Hypothesis* 1:36–42
- Aguzzi A, Brandner S, Marino S, Steinbach JP (1996) Transgenic and knockout mice in the study of neurodegenerative diseases. *J Mol Med* 74:111–26
- Alberts MJ, Latchaw RE, Selman WR, Shephard T, Hadley MN, Brass LM, Koroshetz W, Marler JR, Booss J, Zorowitz RD, Croft JB, Magnis E, Mulligan D, Jagoda A, O'Connor R, Cawley CM, Connors JJ, Rose-DeRenzy JA, Emr M, Warren M, Walker MD (2005) Recommendations for comprehensive stroke centers—a consensus statement from the Brain Attack Coalition. *Stroke* 36:1597–616
- American Physiological Society (2005) Guidelines for reporting statistics in journals published by the American Physiological Society. *Am J Physiol* 287:R247–9
- Astrup J, Symon L, Branston NM, Lassen NA (1977) Cortical evoked potential and extracellular K<sup>+</sup> and H<sup>+</sup> at critical levels of brain ischemia. *Stroke* 8:51–7
- Bebarta V, Luyten D, Heard K (2003) Emergency medicine animal research—does use of randomization and blinding affect the results? *Acad Emerg Med* 10:684–7
- Buchan A, Pulsinelli WA (1990) Hypothermia but not the N-methyl-D-aspartate antagonist, MK-801, attenuates neuronal damage in gerbils subjected to transient global ischemia. *J Neurosci* 10:311–6
- Carmichael ST (2003) Gene expression changes after focal stroke, traumatic brain and spinal cord injuries. *Curr Opin Neurol* 16:699–704
- Chan AW, Altman DG (2005a) Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 365:1159–62
- Chan AW, Altman DG (2005b) Identifying outcome reporting bias in randomised trials on PubMed—review of publications and survey of authors. *BMJ* 330:753
- Chan AW, Krleza-Jeric K, Schmid I, Altman DG (2004) Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian Medical Association Journal* 171:735–40
- Chen R, Cohen LG, Hallett M (2002) Nervous system reorganization following injury. *Neuroscience* 111:761–73
- Cheng YD, Al Khoury L, Zivin JA (2004) Neuroprotection for ischemic stroke—two decades of success and failure. *NeuroRx* 1:36–45
- Chopp M, Li Y (2002) Treatment of neural injury with marrow stromal cells. *Lancet Neurol* 1:92–100
- Clarke R (1997) Issues in experimental design and endpoint analysis in the study of experimental cytotoxic agents *in vivo* in breast cancer and other models. *Breast Cancer Res Treat* 46:255–78
- Cohen J (1997) The earth is round ( $P < 0.05$ ). In: *What if there were no significance tests?* (Harlow L, Mulaik SA,

- Steiger JH, eds), London, Mahwah, NJ: Lawrence Erlbaum Associates, 21–35
- Curry SH (2003) Why have so many drugs with stellar results in laboratory stroke models failed in clinical trials? A theory based on allometric relationships. *Ann NY Acad Sci* 993:69–74
- Danton GH, Dietrich WD (2004) The search for neuroprotective strategies in stroke. *Am J Neuroradiol* 25:181–94
- DeGraba TJ, Pettigrew LC (2000) Why do neuroprotective drugs work in animals but not humans? *Neurol Clin* 18:475–93
- del Zoppo GJ (1998) Clinical trials in acute stroke—why have they not been successful? *Neurology* 51:S59–61
- del Zoppo GJ (2004) Thrombolysis—from the experimental findings to the clinical practice. *Cerebrovasc Dis* 17(Suppl 1):144–52
- Demaria AN (2004) Publication bias and journals as policemen. *J Am Coll Cardiol* 44:1707–8
- Dijkhuizen RM, Nicolay K (2003) Magnetic resonance imaging in experimental models of brain disorders. *J Cereb Blood Flow Metab* 23:1383–402
- Dirnagl U, Iadecola C, Moskowitz MA (1999) Pathobiology of ischaemic stroke—an integrated view. *Trends Neurosci* 22:391–7
- Dirnagl U, Simon RP, Hallenbeck JM (2003) Ischemic tolerance and endogenous neuroprotection. *Trends Neurosci* 26:248–54
- Dirnagl U, Tanabe J, Pulsinelli W (1990) Pre- and post-treatment with MK-801 but not pretreatment alone reduces neocortical damage after focal cerebral ischemia in the rat. *Brain Res* 527:62–8
- Drummond AE, Pearson B, Lincoln NB, Berman P (2005) Ten year follow-up of a randomised controlled trial of care in a stroke rehabilitation unit. *BMJ* 331:491–2
- Emerich DF, Dean III RL, Bartus RT (2002) The role of leukocytes following cerebral ischemia—pathogenic variable or bystander reaction to emerging infarct? *Exp Neurol* 173:168–81
