Journal of Cerebral Blood Flow & Metabolism (2011) 31, 1064–1072

www.jcbfm.com

JCBFM Open

Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the *Journal of Cerebral Blood Flow and Metabolism*

Hanna V Vesterinen[1], Kieren Egan[1], Amelie Deister[2], Peter Schlattmann[3], Malcolm R Macleod[1,4] and Ulrich Dirnagl[2]

[1]*Department of Clinical Neurosciences, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, Scotland, UK;* [2]*Departments of Neurology and Experimental Neurology, Center for Stroke Research, Charité University Medicine Berlin, Berlin, Germany;* [3]*Department of Medical Statistics, Computer Sciences and Documentation, University Hospital of Friedrich-Schiller-University Jena, Jena, Germany;* [4]*Department of Neurology, NHS Forth Valley, Stirling, Scotland, UK*

Translating experimental findings into clinically effective therapies is one of the major bottlenecks of modern medicine. As this has been particularly true for cerebrovascular research, attention has turned to the quality and validity of experimental cerebrovascular studies. We set out to assess the study design, statistical analyses, and reporting of cerebrovascular research. We assessed all original articles published in the *Journal of Cerebral Blood Flow and Metabolism* during the year 2008 against a checklist designed to capture the key attributes relating to study design, statistical analyses, and reporting. A total of 156 original publications were included (animal, *in vitro*, human). Few studies reported a primary research hypothesis, statement of purpose, or measures to safeguard internal validity (such as randomization, blinding, exclusion or inclusion criteria). Many studies lacked sufficient information regarding methods and results to form a reasonable judgment about their validity. In nearly 20% of studies, statistical tests were either not appropriate or information to allow assessment of appropriateness was lacking. This study identifies a number of factors that should be addressed if the quality of research in basic and translational biomedicine is to be improved. We support the widespread implementation of the ARRIVE (Animal Research Reporting *In Vivo* Experiments) statement for the reporting of experimental studies in biomedicine, for improving training in proper study design and analysis, and that reviewers and editors adopt a more constructively critical approach in the assessment of manuscripts for publication.
Journal of Cerebral Blood Flow & Metabolism (2011) **31,** 1064–1072; doi:10.1038/jcbfm.2010.217; published online 15 December 2010

Keywords: ARRIVE; bias; CONSORT; quality; translation; validity

Correspondence: Professor Dr U Dirnagl, Departments of Neurology and Experimental Neurology, Center for Stroke Research Berlin, Charité University Medicine, Berlin 10098, Germany.
E-mail: ulrich.dirnagl@charite.de
The work of PS is supported by the Deutsche Forschungsgemeinschaft DFG (Schl 3-1). HV is supported by a University of Edinburgh CCBS PhD studentship. KE is supported by an MRC Edinburgh Trials Methodology Hub studentship. MM receives funding from the Scottish Chief Scientist's Office, the UK Stroke Association, Chest Heart and Stroke (Scotland), and the MS Society, and is supported by the MRC Trials Methodology Hub. The work of UD is supported by the European Union's Seventh Framework Programme (FP7/2008-2013) under grant agreements no. 201024 and no. 202213 (European Stroke Network), and the German Ministry for Health and Education (BMBF), as well as the Deutsche Forschungsgemeinschaft (Excellence Cluster NeuroCure).
Received 30 September 2010; revised 25 October 2010; accepted 3 November 2010; published online 15 December 2010

Introduction

Translating experimental findings into clinically effective therapies is one of the major bottlenecks of modern medicine. That is, bench findings rarely lead to bedside treatments. This 'translational roadblock' is particularly evident in the cerebrovascular research field, in which despite numerous promising preclinical trials, only few treatments of proven efficacy are available (Dirnagl, 2006; O'Collins *et al*, 2006; Macleod *et al*, 2009). As little has been known about the quality of preclinical studies (Dirnagl, 2006; O'Collins *et al*, 2009), some have suggested that low study quality at various stages in the research process might have reduced the internal validity of experimental studies (Dirnagl, 2006; Zinsmeister and Connor, 2008).

An extensive literature has accumulated with a primary focus of quality assessment of clinical trials, specifically assessing study design, statistical analysis, and trial reporting (Altman, 1998, 2002; Sarter and Fritschy, 2008; Glantz, 1980). This paved the way for establishment of standards for conducting and reporting clinical trials—namely the initialization and implementation of practices such as Cochrane (http://www.cochrane.org), CONsolidated Standards of Reporting Trials (CONSORT) (http://www.consort-statement.org), web-based trial databases (e.g., http://www.clinicaltrials.gov, http://www.controlled-trials.com), and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) (http://www.strobe-statement.org) (Bellolio *et al*, 2008; Moher *et al*, 2001*a*, *b*). These measures have vastly improved the validity of clinical trials and ultimately their impact on patients (Moher *et al*, 2001*a*). However, in the translational preclinical realm, and particularly in translational cerebrovascular medicine, such approaches have been advocated more recently, e.g., Stroke Academia Industry Roundtable (STAIR) (http://www.thestair.org) and Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Stroke (CAMARADES) (http://www.camarades.info) (Dirnagl, 2006; Macleod *et al*, 2009).

