PLOS ONE

RESEARCH ARTICLE

The Researchers' View of Scientific Rigor—Survey on the Conduct and Reporting of *In Vivo* Research

Thomas S. Reichlin, Lucile Vogt, Hanno Würbel*

Division of Animal Welfare, Veterinary Public Health Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland

* hanno.wuerbel@vetsuisse.unibe.ch

Citation: Reichlin TS, Vogt L, Würbel H (2016) The Researchers' View of Scientific Rigor—Survey on the Conduct and Reporting of *In Vivo* Research. PLoS ONE 11(12): e0165999. doi:10.1371/journal.pone.0165999

Editor: Bart O. Williams, Van Andel Institute, UNITED STATES

Received: July 21, 2016

Accepted: October 23, 2016

Published: December 2, 2016

Data Availability Statement: The original data files are available from the figshare database (accession number 10.6084/m9.figshare.3796155).

Funding: This study was funded by the Swiss Federal Food Safety and Veterinary Office (FSVO, www.blv.admin.ch/, Grant No. 2.13.01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Reproducibility in animal research is alarmingly low, and a lack of scientific rigor has been proposed as a major cause. Systematic reviews found low reporting rates of measures against risks of bias (e.g., randomization, blinding), and a correlation between low reporting rates and overstated treatment effects. Reporting rates of measures against bias are thus used as a proxy measure for scientific rigor, and reporting guidelines (e.g., ARRIVE) have become a major weapon in the fight against risks of bias in animal research. Surprisingly, animal scientists have never been asked about their use of measures against risks of bias and how they report these in publications. Whether poor reporting reflects poor use of such measures, and whether reporting guidelines may effectively reduce risks of bias has therefore remained elusive. To address these questions, we asked in vivo researchers about their use and reporting of measures against risks of bias and examined how self-reports relate to reporting rates obtained through systematic reviews. An online survey was sent out to all registered in vivo researchers in Switzerland (N = 1891) and was complemented by personal interviews with five representative in vivo researchers to facilitate interpretation of the survey results. Return rate was 28% (N = 530), of which 302 participants (16%) returned fully completed questionnaires that were used for further analysis. According to the researchers' self-report, they use measures against risks of bias to a much greater extent than suggested by reporting rates obtained through systematic reviews. However, the researchers' self-reports are likely biased to some extent. Thus, although they claimed to be reporting measures against risks of bias at much lower rates than they claimed to be using these measures, the self-reported reporting rates were considerably higher than reporting rates found by systematic reviews. Furthermore, participants performed rather poorly when asked to choose effective over ineffective measures against six different biases. Our results further indicate that knowledge of the ARRIVE guidelines had a positive effect on scientific rigor. However, the ARRIVE guidelines were known by less than half of the participants (43.7%); and among those whose latest paper was published in a journal that had endorsed the ARRIVE guidelines, more than half (51%) had never heard of these guidelines. Our results suggest that whereas reporting rates may underestimate the true use of measures against risks of bias, self-reports may overestimate it. To a large extent, this discrepancy

can be explained by the researchers' ignorance and lack of knowledge of risks of bias and measures to prevent them. Our analysis thus adds significant new evidence to the assessment of research integrity in animal research. Our findings further question the confidence that the authorities have in scientific rigor, which is taken for granted in the harm-benefit analyses on which approval of animal experiments is based. Furthermore, they suggest that better education on scientific integrity and good research practice is needed. However, they also question reliance on reporting rates as indicators of scientific rigor and highlight a need for more reliable predictors.

Introduction

Reproducibility is the cornerstone of the scientific method and fundamental for the ethical justification of in vivo research. Mounting evidence of poor reproducibility (e.g. [1,2]) and translational failure of preclinical animal research [3–5] has therefore raised serious concerns about the scientific validity [6,7] and ethical justification [8,9] of in vivo research. Possible reasons for poor reproducibility include a lack of education [10,11], perverse incentives [12], ignorance of standards of good research practice [2], as well as scientific misconduct and fraud [13]. All of these may result in poor experimental design and conduct, thereby compromising scientific validity [5,14–17].

Poor scientific validity has important scientific, economic, and ethical implications. It hampers scientific and medical progress and leads to translational failure through misguided research efforts (e.g. [3,18–21]). It also increases R&D costs in drug development [22], resulting in higher health care costs (e.g. [17]). Based on estimates of irreproducibility in preclinical research, up to USD 28B/year may be spent in the US alone on irreproducible preclinical research [19]. Furthermore, poor scientific validity imposes unnecessary harm and distress upon research animals (e.g. [8,9]), raises false hopes in patients awaiting cures for their diseases, and puts patients in clinical trials at risk [23].

Much of the evidence of poor experimental design and conduct in animal research rests on systematic reviews and meta-analyses revealing low rates of reporting of measures against risks of bias (e.g., randomization: mean = 27% [range = 9–55%], blinding: 28.7% [0–61%], sample size calculation: 0.5% [0–3%]) in the primary literature (e.g. [23,24–33]). Consequently, reporting guidelines such as the 'Animal Research: Reporting of In Vivo Experiments' (ARRIVE) guidelines [34] (https://www.nc3rs.org.uk/arrive-guidelines) or the revised 'Reporting Checklist for Life Science Articles' by the Nature publishing group (http://www.nature.com/authors/policies/reporting.pdf) were promoted in view of improving the situation. For example, the ARRIVE guidelines consist of a checklist of 20 items of information that all publications reporting animal research should include, including details of methods used to reduce bias such as randomisation and blinding. Despite general consensus about the benefits of such guidelines (> 1000 journals have endorsed the ARRIVE guidelines by September 2016), Baker et al. [35] found that reporting rates of measures against bias remained low in PLoS and Nature journals even after they had endorsed the ARRIVE guidelines. Although reporting rates are generally increasing, they are still rather low [26].

In the past, the reporting of measures against bias such as randomization, blinding, sample size calculation and others was largely optional and–to some extent–this is still the case today. Therefore, reporting rates of these measures may not be reliable indicators of scientific rigor. Whether poor reporting reflects poor scientific validity, however, has never been systematically

studied. Nevertheless, some indications exist that scientific rigor is often lacking, and that risks of bias are associated with poor reporting. For example, in neuroscience research most experimental studies are underpowered, and low statistical power in combination with null hypothesis significance testing and publication bias may lead to inflated effect size estimates from the published literature (e.g. [36]); inappropriate statistical methods often lead to spurious conclusions (e.g. [37,38,39]); and several systematic reviews indicate that low reporting rates of measures against bias are associated with larger effect sizes (e.g. [28,29,30,40]). This has raised concerns that there may be systemic flaws in the way we conduct and report research [2]. Several authors warned that the quality of animal research is (unacceptably) poor (e.g. [41]) and stricter adherence to standards of best research practice is necessary if the scientific validity of animal research is to be improved [15].