- Endres M (2005) Statins and stroke. *J Cereb Blood Flow Metab* 25:1093–110
- Endres M, Dirnagl U (2002) Ischemia and stroke. *Adv Exp Med Biol* 513:455–73
- Evenson KR, Rosamond WD, Morris DL (2001) Prehospital and in-hospital delays in acute stroke care. *Neuroepidemiology* 20:65–76
- Fergusson D, Aaron SD, Guyatt G, Hebert P (2002) Post-randomisation exclusions—the intention to treat principle and excluding patients from analysis. *BMJ* 325:652–4
- Festing MF (2003) Principles—the need for better experimental design. *Trends Pharmacol Sci* 24:341–5
- Festing MF (2006) Design and statistical methods in studies using animal models of development. *ILAR J* 47:5–14
- Fisher M (2004) The ischemic penumbra—identification, evolution and treatment concepts. *Cerebrovasc Dis* 17(Suppl 1):1–6
- Fisher M, Tatlisumak T (2005) Use of animal models has not contributed to development of acute stroke therapies—con. *Stroke* 36:2324–5
- Ford I (1983) Can statistics cause brain damage? *J Cereb Blood Flow Metab* 3:259–62
- Fukuda S, del Zoppo GJ (2003) Models of focal cerebral ischemia in the nonhuman primate. *ILAR J* 44:96–104
- Gardner MJ, Altman DG (1995) *Statistics with confidence—confidence intervals and statistical guidelines*. London: British Medical Journal
- Germano IM, Bartkowski HM, Cassel ME, Pitts LH (1987) The therapeutic value of nimodipine in experimental focal cerebral ischemia. Neurological outcome and histopathological findings. *J Neurosurg* 67:81–7
- Goodman SN, Berlin JA (1994) The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 121:200–6
- Green AR (2002) Why do neuroprotective drugs that are so promising in animals fail in the clinic? An industry perspective. *Clin Exp Pharmacol Physiol* 29:1030–4
- Greenwald AG, Gonzalez R, Harris RJ, Guthrie D (1996) Effect sizes and *p* values—what should be reported and what should be replicated? *Psychophysiology* 33:175–83
- Hacke W, Kaste M, Fieschi C, Toni D, Lesaffre E, von Kummer R, Boysen G, Bluhmki E, Hoxter G, Mahagne MH (1995) Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Study (ECASS). *JAMA* 274:1017–25
- Hayashino Y, Noguchi Y, Fukui T (2005) Systematic evaluation and comparison of statistical tests for publication bias. *J Epidemiol* 15:235–43
- Heard K, Bebera VS, Lowenstein SR (2005) Methodological standards in human versus animal clinical trials. *JAMA* 294:40–1
- Hill MD, Barber PA, Demchuk AM, Sevick RJ, Newcommon NJ, Green T, Buchan AM (2000) Building a 'brain attack' team to administer thrombolytic therapy for acute ischemic stroke. *CMAJ* 162:1589–93
- Hjort N, Butcher K, Davis SM, Kidwell CS, Koroshetz WJ, Rother J, Schellinger PD, Warach S, Ostergaard L (2005) Magnetic resonance imaging criteria for thrombolysis in acute cerebral infarct. *Stroke* 36:388–97
- Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M (2001) Nimodipine in animal model experiments of focal cerebral ischemia—a systematic review. *Stroke* 32:2433–8
- Horn J, Limburg M (2001) Calcium antagonists for ischemic stroke—a systematic review. *Stroke* 32:570–6
- Hossmann KA (2004) Genetically modified animals in molecular stroke research. *Acta Neurochir Suppl* 89:37–45
- Hoyte L, Kaur J, Buchan AM (2004) Lost in translation—taking neuroprotection from animal models to clinical trials. *Exp Neurol* 188:200–4
- Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* 37:243–53
- Indredavik B, Bakke F, Slordahl SA, Rokseth R, Haheim LL (1999) Stroke unit treatment. 10-year follow-up. *Stroke* 30:1524–7
- Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2:e124
- Johnson PD, Besselsen DG (2002) Practical aspects of experimental design in animal research. *ILAR J* 43:202–6
- Jonas S, Tran AQ, Eisenberg E, Azam M, Viera D, Grumet S (1997) Does effect of a neuroprotective agent on volume of experimental animal cerebral infarct predict effect of the agent on clinical outcome in human stroke? *Ann NY Acad Sci* 825:281–7:281–7

- Kaste M (2005) Use of animal models has not contributed to development of acute stroke therapies—pro. *Stroke* 36:2323–4
- Kidwell CS, Alger JR, Saver JL (2003) Beyond mismatch—evolving paradigms in imaging the ischemic penumbra with multimodal magnetic resonance imaging. *Stroke* 34:2729–35