Despite the lessons learned through clinical trial quality assessment, few studies have formally investigated such factors in the basic science and translational preclinical trial realms (Dirnagl, 2006). Those that have, included publications from more than one scientific journal and assessed publications across various research fields (Kilkenny *et al*, 2009; Ioannidis, 2005; Nieminen *et al*, 2006; Schroter *et al*, 2008). To our knowledge, no systematic investigation focusing on study design, statistical analysis, and reporting in the field of translational cerebrovascular research has been carried out.

The purpose of this study was to assess the *status quo* of scientific reporting, experimental design, and statistical analysis in the field of experimental cerebrovascular medicine. In a first attempt, we focused on the *Journal of Cerebral Blood Flow and Metabolism* (JCBFM), one of the leading journals in the field of experimental cerebrovascular research.

We chose JCBFM, the official journal of the International Society for Cerebral Blood Flow and Metabolism, as it stands at the interface between basic and clinical neurovascular research. It features timely and relevant high-quality research while highlighting experimental, theoretical, and clinical aspects of brain circulation, metabolism, and imaging (http://www.nature.com/jcbfm). According to the 2009 Journal Citation Report, the journal has an impact factor of 5.457, and ranks 29th of 230 journals in neuroscience, 14th of 105 journals in endocrinology and metabolism, and 10th out of 61 in hematology (Thomson Reuters, 2010). As an internationally esteemed source of information in translational cerebrovascular medicine, JCBFM is well suited to survey reporting and quality aspects in this field; our prediction was that, because JCBFM ratings are high and it is a leading journal in the field, the quality of experimental studies (based on design, statistics, and reporting) would also be very high.

Methods

Publication Searches

To assess the quality of research reporting, experimental design, and statistical analysis in all original articles published in JCBFM from January 2008 through December 2008, a full version of JCBFM volume 28 (numbers 1 to 12) was secured in both electronic and hard copy format. A nonblinded reviewer (HV) then systematically identified original scientific publications within JCBFM volume 28.

Original publications were categorized according to study type: animal (including rodents, primates, canines, and birds), *in vitro*, or human studies, review articles, commentaries, communications, errata, and corrigenda.

Inclusion and Exclusion Criteria for Quality Assessment

All original articles published in JCBFM volume 28 were eligible for inclusion. Review articles (8), commentaries or communications (21), or errata and corrigenda (8) were excluded.

Publication Quality Assessment

Questionnaire: We developed a checklist questionnaire to capture the key aspects of the reporting of (1) experimental design, (2) experimental analysis and statistics, and (3) the overall quality of reporting. Such checklists for reporting standards are commonly used in other research domains (particularly in clinical trials), and we began by creating a catalog of possible checklist items from publications in these other domains. Some items thus identified were clearly not relevant to original articles in JCBFM, and after exclusion of these, we selected 15 main items and 9 supplementary items which, in our view, captured most of the important aspects of study reporting, which might reasonably be expected from publications in JCBFM (see Table 1).

Assessment Process

For all items apart from the reporting of specific test statistics (question 8) and the appropriateness of the statistical tests used (question 8a), two nonblinded reviewers (KE and HV) independently extracted data for publication quality by reading each original article and then reporting whether the specific checklist item was met by the publication (yes), not met by the publication (no), not applicable to that publication (n.a.), or unknown.

Table 1 Questionnaire used to assess the quality of publications

Category	*Question*
Design	
1	Was a primary/research hypothesis stated?
1a	Was an aim/purpose of study stated?
2	Was the design randomized? (Dirnagl, 2006)
3	Was allocation concealed?
4	Was outcome assessed blinded?
5	Was a statement about sample size given? (e.g., *a priori* power analysis) (Altman, 2002)
6	Was study design stated? (Andersen, 1990)
7	Were inclusion and exclusion criteria stated? (Altman *et al*, 1983; Andersen, 1990)
Analysis and statistics	
8	Were specific test statistics reported? (Altman, 2002)
8a	Were statistical tests appropriate for study design?
9	Was a measure of variance reported? (Andersen, 1990)
9a	Were s.d. reported?
9b	Were s.e.m. reported?
9c	Were confidence intervals reported?
10	Were the units of analysis specified? (Andersen, 1990; Altman, 2002)
10a	Were individual data points reported (e.g., plot)?
10b	Were raw data given?
Reporting	
11	Were numerical values only given in graphs? (regarding primary hypothesis/main experiment) (Altman, 1998)
12	Was mortality/number of dead quantified and stated? (Andersen, 1990)
13	Was the source of experimental organism/cells given? (species, strain, etc.)
13a	Was the laboratory/company stated where experimental organism was acquired from?
13b	Was the age of the experimental organism given?
13c	Was the weight of the experimental organism stated?
14	Was a control group reported? (Andersen, 1990)
15	Was a conflict of interest statement given?