In light of the many studies published on poor reporting of measures against bias and the level of attention they received [5,7,18,42], it is surprising that so far no study has investigated the relationship between what researchers do in the laboratory and what they report in their publications. The primary aim of the present study, therefore, was to assess the researchers' view of the quality of experimental conduct and how this relates to what they report in the primary literature. Using a questionnaire sent out to all registered animal scientists actively involved with ongoing animal experiments in Switzerland, we assessed (i) the researchers' awareness and knowledge of risks of bias in animal research, (ii) the measures they take to avoid bias in their own research, and (iii) how they report these measures in their publications. To aid interpretation of the results, we also conducted qualitative interviews with a small subset of these researchers to get insight into personal viewpoints, underlying motivations, and compliance with quality standards.

Methods

Online survey

An anonymous online survey was developed using the free software Limesurvey [43]. The survey contained a total of 21 questions divided into seven sections. Thus, participants were asked about (i) their area of research, and the species they were mainly working with; (ii) their work institution, including certification; (iii) experimental design and conduct, including which of seven primary measures against risks of bias (Table 1) participants generally apply to their own research; (iv) the journal of their latest scientific publication and which of the seven measures against bias listed in Table 1 they had reported in that publication (or the reasons for not reporting them); (v) awareness of risks of bias, and knowledge about measures to prevent them; (vi) familiarity with the ARRIVE (or similar) guidelines, and whether they adhered to them; and (vii) the participants' personal research experience. The questionnaire was piloted among five animal researchers to ensure clarity. Participants had to answer all questions of a section before being able to move on to the next section, however, for most questions they had the option of not answering questions by ticking 'no answer', 'do not know', or 'not relevant'. The full questionnaire is available in S1 Text.

Study population and data collection

The online survey was set up as a partially closed survey, for which potential participants (N = 1891) were invited via email. Email addresses were provided by the Swiss Federal Food Safety and Veterinary Office (FSVO) and included all researchers involved in ongoing animal experiments in Switzerland, which were registered by the FSVO as experimenters, study directors, or resource managers of animal facilities. The questionnaire was online for seven weeks;

Table 1. List of measures against risks of bias included in this study.

Measure	Definition	Bias
Allocation Concealment	Concealment of allocation sequence from those assigning subjects to treatment groups, until the moment of assignment.	Selection Bias
Randomization	Allocation of study subjects randomly to treatment groups across the comparison, to ensure that group assignment cannot be predicted.	Selection Bias
Blinding	Keeping the persons involved in an experiment (i.e. experimenter, data collector, outcome assessors) unaware of the treatment allocation.	Attrition Bias, Detection Bias, Performance Bias
Sample Size Calculation	Appropriate *a priori* determination of number of study subjects for a given test setup that allows for a detection of a treatment effect given the power to find an effect of a defined size.	"Avoiding wastage of animals"
Inclusion and Exclusion Criteria	*A priori* defined characteristics which describe on which basis subjects will be included in the study or how they need to be treated in case of attrition.	Attrition Bias, Selective Reporting
Primary Outcome	*A priori* defined main variable of interest, on which the treatment effect is measured; with sample size calculation being based on it.	Selective Reporting
Statistical Analysis Plan	*A priori* definition of statistical methods by which the primary outcome variable is analyzed at the end of the study.	Attrition Bias, Selective Reporting

Definitions adapted from van der Worp et al. [5], CONSORT (www.consort-statement.org), the Cochrane Collaboration (methods.cochrane.org).

doi:10.1371/journal.pone.0165999.t001

after five weeks a reminder for participation was sent to all addressees to increase response rate.

Ethics statement

Given that there were no known risks associated with this research study, participants of the survey and the interviews were not a vulnerable group of people, and complete confidentiality was guaranteed, we saw no need for formal ethical review before the study began.

Data analysis

The online survey generated 530 questionnaires (return rate: 28%), of which 302 (57%) were fully completed while 228 were only partially completed. Partially completed questionnaires were only used for assessing a potential bias in the sample of fully completed questionnaires, while only the latter (16% of the total sample) were used for further analysis. Survey data were exported to MS Excel, checked for inconsistencies, and revised if necessary with suitable correction rules. Each question of the survey was analyzed quantitatively in terms of proportions of the answers given by the participants.

Besides analyzing each question separately, internal validity scores (IVS) were calculated for each participant for a) experimental conduct (IVS_{Exp}), and b) reporting in the latest publication (IVS_{Pub}). The scores were based on the measures against risks of bias (Table 1) and were equal to the number of these measures claimed to be applied (under a, Eq 1) or reported (under b, Eq 2) by the participant, divided by the number of measures that were applicable to a) or b), respectively.

$$\mathrm{IVS}_{\mathrm{Exp}} = \text{number of } yes \text{ and } depends \,/\, (7 - \text{number of } no\ answer) \tag{1}$$

$$\mathrm{IVS}_{\mathrm{Pub}} = \text{number of } yes\,(\text{full details}) - (6 - \Sigma(\text{numbers of } no\ answer + does\ not\ apply\ to\ last\ manuscript + have\ not\ published\ yet)). \tag{2}$$

For an overview of all possible answer options, please refer to the copy of the online survey in S1 Text. Due to a mistake in the way the questions regarding allocation concealment were formulated, data for this measure were excluded from both IVS in the case of a direct

comparison between the scores (change of denominator in Eq 1 to "6—number of *no answer*"). In addition, the inuence of several independent variables (descriptors of the participants derived from the online survey) on these scores was investigated through an information theoretic modelling approach using generalized linear models (*glm*). The Bayesian Information Criterion (BIC) was used to compare candidate models [44] and to retrieve the model which best described the data [45]. The two scores were modelled with the following main effects (descriptors): Knowledge of the ARRIVE guidelines (binary; yes, no), host institution (categorical; academia, industry, governmental, private), animal research experience (continuous; no. of years), the authority (cantonal veterinary ofce) responsible for approving the participants' applications for animal experiments (categorical; 13 cantons), eld of research (categorical: basic, applied, other), and the research discipline (categorical: Animal Welfare, Cell Biology/Biochemistry/Molecular Biology, Ethology, Human Medicine, Para-clinics, Veterinary Medicine, Zoology, other discipline).