- Krams M, Lees KR, Berry DA (2005) The past is the future—innovative designs in acute stroke therapy trials. *Stroke* 36:1341–7
- Kyzas PA, Loizou KT, Ioannidis JP (2005) Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* 97:1043–55
- Langhorne P, Dennis MS (2004) Stroke units—the next 10 years. *Lancet* 363:834–5
- Lee DS, Nguyen QT, Lapointe N, Austin PC, Ohlsson A, Tu JV, Stewart DJ, Rouleau JL (2003) Meta-analysis of the effects of endothelin receptor blockade on survival in experimental heart failure. *J Cardiac Fail* 9:368–74
- Lees KR (2001) Advances in neuroprotection trials. *Eur Neurol* 45:6–10
- Lees KR, Hankey GJ, Hacke W (2003) Design of future acute-stroke treatment trials. *Lancet Neurol* 2:54–61
- Leker RR, Constantini S (2002) Experimental models in focal cerebral ischemia—are we there yet? *Acta Neurochir Suppl* 83:55–59:55–9
- Liang D, Dawson TM, Dawson VL (2004) What have genetically engineered mice taught us about ischemic injury? *Curr Mol Med* 4:207–25
- Lindvall O, Kokaia Z (2004) Recovery and rehabilitation in stroke—stem cells. *Stroke* 35:2691–4
- Lo EH, Dalkara T, Moskowitz MA (2003) Mechanisms, challenges and opportunities in stroke. *Nat Rev Neurosci* 4:399–415
- Lo EH, Moskowitz MA, Jacobs TP (2005) Exciting, radical, suicidal—how brain cells die after stroke. *Stroke* 36:189–92
- Macleod MR, Ebrahim S, Roberts I (2005a) Surveying the literature from animal experiments—systematic review and meta-analysis are important contributions. *BMJ* 331:110
- Macleod MR, O'Collins T, Horkey LL, Howells DW, Donnan GA (2005b) Systematic review and meta-analysis of the efficacy of melatonin in experimental stroke. *J Pineal Res* 38:35–41
- Macleod MR, O'Collins T, Horkey LL, Howells DW, Donnan GA (2005c) Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. *J Cereb Blood Flow Metab* 25:713–21
- Macleod MR, O'Collins T, Howells DW, Donnan GA (2004) Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke* 35:1203–8
- Macleod MR, Sena PC, Wheble PCR, O'Collins T, Horkey LL, van der Worp HB, Howells DW, Sandercock P, Donnan GA (2005d) Sources of bias in animal models of neurological disease. *Abstract Viewer/Itinerary Planner* Washington, DC: Society for Neuroscience, 2005 Online, Program No. 702.1
- Majzoub JA, Muglia LJ (1996) Knockout mice. *N Engl J Med* 334:904–7
- Meisel C, Schwab JM, Prass K, Meisel A, Dirnagl U (2005) Central nervous system injury-induced immune deficiency syndrome. *Nat Rev Neurosci* 6:775–86
- Mergenthaler P, Dirnagl U, Meisel A (2004) Pathophysiology of stroke—lessons from animal models. *Metab Brain Dis* 19:151–67
- Moher D, Schulz KF, Altman DG (2001) The CONSORT statement—revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 357:1191–4
- Mulaik SA, Raju NS, Harshman RA (1997) There is a time and a place for significance testing. In: *What if there were no significance tests?* (Harlow L, Mulaik SA, Steiger JH, eds), London, Mahwah, NJ: Lawrence Erlbaum Associates, 66–115
- Myhrer T (2003) Neurotransmitter systems involved in learning and memory in the rat—a meta-analysis based on studies of four behavioral tasks. *Brain Res Brain Res Rev* 41:268–87
- Nava-Ocampo AA, Reyes-Perez H, Bello-Ramirez AM, Mansilla-Olivares A, Ponce-Monter H (2000) For ischemic brain damage, is preclinical evidence of neuroprotection by presynaptic blockade of glutamate release enough? *Med Hypotheses* 54:77–9
- O'Collins VE, Macleod MR, Donnan GA, Horkey LL, Van der Worp BH, Howells DW (2006) 1,026 Experimental treatments in acute stroke. *Ann Neurol* (Epub ahead of print)
- O'Hagan A, Luce BR (2003) *A primer on Bayesian Statistics in Health Economics. The Bayesian Initiative in Health Economics & Outcomes Research*. Electronic publication URL <http://www.bayesian-initiative.com>
- Phillips CV (2004) Publication bias *in situ*. *BMC Med Res Methodol* 4:20:20
- Piccioito MR, Wickman K (1998) Using knockout and transgenic mice to study neurophysiology and behavior. *Physiol Rev* 78:1131–63
- Pound P, Ebrahim S, Sandercock P, Bracken MB, Roberts I (2004) Where is the evidence that animal research benefits humans? *BMJ* 328:514–7