Assessing Test Statistics

An independent, blinded expert reviewer (PS) assessed for the presence of specific test statistics (item 8) and whether the analytical test statistics were appropriate for the underlying experimental design (item 8a). In a statistical analysis, the method of choice depends on the type of data, e.g., numerical or categorical, as well as on the structure and distribution of data. Thus, it was investigated whether the chosen analysis was suitable for data at hand. For example, for a comparison of three groups with repeated continuous measurements, it was checked whether an appropriate method, such as an analysis of variance with repeated measures was chosen. To do so, statistical details reported in the selected publications were extracted and given to the reviewer alongside the publications. Moreover, the order in which publications were presented was randomized and the reviewer was blinded to authors, institution, journal, volume number, and digital object identifier number of the publications.

Interobserver Agreement

κ-Statistics, representing the extent of agreement between the two scorers, were calculated for each item, except for 8a. As κ is highly affected by the prevalence of *t*-positive scores, we also calculated separate indices of the proportionate agreement in the observers' positive ('yes') and negative ('no') decisions (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990).

Results

A total of 193 original scientific articles were published in JCBFM in the year 2008. Of these, 95 (49%) described animal studies, 49 (25%) described *in vitro* experiments, 34 (18%) included human participants, 8 (4%) were review articles, and 29 (15%) were of other types. Of the 193 total publications, 156 were original studies. Quality assessment of these studies was performed using the checklist presented in Table 1. The proportion of animal, cell culture, and human studies meeting each of the checklist items is given in Table 2, and is summarized in Figure 1.

Interobserver Agreement

To ensure appropriate robustness to the assessment, each item was determined by two investigators blinded to the assessment of the other. Interobserver agreement is given in Supplementary Table 1.

Discussion

Systematic assessment of all 156 original articles published in JCBFM in 2008 revealed a surprisingly high prevalence of deficiencies in the reporting of key components of scientific quality: design, reporting, and statistics. This is the first systematic study of this type in research on physiology and pathophysiology of brain metabolism and blood flow, but several studies and commentaries have already hinted that methodological and reporting problems are prevalent, and that this might be an important contributor to the 'translational roadblock' that exists in the field (Dirnagl, 2006; Sena *et al*, 2007; Phillips *et al*, 2008; Fisher *et al*, 2009; Crossley *et al*, 2008; Jerndal *et al*, 2010; Dirnagl and Macleod, 2009). Other investigators have found similar disappointing quality indicators when surveying publication practice of animal experiments in general (Kilkenny *et al*, 2009), medical research in general (Altman, 2002), and statistics in experimental and clinical medical papers (García-Berthou and Alcaraz, 2004; Williams *et al*, 1997; Holmes, 2004; Zinsmeister and Connor, 2008; Phillips *et al*, 2008; Hoffmann, 1984; Glantz, 1980).

Scientific publications are the main source of information in research. Original publications need 'to convey to the reader relevant information concerning the design, conduct, analysis, and generalization of the trial. This information should provide the reader with the ability to make informed judgments regarding the internal and external validity of the trial. Accurate and complete reporting also benefits editors and reviewers in their deliberations regarding submitted manuscripts.' (Begg *et al*, 1996). Ethical considerations regarding the use of animals in research and the well-being of patients dictate

Table 2 The total number of animal, *in vitro*, and human studies meeting each of the checklist items (yes, not applicable (n.a.), and unknown (U)) as assessed by both reviewers in relevant publications (those which were not n.a.) (summarized in Figure 1)