Starting with the full model (all descriptors including the interaction term knowledge of ARRIVE x institution), single term deletion was performed by a stepwise backwards procedure (*drop1* function), eliminating the descriptor with the largest p-value to produce a set of candidate models for the model selection process. Besides this set of candidate models, we also included all univariate models (single descriptors) as well as the null model (only intercept; total of 12 models). The model comparison was conducted using the function *model.sel* from the R-package MuMIn [46]. The model with the lowest BIC was chosen as the one fitting the data best. Model estimates and 95% confidence intervals were corrected for overdispersion of the data (*glm* link function = quasibinomial).

In order to investigate whether the IVS_{Exp} and IVS_{Pub} were correlated, a Spearman's Rank Correlation was performed with the reduced IVS scores (only considering six validity criteria, i.e., without allocation concealment). Mean differences in IVS between participants of certified institutions vs. non-certified institutions were investigated with a Wilcoxon Rank Sum Test for both scores. Values for IVS are presented as means ± SD. All statistical analysis were performed using the statistical software R, Version 3.0.3 [47].

Personal Interviews

The online survey was complemented by interviews with five selected researchers representing the diversity of institutions and areas of research among the participants of the survey. The interviews are not described in the main text of this article, however, complete information about the methods, study population, analysis and results of the interviews is provided in S2 Text.

Results

Study population

The 302 participants returning fully completed questionnaires had an average of 15.5 (SD ± 8.6) years of experience in animal research. Most of them were affiliated with academic institutions (74.5%, N = 225), 14.9% (N = 45) with industry, 4.6% (N = 14) with governmental research institutions, and 6% (N = 18) with private research institutions. Only 23.8% of the participants (N = 72) indicated that their institution was formally certified, and 27.2% (N = 82) that it was not, while almost half of the participants (44.7%) did not know this or did not answer this question (4.3%) (for a complete table, see S1 Table). Among the 72 participants indicating that their institution was certified, academics were relatively underrepresented with only 45.8% compared to 74.5% among the total sample, whereas researchers from

Table 2. Research Disciplines of Survey Participants.

Discipline	Number	Proportion [%]
Human Medicine	108	36
Biochemistry, Cell or Molecular Biology	105	35
Veterinary Medicine	27	9
Para-clinics	22	7
Other	15	5
Animal Welfare	12	4
Zoology	11	4
Ethology	2	1

doi:10.1371/journal.pone.0165999.t002

pharmaceutical industry (29.2% vs. 14.9%), governmental institutions (8.3% vs. 4.6%), and private institutions (16.7% vs. 6%) were relatively overrepresented.

Most participants (58.9%, N = 178) attributed their work to basic research and 40.4% (N = 122) to applied research, while two participants (0.7%) were undecided. The large majority of participants (86.8%) were engaged in biomedical or medical research (for details see Table 2), and the animals used as experimental subjects were mainly mice (60.6%) and rats (15.6%) (for details see Table 3). While 14 participants (4.6%) had not yet published their first paper, most participants (57%, N = 172) had published between 1 and 20 papers, and 116 participants (38.4%) had published more than 20 papers.

Measures to avoid bias

When asked which of the seven measures against risks of bias the participants normally used in the conduct of their experiments (including the answers 'yes' and 'depends'), a large majority ticked primary outcome variable (90%, N = 264), inclusion and exclusion criteria (84%, N = 245), randomization (86%, N = 248), and statistical analysis plan (82%, N = 240). More than half also ticked sample size calculation (69%, N = 203) and allocation concealment (52%,

Table 3. Primary Research Species of Survey Participants.

Primary Species	Number	Proportion [%]
Mice	183	60.60
Rats	47	15.56
Fish	15	4.97
Cattle	10	3.31
Dogs	9	2.98
Birds (incl. Poultry)	7	2.32
Amphibians, Reptiles	6	1.99
Donkeys, Horses	5	1.66
Pigs	5	1.66
Sheep, Goats	5	1.66
Primates	3	0.99
Rabbits	2	0.66
Cats	1	0.33
Guinea Pigs	1	0.33
Invertebrates	1	0.33
Other Mammals	1	0.33
Other Rodents	1	0.33

doi:10.1371/journal.pone.0165999.t003

N = 143), whereas less than half (47%, N = 135) ticked blinded outcome assessment (see Fig 1A white bars). These proportions were corrected for the number of participants ticking 'no answer' (for full results including absolute numbers see S2 Table).

To put these numbers in relation to reporting rates derived from published papers, we asked participants to state explicitly which of these seven measures against risks of bias they had reported in their latest published research article. Most of the participants indicated that they had reported in full details a statistical analysis plan (71%, N = 180) and the primary outcome variable (78%, N = 177), whereas reporting rates for inclusion and exclusion criteria (45%, N = 97), randomization (44%, N = 87), sample size calculation (18%, N = 40) and blinding (27%, N = 49) were considerably lower (Fig 1A grey bars). Again, reporting rates were corrected for the number of participants having ticked 'does not apply to last manuscript', 'have not published so far', and 'no answer'.

For each of these bias avoidance measures, between 5.3% (statistical analysis) and 27.5% (blinding) of the 302 participants considered these measures to be irrelevant with respect to their latest publication (see "NA" in S2B Table). The most common reason (chosen from a drop-down list) for not reporting measures against risks of bias in their latest publications was that it was 'not necessary' (from 30% for sample size calculation up to 80% for statistical analysis). Additional reasons were that it was 'not common' (up to 39% for sample size calculation), that they 'did not think of it' (up to 19% for primary outcome variable) or space limitations by the journals (up to 8% for sample size calculation).