- Reichardt CS, Gollob HF (1997) When confidence intervals should be used instead of statistical tests, and vice versa. In: *What if there were no significance tests?* (Harlow L, Mulaik SA, Steiger JH, eds), London, Mahwah, NJ: Lawrence Erlbaum Associates, 259–83
- Richard GA, Odegren T, Ashwood T (2003) Animal models of stroke—do they have value for discovering neuroprotective agents? *Trends Pharmacol Sci* 24:402–8
- Rivers CS, Wardlaw JM (2005) What has diffusion imaging in animals told us about diffusion imaging in patients with ischaemic stroke? *Cerebrovasc Dis* 19:328–36
- Roberts I, Kwan I, Evans P, Haig S (2002) Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. *BMJ* 324:474–6
- Roitberg B (2004) Transplantation for stroke. *Neurol Res* 26:256–64
- Samsa GP, Matchar DB (2001) Have randomized controlled trials of neuroprotective drugs been underpowered? An illustration of three statistical principles. *Stroke* 32:669–74
- Sandercock P, Roberts I (2002) Systematic reviews of animal experiments. *Lancet* 360:586
- Savitz SI, Dinsmore JH, Wechsler LR, Rosenbaum DM, Caplan LR (2004) Cell therapy for stroke. *NeuroRx* 1:406–14
- Savitz SI, Rosenbaum DM, Dinsmore JH, Wechsler LR, Caplan LR (2002) Cell transplantation for stroke. *Ann Neurol* 52:266–75
- Schmidt FL, Hunter JE (1997) Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: *What if there were*

- no significance tests? (Harlow L, Mulaik SA, Steiger JH, eds), London, Mahwah, NJ: Lawrence Erlbaum Associates, 37–64
- Sterne JA, Davey SG (2001) Sifting the evidence—what's wrong with significance tests? *BMJ* 322:226–31
- Stroke Therapy Academic Industry Roundtable (1999) Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 30:2752–8
- Stroke Therapy Academic Industry Roundtable (2001) Recommendations for clinical trial evaluation of acute stroke therapies. *Stroke* 32:1598–606
- Stroke Unit Trialists' Collaboration (1997) Collaborative systematic review of the randomised trials of organised inpatient (stroke unit) care after stroke. *BMJ* 314:1151–9
- Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR (2000) Empirical assessment of effect of publication bias on meta-analyses. *BMJ* 320:1574–7
- Symon L, Crockard HA, Dorsch NW, Branston NM, Juhasz J (1975) Local cerebral blood flow and vascular reactivity in a chronic stable stroke in baboons. *Stroke* 6:482–92
- The National Institute of Neurological Disorders Stroke rt-PA Stroke Study Group (1995) Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med* 333:1581–7
- Traystman RJ (2003) Animal models of focal and global cerebral ischemia. *ILAR J* 44:85–95
- van der Worp HB, de Haan P, Morrema E, Kalkman CJ (2005) Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. *J Neurol* 252:1108–14
- Veitch E (2005) Tackling publication bias in clinical trial reporting. *PLoS Med* 2:e367
- Wahlgren NG, Ahmed N (2004) Neuroprotection in cerebral ischaemia—facts and fancies—the need for new approaches. *Cerebrovasc Dis* 17(Suppl 1): 153–66
- Warach S (2003) Stroke neuroimaging. *Stroke* 34:345–7
- Ward NS, Cohen LG (2004) Mechanisms underlying recovery of motor function after stroke. *Arch Neurol* 61:1844–8
- Willmot M, Gibson C, Gray L, Murphy S, Bath P (2005a) Nitric oxide synthase inhibitors in experimental ischemic stroke and their effects on infarct size and cerebral blood flow—a systematic review. *Free Radical Biol Med* 39:412–25
- Willmot M, Gray L, Gibson C, Murphy S, Bath PM (2005b) A systematic review of nitric oxide donors and L-arginine in experimental stroke; effects on infarct size and cerebral blood flow. *Nitric Oxide* 12:141–9
- Wolfer DP, Crusio WE, Lipp HP (2002) Knockout mice—simple solutions to the problems of genetic background and flanking genes. *Trends Neurosci* 25:336–40
- Xu Z, Ford GD, Croslan DR, Jiang J, Gates A, Allen R, Ford BD (2005) Neuroprotection by neuregulin-1 following focal stroke is associated with the attenuation of ischemia-induced pro-inflammatory and stress gene expression. *Neurobiol Dis* 19:461–70
- Zivin JA, Fisher M, DeGirolami U, Hemenway CC, Stashak JA (1985) Tissue plasminogen activator reduces neurological damage after cerebral embolism. *Science* 230:1289–92