Category	*Question*	*Animal (n = 190)*			In vitro *(n = 98)*			*Human (n = 68)*			*Overall (n = 312)*		
		Yes	*n.a.*	*U*	*Yes*	*n.a.*	*U*	*Yes*	*n.a.*	*U*	*Yes*	*n.a.*	*U*
Design													
1	Primary/research hypothesis stated?	51 (27)	—	—	32 (33)	—	—	24 (35)	—	—	93 (30)	—	—
1a	Aim/purpose of study stated?	181 (95)	—	—	90 (92)	—	—	67 (99)	—	—	300 (96)	—	—
2	Randomization?	39 (22)	13 (7)	—	14 (15)	3 (3)	—	5 (8)	6 (9)	—	46 (15)	20 (6)	—
3	Allocation concealment?	14 (8)	6 (3)	—	2 (2)	3 (3)	—	3 (5)	3 (4)	—	17 (6)	10 (3)	—
4	Blinded assessment of outcome?	28 (15)	—	—	13 (13)	—	—	17 (25)	—	—	46 (15)	—	—
5	Statement about sample size given	2 (1)	1 (1)	—	2 (2)	1 (1)	—	0 (0)	—	—	2 (1)	1 (0)	—
6	Is study design stated?	6 (3)	—	—	2 (2)	—	—	9 (13)	—	—	16 (5)	—	—
7	Are inclusion and exclusion criteria stated?	26 (14)	—	—	11 (11)	—	—	29 (43)	—	—	58 (19)	—	—
Analysis and statistics													
8	Are specific test statistics reported?	84 (88)	—	—	36 (73)	—	—	31 (91)	—	—	132 (85)	—	—
8a	Are applied statistical tests appropriate for study design?	73 (77)	—	—	42 (88)	1 (2)	—	30 (88)	—	—	127 (81)	—	—
9	Measure of variance reported?	183 (97)	2 (1)	—	95 (99)	2 (2)	—	66 (99)	1 (1)	3 (4)	302 (97)	2 (1)	—
9a	s.d. reported?	86 (47)	8 (4)	—	36 (38)	2 (2)	14 (15)	44 (66)	1 (1)	3 (4)	150 (50)	9 (3)	25 (8)
9b	s.e.m. reported?	92 (51)	8 (4)	—	47 (49)	2 (2)	14 (15)	24 (36)	1 (1)	3 (4)	140 (46)	9 (3)	25 (8)
9c	Confidence interval (CI) reported?	5 (3)	8 (4)	—	1 (1)	2 (2)	14 (15)	5 (7)	—	—	10 (3)	9 (3)	25 (8)
10	Units of analysis given?	85 (34)	1 (1)	—	37 (38)	1 (1)	—	40 (62)	3 (4)	—	140 (46)	5 (2)	—
10a	Are individual data points reported	24 (13)	—	—	16 (16)	—	—	22 (32)	—	—	59 (19)	—	—
10b	Are raw data given?	13 (7)	—	—	11 (11)	—	—	13 (19)	—	—	34 (11)	—	—
Reporting													
11	Numerical values only given in graphs?	110 (59)	2 (1)	—	50 (52)	2 (2)	—	28 (41)	—	—	163 (53)	2 (1)	—
12	Mortality/number of dead stated?	16 (9)	19 (10)	—	6 (8)	24 (24)	—	4 (8)	19 (28)	—	22 (8)	53 (17)	—
13	Source of experimental organism/cells given?	158 (92)	19 (10)	—	82 (93)	10 (10)	—	23 (100)	45 (66)	—	224 (93)	71 (23)	—
13a	Laboratory/company stated where experimental organism was acquired from?	91 (53)	19 (10)	—	51 (57)	9 (9)	—	13 (54)	44 (65)	—	133 (55)	68 (22)	—
13b	Age of experimental organism given?	75 (42)	11 (6)	—	38 (46)	16 (16)	—	50 (77)	3 (4)	—	145 (50)	23 (7)	—
13c	Weight of experimental organism stated?	108 (60)	11 (6)	—	38 (46)	16 (16)	—	19 (29)	3 (4)	—	149 (52)	23 (7)	—
14	Is a control group reported?	162 (86)	1 (1)	—	80 (82)	1 (1)	—	49 (73)	1 (1)	—	254 (82)	3 (1)	—
15	Conflict of interest statement given?	54 (28)	—	—	23 (23)	—	—	31 (46)	—	—	94 (30)	—	—

Some publications reported more than one type of subject (animal, *in vitro*, and human) and are therefore represented more than once. Values in brackets represent percentages. Only one reviewer assessed questions 8 and 8a, and therefore, values are for half the sample size given in column headings.

that experiments be conducted and analyzed according to good laboratory or clinical practice (GLP, GCP), and that reporting of the results be comprehensive, accurate, and transparent. More than a decade ago, following an analysis of deficiencies in the quality and reporting of randomized clinical trials, journal editors, epidemiologists, and statisticians have published the CONSORT statement and the CONSORT checklist of items to include when reporting a randomized trial (Moher *et al*, 2001*a*). Since then, the quality of reporting and quality in general of randomized clinical trials has greatly improved, which has at least in part been attributed to this process (Plint *et al*, 2006; Hopewell *et al*, 2010). The current study was conducted to further raise the awareness of quality issues in neuroscience research, and to further the implementation of a CONSORT-like statement in experimental medical research.

Specific Checklist Items

Hypothesis/Purpose: A clear statement on the objective of the study or its main hypothesis being tested is critical for the reader to assess the appropriateness of the study design, methods, analysis, and implications. Surprisingly, only 30% of publications explicitly stated a primary research hypothesis; however, 97% indicated the aim or purpose; 44% indicated both, and only 3% gave neither a hypothesis nor an aim (percentages given herein and below represent the mean of both reviewers).

Interestingly, clinical studies involving human participants, wherein the regulatory environment is more strict, were only slightly better than experimental laboratory or *in vitro* studies (see Figure 1, Table 2). Where a study hypothesis is well defined and an analysis protocol agreed in advance of experiments being conducted, then the risk of chance associations being considered of significance are diminished; this is the laboratory equivalent of the *post hoc* subgroup analysis of clinical trial data, which can at best be considered hypothesis generating only. Where a clear hypothesis is stated, data presented that are not relevant to this can be interpreted appropriately (Andersen, 1990).