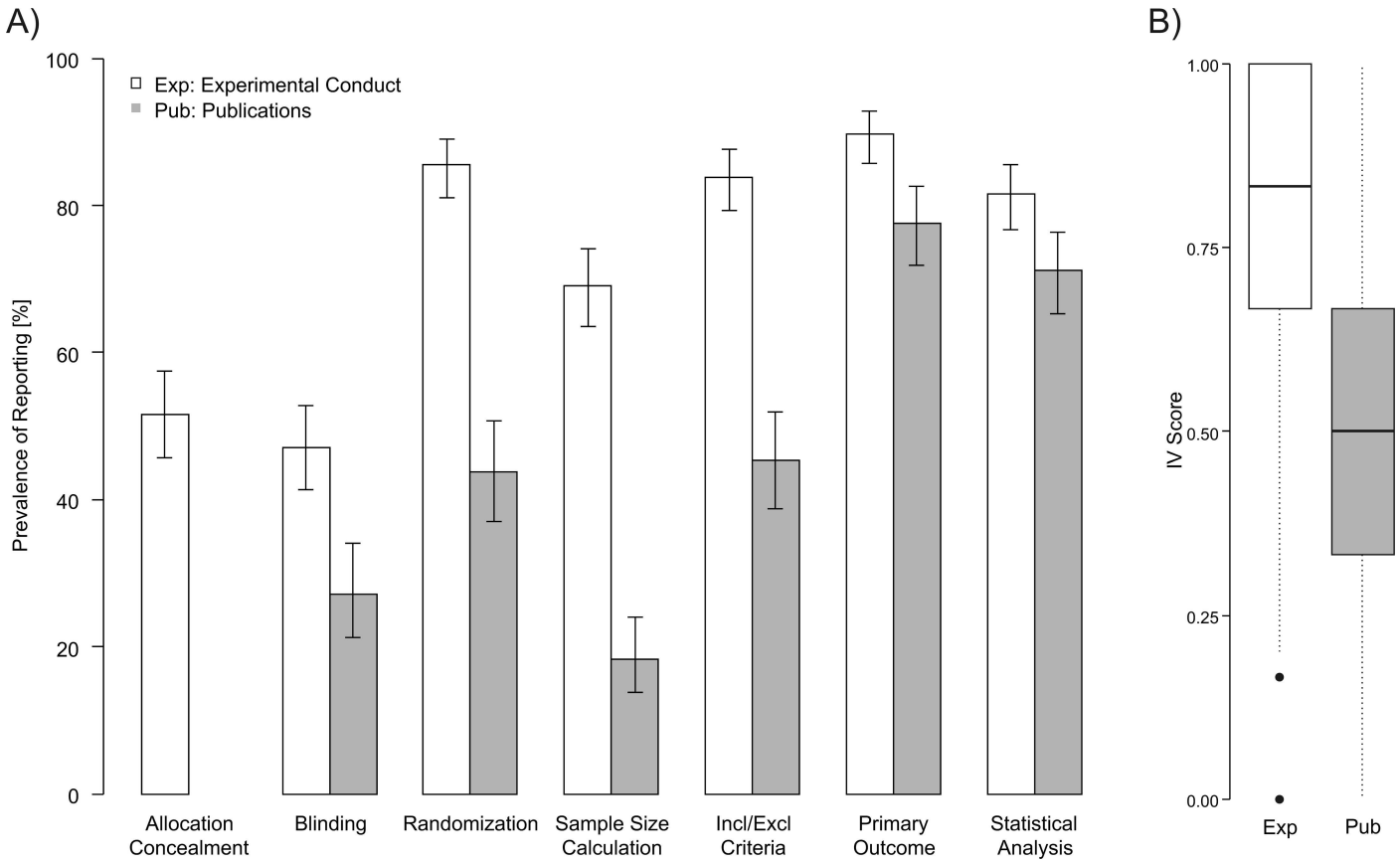

Fig 1. Prevalence of the measures used and reported to avoid risks of bias by the participants to online survey. (A) Prevalence of use of bias avoidance measures during experimental conduct and reporting in the participants' latest publication (percentages are corrected for 'no answer', 'does not apply to last manuscript' and 'have not published so far'. (B) Internal validity scores (IVS) for experimental conduct and reporting in publications.

doi:10.1371/journal.pone.0165999.g001

IV scores

The mean IVS_{Exp} based on all seven measures against risks of bias was 0.73 (SD ± 0.24, N = 301). However, to facilitate comparison with the IVS_{Pub}, we also calculated an IVS_{Exp} based on six measures only (excluding allocation concealment), resulting in a mean IVS_{Exp} of 0.76 (SD± 0.23; N = 301) compared to a mean IVS_{Pub} of 0.49 (SD ± 0.29, N = 261). There was a weak but significant positive correlation between IVS_{Exo} and IVS_{Pub} (Spearman's Rank Correlation: S = 2322088, Rho = 0.22, p- value = 0.0004).

Variation in IVS_{Exp} was best explained by knowledge of the ARRIVE guidelines and by the participants' field of research (BIC = 1132, $BIC_{weights}$ = 0.967; ΔBIC to second best model [only including ARRIVE knowledge] = 7.38). There was a positive effect of 'ARRIVE knowledge' on the IVS_{Exp} compared to 'no knowledge' (model estimate = 0.406, 95% CI = 0.133–0.683), and negative effects of 'basic research' (model estimate = -0.299, 95% CI = -0.579–-0.0223) and 'other research' (model estimate = -2.052, 95% CI = -3.912–-0.526) compared to 'applied research' (Fig 2A).

The model including knowledge of the ARRIVE guidelines performed best in explaining variation in the IVS_{Pub} (BIC = 874.24, $BIC_{weights}$ = 0.995, ΔBIC to second best model [null model with intercept only] = 11.4). Again, knowledge of the ARRIVE guidelines had a positive effect on the IVS_{Pub} compared to 'no knowledge' (model estimate = 0.461, 95% CI = 0.201–0.723) (Fig 2B). An overview of the models and selection procedure can be found in S3 Table.