Randomization/Allocation Concealment/Blinded Assessment of Outcomes: Randomization, allocation concealment, and blinded assessment of outcomes are the key measures to reduce bias and to improve the internal validity of a study. Nevertheless, <20% of the relevant publications

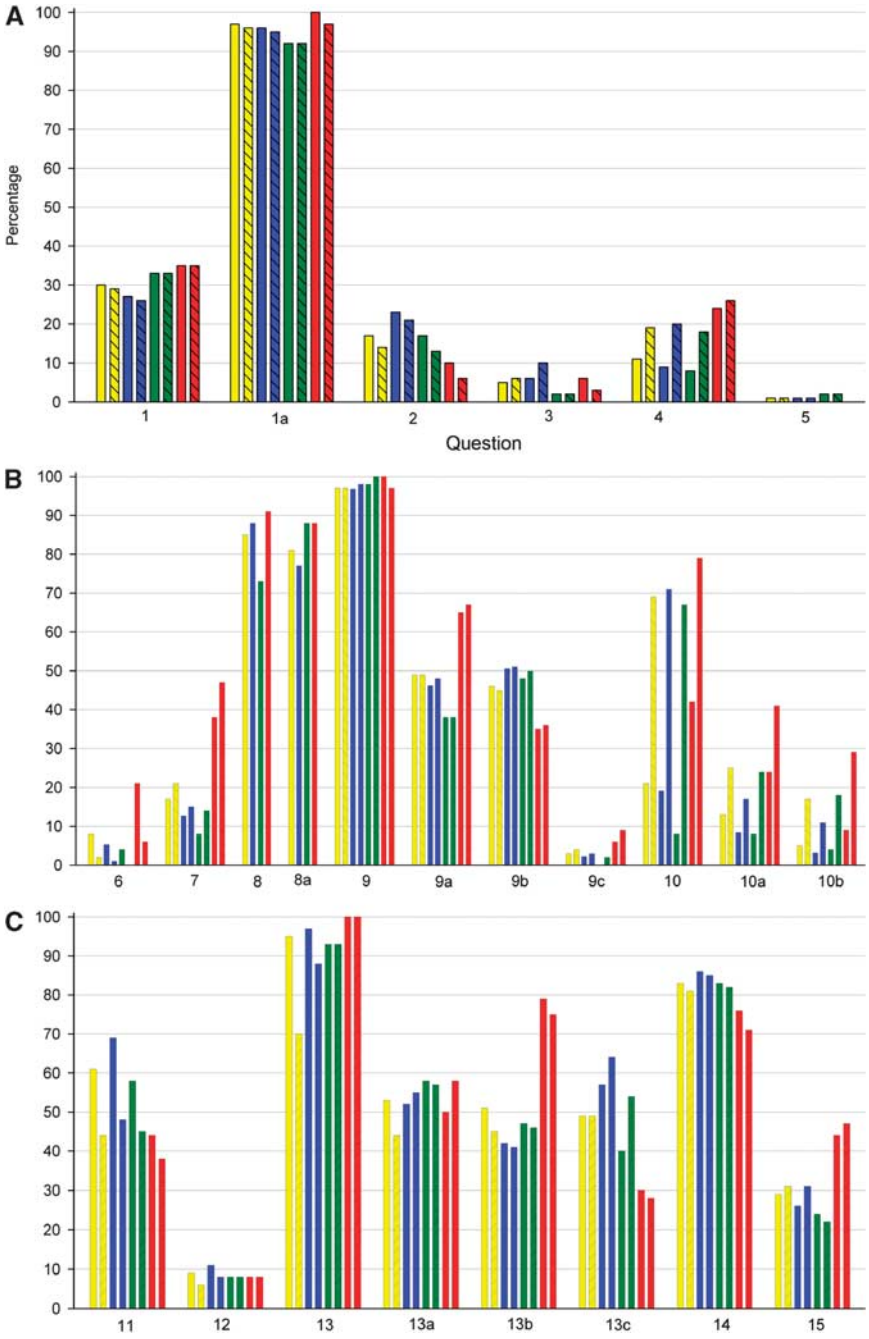

Figure 1 Comparison of the results obtained by two reviewers (HV: solid color; KE: diagonal stripes) assessing the publications presenting all included publications ($N = 156$, yellow), animal studies ($N = 95$, blue), *in vitro* studies ($N = 49$, green), and those involving human participation ($N = 34$, red). *x* axis values indicate the corresponding question number as it appeared on the questionnaire (see Table 1), and *y* axis values represent the number of publications expressed in percentages (0% to 100%). For numerical values, see Table 2.

reported randomization, allocation concealment, or blinded assessment of outcomes. In the few studies in which randomization was mentioned, it was unclear whether proper procedures were followed. We cannot exclude the possibility that some studies may have used such measures but did not report these. Moreover, in some studies, randomization or blinded assessment may not have been feasible, or indicated. However, as reporting of these items is not common, it is often not possible for readers to assess whether such studies are indeed flawed. It should be noted that in two recent studies focusing on experimental stroke studies, roughly 50% of the included studies reported randomization (Minnerup *et al*, 2010; Philip *et al*, 2009).

Sample Size: Only 1% of the studies reported either sample size calculations or power analyses, or the effect size that could have been detected, given the variance of data and the preset levels for α (risk of committing a type I error, or false positive) and β (risk of committing a type II error, or false negative) (Sterne *et al*, 2001; Mulaik *et al*, 1997; Schlattmann and Dirnagl, 2010*a*). The statistical power of a study informs us not only about the risk of false negatives but it is also directly related to the reproducibility and positive-predictive value of the experimental results (Ioannidis, 2005). Again, in some of the studies in this survey, *a priori* sample size calculations would not have made sense. However, in the overwhelming number of studies, they would clearly have been helpful in designing experiments. Most studies included here were in fact grossly underpowered; with sample sizes of 10 per group (which is at the high end of sample sizes found), an α-value of 0.05 and a β-value of 0.8, effect sizes of 1.33 times the s.d. can be detected (two-sided, independent samples *t*-test). As variance in most of the reported experiments is quite high (e.g., s.d. 30% of the mean), this implies that these studies are only powered reliably to detect a 32% mean change in outcomes, such as infarct volume or neurobehavioral score.

Study Design: In this sample of publications, assessing and categorizing study design was very difficult. In our view, this was not so much to do with problems in individual manuscripts but rather to do with the lack of a well-established vocabulary to describe study design in experimental life sciences. This contrasts with the well-established nomenclature for clinical studies, in which categories such as randomized-controlled trial, cohort study, diagnostic trial, screening trial, phase I to IV trial, etc., help to quickly understand the design and assess the implications of a study. We propose to introduce an analogous terminology to the field of experimental medicine, such as phase I experimental trial (safety, tolerability, pharmacokinetics, pharmacodynamics, dose finding; not necessarily in a disease model; small group sizes), phase II experimental trial (dosing and safety in disease model, efficacy, proof of concept), and phase III experimental trial (efficacy, larger group sizes, confounders). If applicable, experimental trial types should be further specified as 'randomized' and 'controlled', if applicable.

Inclusion and Exclusion Criteria/Mortality/Control Group: Reporting inclusion and exclusion criteria in the 'Methods' section, and exclusions, dropouts, or mortality (including the reasons behind it) of the experimental groups in the 'Results' section is another key element to prevent bias and to improve the internal validity of a study. We found that $<20\%$ (29/156) of the publications reported inclusion and exclusion criteria.

Mortality was reported in only 8% (20/129) of studies in which it was relevant; a further 27 studies reported for instance imaging findings, human participants, or purely *in vitro* experiments. We propose that it should be made mandatory to publish mortality rates in animal experiments, as this would prevent the masking of a severe bias which could, e.g., result from excluding severely affected animals in only one of the experimental groups. Control groups are crucial measures to safeguard that an effect is a true effect of the manipulation or condition under study, and to minimize the effect of other, unintended variables on the results. More than 80% of the publications reported the use of control groups. Assuming that in a fraction of articles, control groups would not have made sense, this figure is comforting as it points to a widespread use of this key element of the scientific method.

Experimental Animals and Study Subjects: To understand the implications of a study, and potentially to be able to reproduce experiments, we need to be well informed about the specific characteristics of the experimental animal species, strain, and substrain, supplier, genetic background, age, weight, and sex, among others, which must be considered critical information for reviewers and readers alike.

General Reporting of Statistical Analysis: Specific test statistics were reported in 85% of the assessed publications. Of those, expert statistical assessment revealed that 81% (127/156 relevant publications) of the cases used appropriate approaches and test. The remaining articles either used inappropriate statistical analyses or did not supply the reader with sufficient information to comprehend the reported results. The lack of information made it difficult to evaluate the appropriateness of given statistical tests, and moreover, of the 156 assessed studies, 3 did not provide enough information to judge the methodological quality.

When performing statistical analysis, the method of choice depends on the type of data, e.g., numerical or categorical, as well as on the structure and distribution of data. In all studies assessed, the choice of statistical method was appropriate for the type of data under study.

However, in terms of structure, the results were different. Among others factors, the structure of the data is given by the number of groups considered, e.g., one, two, or more than two groups. If the number of groups is larger than two, applying multiple *t*-tests without correcting for multiple comparisons is not appropriate. This error occurred in 4 of the 156 studies reviewed.

Often, a more complex design is chosen, e.g., when several measurements per animal or subject are taken. In this case, the data points are dependent as we are measuring on the same subject. This introduces a correlation between individual measurements. In such cases, statistical tests that require independent data, such as the unpaired *t*-test or an analysis of variance, are not appropriate. This error occurred in 22 of the 156 assessed studies. As part of a series on statistics in cerebrovascular research, JCBFM presents methods that are appropriate for dependent data (Schlattmann and Dirnagl, 2010*a*, *b*).

Measure of Variance: A high number of publications include a measure of variance. More than 90% (151/155) of the relevant included publications provide the assessor with s.e.m. (46%), s.d. (49%), or confidence interval (4%). A further 9% of relevant publications did not state the type of variance reported. Although popular because it

produces smaller whiskers in graphs for descriptive purposes, s.e.m. is not an acceptable measure. The latter is an estimate for the precision of estimating the mean, not a description of the sample (Altman and Bland, 2005; Schlattmann and Dirnagl, 2010*a*).

Units of Analysis/Data Points: Most of the assessed studies represent data numerically, and more or less directly indicate the unit of analysis (e.g., an Eppendorf tube, right or left hemisphere, a single animal, a group or a cage of animals). Nonetheless, in some cases, the experimental unit/unit of analysis could not be identified.