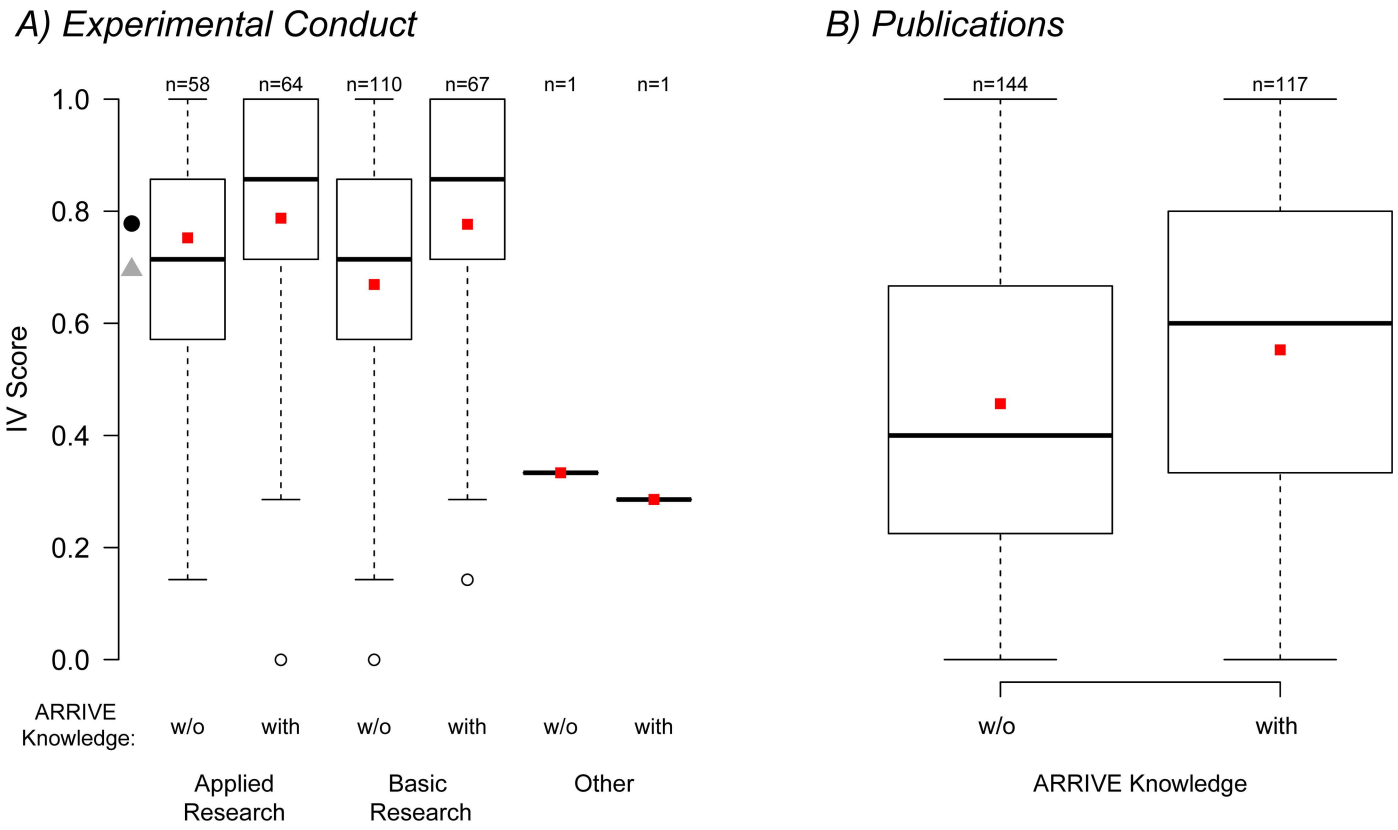

Fig 2. Boxplot of IVS versus descriptors of model selection process. Descriptors are selected for the models with lowest BIC, thus best explaining the variation in IVS for (A) experimental conduct and (B) for publications. For (A) one value is missing as no IVS could be calculated, and for (B) 41 values are missing, because participants ticked 'have not yet published', gave 'no answer' or declared that these questions 'do not apply to last manuscript'. For experimental conduct (A), the model including ARRIVE knowledge and research field best explained the IVS_{Exp}, whereas for publications (B), the model including only ARRIVE knowledge best explained the IVS_{Pub}. Red squares indicate the mean IVS; black circle the mean of IVS_{Exp} of participants with ARRIVE knowledge; grey triangle the mean IVS_{Exp} of participants without ARRIVE knowledge. Whiskers are 1.5*interquartile range.