Data for individual study objects (instead of only presenting group data) were found in 30 publications, and raw data were provided in 17 publications. Reporting raw data and individual data points can help the reader to assess the quality and dispersion of results. As most studies were reporting on rather low numbers of subjects, and because most journals allow almost unlimited publication of additional data as Supplementary material on the website of the journal, we strongly encourage authors to provide as much detailed information as useful and possible. This, in many cases, includes the plotting of individual data points in scatter plots and the listing of numerical values in tables (Schlattmann and Dirnagl, 2010*a*).

Limitations of this Study: Our survey has a number of limitations. None of the dichotomized percentages (yes/no) of the items of our questionnaire (Table 1) can be interpreted as quality statements in themselves. For example, if it is said that 85% of all studies did not use allocation concealment, then it must be kept in mind that not all studies may have allowed designs in which allocation could be concealed, or allocation concealment may have not made sense. We have repeatedly alluded to this constraint in the above discussion. In any case, these numbers present a first overview on the reporting and statistics practices of the surveyed volume.

Our approach was to define criteria and then apply them to all publication categories (such as experimental, animal, and clinical) rather than to define criteria specific to each. Although there might be concerns that not all criteria would be appropriate in each publication category, in fact this was not the case; each criterion was scored as being present in at least one publication from each category and at worse, one-third of publications in a category could be scored (source of experimental organism, human studies). Even with criteria specific to each publication category, we would have some in which criteria were not appropriate, and we believe that our approach has the benefit of being simple and broadly applicable.

It might be argued that some of the questions used in this survey may not be answered in an unambiguous manner. For example, we might have answered question 1: Primary/research hypothesis stated? with 'no', because the hypothesis was not explicitly phrased as such ('In our study we tested the hypothesis that.'), but the author of the study might object because he might insist that the hypothesis reveals itself allusively from the text. However, we argue that a suitably skilled reader should be clearly presented with all the relevant information of a scientific communication, which includes aims, purpose, and hypotheses, to be able to understand, analyze, and potentially replicate the findings.

To assess potential bias of the assessor, we used two assessors, and found a very high degree of interrater agreement. In all categories, questions were more often answered by the scorers with 'no' than with 'yes'. For many questions, a score of 'yes' was low. This in some instances lead to low κ-values despite high agreement between scorers (Supplementary Table 1), as κ is not reliable for rare observations because it is affected by the prevalence of observations.

Restricting our analysis to JCBFM may introduce bias, which precludes generalization of our results to studies published in other journals. This is very unlikely, as JCBFM is one of the top-ranked journals in the field of experimental cerebrovascular and stroke research. Standards for publication, authors, reviewers, etc., are similar to other journals in the field, such as *Stroke*. We have opted to restrict our analysis to this journal, as its scope was ideally suited to survey experimental studies and clinical proof-of-concept studies. We hypothesize that surveying related journals, and even scientific journals in other experimental-translational areas in life sciences would yield very similar results.

Conclusions

In this systematic survey, we found indicators for deficiencies in the design, reporting, and statistical analysis of the original articles in a recent volume of JCBFM, one of the leading journals in the cerebrovascular field. There is ample indirect and direct evidence that this is not a problem unique to this particular journal, or even research field. We, along with others (e.g., Fisher *et al*, 2009; Macleod *et al*, 2009), believe that quality issues in experimental life sciences are an important reason why we are currently facing roadblocks to translation from bench to bedside (and *vice versa*). Only a joint effort of the scientific community (authors, readers, reviewers, editors, professional societies, etc.) can improve the current situation. The results of our study, together with a thorough analysis of the measures that were successfully taken in clinical medicine to improve study quality, indicate some of the action points.

We propose that, in analogy to the CONSORT statement, a set of standards for reporting of experimental studies in biomedicine is used. Like with CONSORT, journal editors need to adopt these standards. Journals need to educate their readers and create awareness in areas, such as proper study design and analysis. Indeed, Kilkenny *et al* (2010) have recently proposed such reporting guidelines (Animal Research Reporting *In Vivo* Experiments, ARRIVE; http://www.nc3rs.org/ARRIVE) using the CONSORT statement as a foundation. Reviewers need to be more critical regarding missing hypotheses, deficiencies of experimental design, insufficient statistical power, or inadequate information in the submitted articles. Most importantly, the upcoming generation of scientists, the students, need to be trained in

GSP, general principles of study design and biostatistics, etc. It has been pointed out that experimental biomedicine should not be constrained by rules and regulations. More specifically, some scientists argue that hypotheses, power calculations, or *a priori* defining the design of a study or its analysis might suffocate the ingenuity of the experimentalist. Quite to the contrary: scientific brilliance or serendipity are traits of the investigator that are unrelated to the standards of scientific publishing. Even the most ingenious finding must be supported by solid and reproducible experiments, which are communicated in a comprehensible manner.

Disclosure/conflict of interest

UD is currently the editor in chief of the *Journal of Cerebral Blood Flow and Metabolism*. MM is currently an Assistant Editor of the journal *Stroke*.