doi:10.1371/journal.pone.0165999.g002

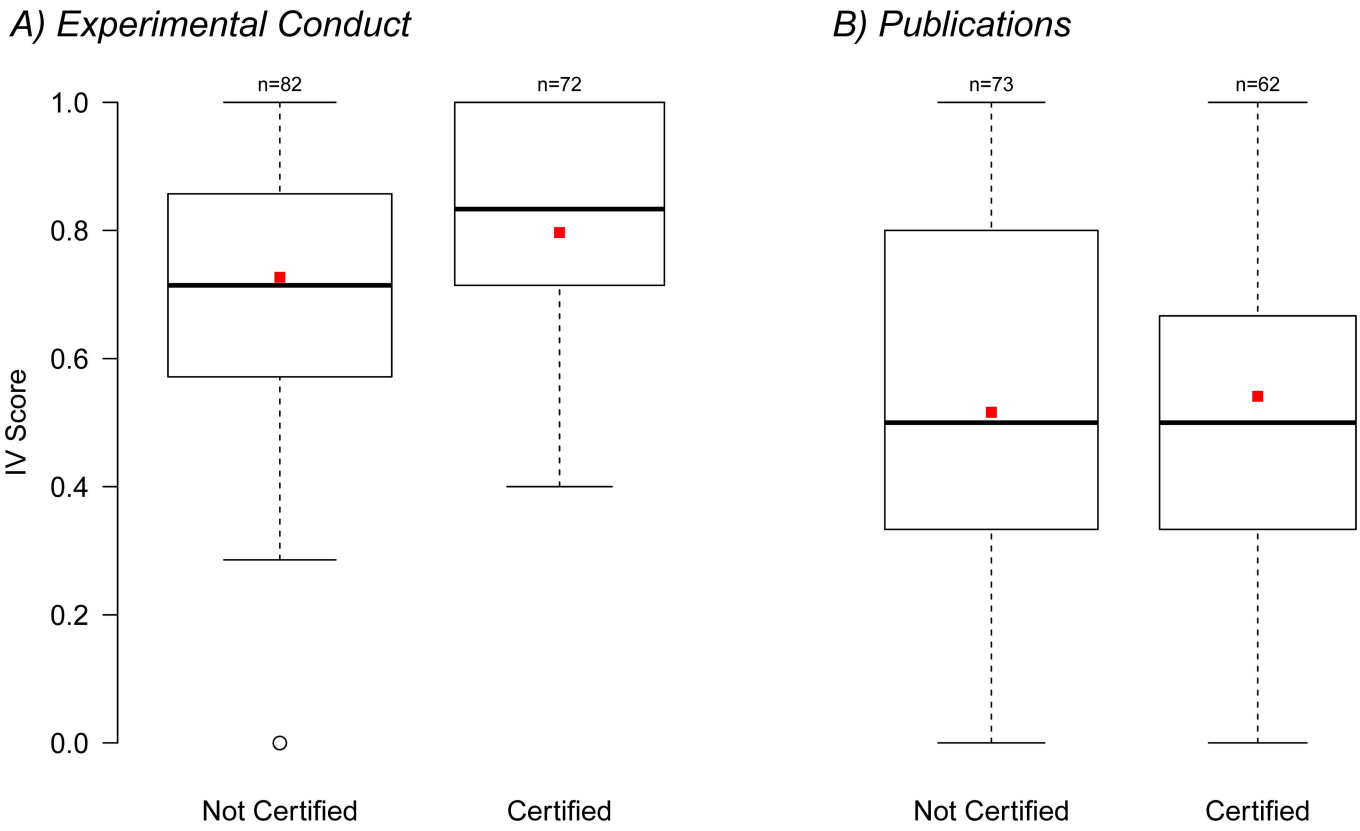

Fig 3. Boxplot of IVS versus certification of institution. Comparison of IVS for (A) experimental conduct, and (B) for the reporting in publications with respect to working institutions being certified. Mean IVS are slightly but non-significantly higher for participants working for certified institutions.

doi:10.1371/journal.pone.0165999.g003

The IVS_{Exp} was slightly higher in participants from certified institutions (mean IVS_{Exp} certified = 0.81, SD ± 0.30, N = 72) compared to non-certified institutions (mean IVS_{Exo} non-certified = 0.73 ± 0.23, N = 82; Fig 3A), however, this difference was not significant (Wilcox Rank Sum Test, W = 2490, p = 0.087). Similarly, IVS_{Pub} was slightly but not significantly higher (W = 2144, p = 0.60) in participants working at certified institutions (mean IVS_{Pub} = 0.54 ± 0.27, N = 62; mean IVS_{Pub} = 0.51 ± 0.30, N = 73; Fig 3B).

Awareness of risk of bias and measures aimed to avoid them

As summarized in Table 4, most participants indicated that they were aware of risks of bias caused by selective reporting (67.5%, N = 204), selection bias (65.2%, N = 197), and detection bias (61.9%, N = 187), and that they avoid these risks routinely in their research. Furthermore, about half of the participants indicated being aware of publication bias (57.6%, N = 174) and performance bias (48.7%, N = 147), whereas less than one third (29.8%, N = 90) indicated being aware of attrition bias. However, depending on the type of bias only between 15.6% and 41.7% of the participants indicated being concerned about these biases with respect to their own research, and between 15.2% and 35.8% of the participants indicated that these biases did not apply to their own research. Moreover, 10.9% (N = 33) of the participants indicated not being aware of any of these biases, and 24.2% (N = 73) indicated that they were not concerned about any of these biases with respect to their own research (Table 4).

Next, we assessed the participants' knowledge of specific measures against risks of bias. Fig 4 presents their responses when asked what measures they would take to avoid the different types of

Table 4. Participants' Assessment of Different Types of Experimental Biases.

	A) Awareness	B) Concerned	C) Avoidance	D) Not Relevant
Selective Reporting	67.5 (204)	37.4 (113)	58.9 (178)	17.5 (53)
Selection Bias	65.2 (197)	39.7 (120)	58.3 (176)	17.9 (54)
Detection Bias	61.9 (187)	41.7 (126)	52.3 (158)	15.2 (46)
Publication Bias	57.6 (174)	35.4 (107)	41.7 (126)	24.5 (74)
Performance Bias	48.7 (147)	29.1 (88)	45.7 (138)	22.8 (69)
Attrition Bias	29.8 (90)	15.6 (47)	22.8 (69)	35.8 (108)
None of Above	10.9 (33)	24.2 (73)	13.9 (42)	42.7 (129)
Other Types of Bias	3.0 (9)	1.7 (5)	2.0 (6)	1.7 (5)

Questions included (A) what biases participants were generally aware of, (B) which biases they were concerned with in their own research, (C) which types of bias they were trying to avoid routinely, and (D) which of these biases did not apply to their own research. Shown are percentages with the absolute numbers in brackets.

doi:10.1371/journal.pone.0165999.t004

bias. As indicated by the distribution of responses across the different panels, apart from publication bias (panel D), there was no clear pattern of preference for effective over ineffective measures.

The ARRIVE guidelines were known by 43.7% of the participants (N = 132), of which 24 indicated that they were familiar with these guidelines, 35 that they had read them, and 73 that they had heard of them. However, the majority of participants (56.3%, N = 170) indicated that they had never heard of the ARRIVE guidelines before (Fig 5). Among the 132 participants being aware of the ARRIVE guidelines, most indicated that they adhere to them either generally (30.3%, N = 40) or occasionally 34.8%, N = 46), while 15.2% (N = 20) answered that they did not adhere to them and 19.7% (N = 26) did not answer this question.

Consulting the NC3rs Website (https://www.nc3rs.org.uk/arrive-animal-research-reporting-vivo-experiments#journals, accessed July 6th 2015), the journals in which participants had published their latest paper was checked for endorsement of the ARRIVE guidelines. Of all participants having published at least one research paper (N = 288), 79 (27.4%) had published their latest paper in a journal that had endorsed the ARRIVE guidelines (86.1% [N = 68] from academia, 6.3% [N = 5] from governmental institutions, 5.1% [N = 4] from industry, 2.5% [N = 2] from private research institutions). Among these participants, 16.5% (N = 13) indicated that they were familiar with the ARRIVE guidelines, 11.4% (N = 9) that they had read them, and 21.5% (N = 17) that they had heard of them. However, more than half of the participants who had last published in a journal endorsing the ARRIVE guidelines (51%, N = 40) indicated that they had never heard of these guidelines.

Apart from the ARRIVE guidelines, 235 participants (77.8%) indicated that they adhered to other guidelines either regularly (N = 217, 71.8%) or occasionally (N = 18, 6%), while 26 participants (8.6%) never followed any guidelines (41 participants [13.6%] did not answer this question). Among the 235 participants who did adhere to other guidelines either regularly or occasionally, 72.3% (N = 170) referred to internal SOPs (Standard Operating Procedure), 59.6% (N = 140) to specific journal guidelines, and 24.2% (N = 57) to various other guidelines (multiple answers were possible).

Assessment of possible bias in study sample

To assess whether our study sample of fully completed questionnaires (N = 302; 16% of total survey population) might be biased, we exploited our sample of partially completed questionnaires (N = 228) and compared participant characteristics between these two samples as well as the primary outcome variable of this study, IVS_{Exp}.

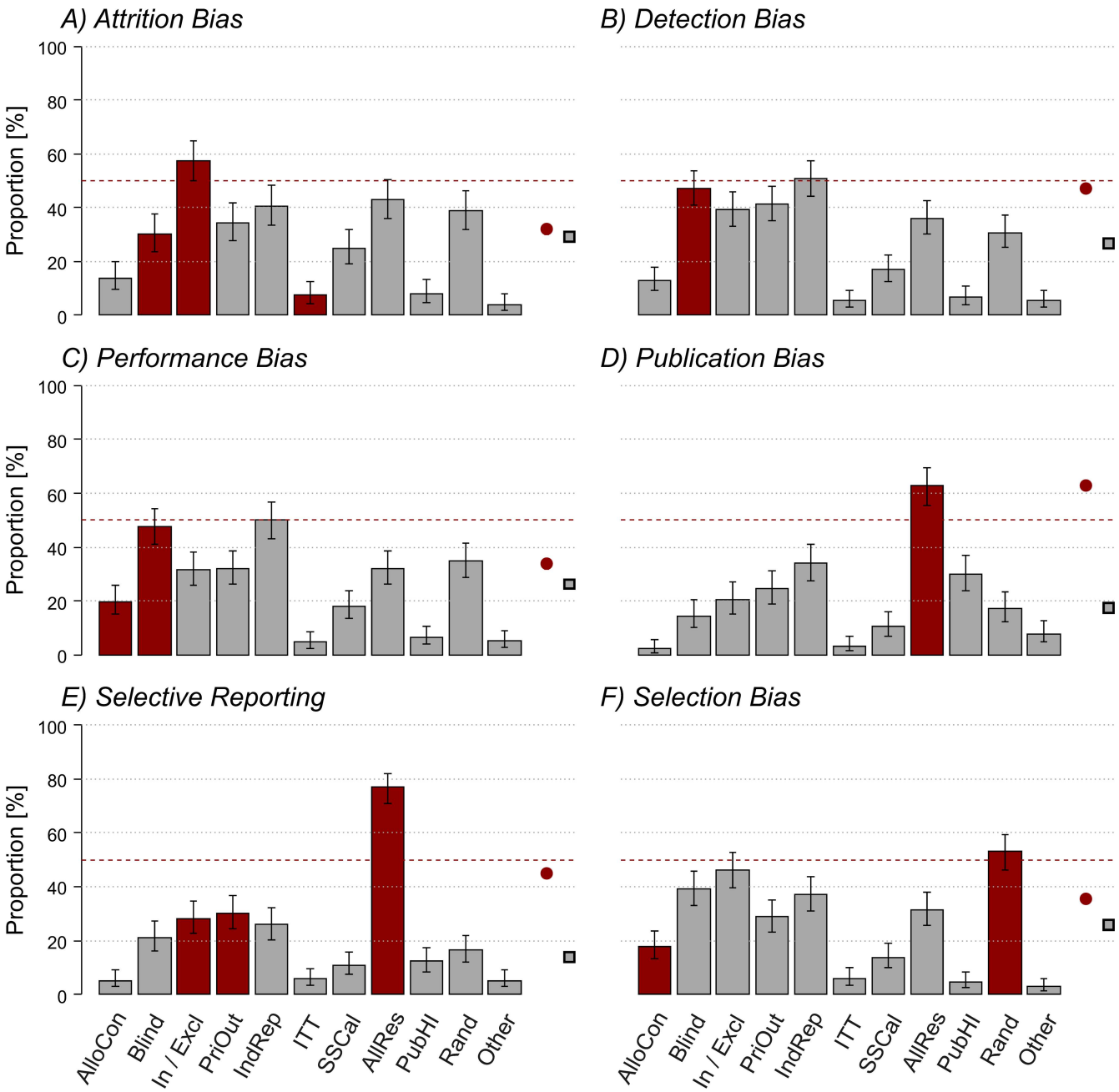

Fig 4. Experimental biases and measures to avoid them. Bars indicate percentage of participants (y-axis) giving that answer (corrected for participants choosing 'no answer'), red bars indicate effective measures to avoid a given bias (A-F), respectively. The red circle is indicative of the mean of effective measures (*sensu stricto* according to Table 1), while the grey rectangle is the mean of ineffective measures. Number of participants answering to questions: attrition bias N = 172; detection bias N = 224; performance bias N = 212; publication bias N = 180; selective reporting N = 213; selection bias N = 219. The list of possible answers (x-axis) included: AlloCon = allocation concealment; Blind = blinding; In / Excl = inclusion / exclusion criteria; PriOut = primary outcome variable; IndRep = independent replication; ITT = Intention-to-Treat analysis; SSCal = sample size calculation; AllRes = reporting of all results; PubHI = publishing in high impact journals; Rand = randomization; Other = other measures.

doi:10.1371/journal.pone.0165999.g004

Similar proportions of participants returning partially vs. fully completed questionnaires ascribed their research to basic research (51.2% vs. 58.9%), applied research (45.5% vs. 40.4%), or were undecided (2.4% vs. 0.7%), respectively (N = 167 partially completed questionnaires).