References

Altman DG (1998) Statistical reviewing for medical journals. *Stat Med* 17:2661–74

Altman DG (2002) Poor-quality medical research what can journals do? *Am Med Assoc* 287:2765–7

Altman DG, Bland JM (2005) Standard deviations and standard errors. *Br Med J* 331:903

Altman DG, Gore SM, Gardner MJ, Pocock SJ (1983) Statistical guidelines for contributors to medical journals. *Br Med J (Clin Res Ed)* 286:1489–93

Andersen B (1990) *Methodological Errors in Medical Research*. Oxford, UK: Blackwell Science Ltd

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF (1996) Improving the quality of reporting of randomized clinical trials. The CONSORT statement. *JAMA* 276:637–9

Bellolio MF, Serrano LA, Stead LG (2008) Understanding statistical tests in the medical literature: which test should I use? *Int J Emerg Med* 1:197–9

Cicchetti DV, Feinstein AR (1990) High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 43:551–8

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U (2008) Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke* 39: 929–34

Dirnagl U (2006) Bench to bedside: the quest for quality in experimental stroke research. *J Cereb Blood Flow Metab* 26:1465–78

Dirnagl U, Macleod MR (2009) Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. *Br J Pharmacol* 157:1154–6

Feinstein AR, Cicchetti DV (1990) High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 43:543–9

Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, Lo EH, STAIR Group (2009) Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* 40:2244–50

García-Berthou E, Alcaraz C (2004) Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol* 4:13

Glantz SA (1980) Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 61:1–7

Hoffmann O (1984) Application of statistics and frequency of statistical errors in articles in Acta Neurochirurgica. *Acta Neurochirurgica* 71:307–15

Holmes TH (2004) Ten categories of statistical errors: a guide for research in endocrinology and metabolism. *Am J Physiol Endocrinol Metab* 286:E495–501

Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG (2010) The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ* 340:c723

Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:e124

Jerndal M, Forsberg K, Sena ES, Macleod MR, O'Collins VE, Linden T, Nilsson M, Howells DW (2010) A systematic review and meta-analysis of erythropoietin in experimental stroke. *J Cereb Blood Flow Metab* 30:961–8

Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG (2010) Improving bioscience research reporting—ARRIVE-ing at a solution. *PLoS Biol* 8:e1000412

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE* 4:e7824

Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, Buchan A, van der Worp HB, Traystman RJ, Minematsu K, Donnan GA, Howells DW (2009) Good laboratory practice: preventing introduction of bias at the bench. *J Cereb Blood Flow Metab* 29:221–3

Minnerup J, Wersching H, Diederich K, Schilling M, Ringelstein EB, Wellmann J, Schäbitz WR (2010) Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. *J Cereb Blood Flow Metab* 30:1619–24

Moher D, Jones A, Lepage L (2001*b*) Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 285:1992–5

Moher D, Schulz KF, Altman DG (2001*a*) The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Med Res Methodol* 1:1471–2288

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: Lawrence Erlbaum Associates, 66–115

Nieminen P, Carpenter J, Rucker G, Schumacher M (2006) The relationship between quality of research and citation frequency. *BMC Med Res Methodol* 6:42

O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW (2006) 1,026 experimental treatments in acute stroke. *Ann Neurol* 59:467–77

O'Collins VE, Donnan GA, Macleod MR, Howells DW (2009) Scope of preclinical testing versus quality control within experiments. *Stroke* 40:e497

Philip M, Benatar M, Fisher M, Savitz SI (2009) Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. *Stroke* 40:577–81

Phillips C, MacLehose R, Kaufman J (2008) Errors in statistical tests 3. *Emerg Themes Epidemiol* 5:9

Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, Gaboury I (2006) Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* 185:263–7

Sarter M, Fritschy JM (2008) Reporting statistical methods and statistical results in EJN. *Eur J Neurosci* 28:2363

Schlattmann P, Dirnagl U (2010*a*) Statistics in experimental cerebrovascular research—comparison of two groups with a continuous outcome variable. *J Cereb Blood Flow Metab* 30:474–9

Schlattmann P, Dirnagl U (2010*b*) Statistics in experimental cerebrovascular research—comparison of more than two groups with a continuous outcome variable. *J Cereb Blood Flow Metab* 30:1558–63

Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R (2008) What errors do peer reviewers detect, and does training improve their ability to detect them? *JRSM* 101:507

Sena E, van der Worp HB, Howells D, Macleod M (2007) How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 30:433–9

Sterne JAC, Egger M, Davey Smith G (2001) Investigating and Dealing with Publication and Other Biases. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Publishing Group, 189–208

Thomson Reuters (2010) 2009 Journal Citation Reports® Science Edition

Williams JL, Hathaway CA, Kloster KL, Layne BH (1997) Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am J Physiol Heart Circ Physiol* 273:H487–93

Zinsmeister AR, Connor JT (2008) Ten common statistical errors and how to avoid them. *Am J Gastroenterol* 103:262–6

Supplementary Information accompanies the paper on the Journal of Cerebral Blood Flow & Metabolism website (http://www.nature.com/jcbfm)