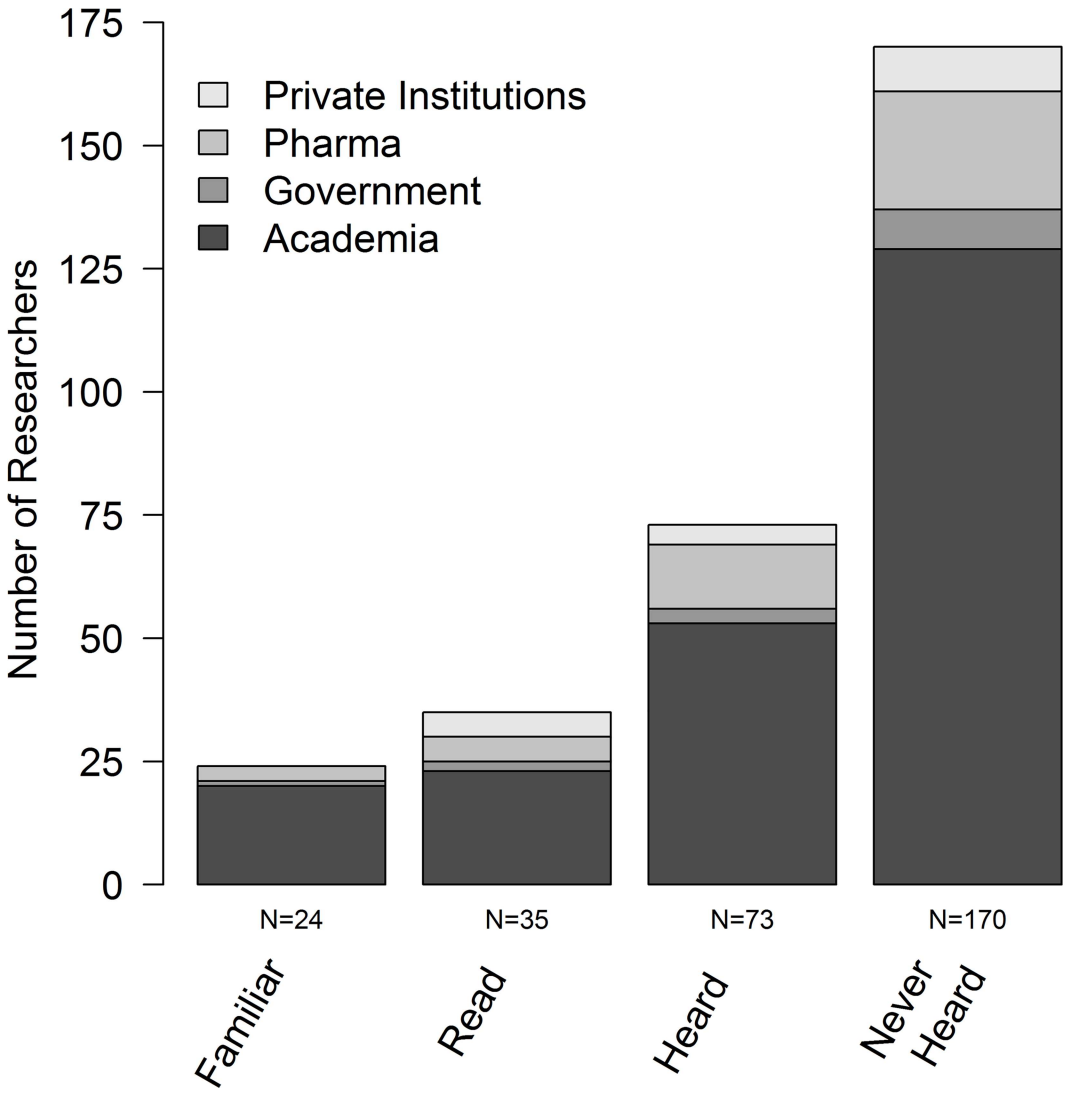

Fig 5. Knowledge of ARRIVE Guidelines by participants to the online survey. Shown are absolute numbers of participants being 'familiar' with the guidelines, 'having read' and 'having heard of' them, and having 'never heard of' the guidelines, split according to the participants' affiliations (private institutions, pharma, governmental institutions and academia).

doi:10.1371/journal.pone.0165999.g005

Also, similar proportions were involved in biomedical and medical research (85.6% vs. 86.8%; N = 167); were using mice (71.5% vs. 60.6%) and rats (13.9% vs. 15.6%; N = 165); and were affiliated with academic institutions (71.1% vs. 74.5%), industry (15.1% vs. 14.9%), governmental research institutions (5.7% vs. 4.6%), and private research institutions (8.2% vs. 6%; N = 159), respectively.

In terms of the primary outcome variable, IVS_{Exp}, the comparison between our study sample of fully completed questionnaires and that of partially completed questionnaires for which the necessary answers were available yielded identical results, with a mean IVS_{Exp} of 0.73 (SD ± 0.24; N = 301) for the study sample and a mean IVS_{Exp} of 0.73 (SD ± 0.20; N = 99) for the sample of partially completed questionnaires.

Discussion

Summary of results

Low reporting rates of measures against risks of bias in the primary literature are widely considered as a proxy measure of poor experimental conduct. Reporting guidelines (e.g., ARRIVE) have thus become a major weapon in the fight against risks of bias in animal research. Here we studied, for the first time, how reporting rates of measures against risks of bias in in vivo research (e.g. [23,26,28–30,33]) relate to the rates at which such measures are implemented, according to researchers' self-reports. Our findings indicate that scientific rigor of animal research may be considerably better than predicted by reporting rates, as researchers may be using measures against risks of bias to a much greater extent than suggested by systematic reviews of the published literature. The large discrepancy suggests that reporting rates may be poor predictors of scientific rigor in animal research. This is further supported by our finding that the rates at which researchers claimed to have reported measures against bias in their latest publication were considerably lower than the rates at which they claimed to have used these measures in their research.

On the other hand, we found a weak but positive correlation between self-reported use and self-reported reporting of measures against risks of bias, supporting findings from systematic reviews indicating that higher reporting rates reflect more rigorous research (e.g. [26,29,48]). Furthermore, self-reported reporting rates of measures against risks of bias in the researchers' latest publication were considerably higher than the reporting rates commonly found by systematic reviews. Taken together, these findings suggest that whereas reporting rates may underestimate scientific rigor, self-reports may overestimate it. The latter is further supported by our finding that the researchers' knowledge of risks of bias, and effective measures to prevent them, was rather limited. Thus, the discrepancy between reporting rates and self-reports may be partly explained by the researchers' ignorance of potential risks of bias and measures to prevent them.

Our findings, therefore, highlight a need for better education and training of researchers in good research practice to raise their awareness of risks of bias and improve their knowledge about measures to avoid them. Furthermore, they indicate a need for more reliable predictors of scientific rigor.

Validity of self-reports

The researchers' self-reports of their use of measures against risks of bias should be interpreted with caution, as self-reports may not necessarily reflect the true quality of experimental conduct. That the reporting rates of measures against bias claimed by the researchers for their latest publication were considerably higher than the reporting rates generally found by systematic reviews of the published literature (e.g., randomization 44% vs. 27%, sample size calculation 18% vs. 0.5%) indicates that the researchers' self-reports should not be taken at face value. There are two main ways in which the self-reports may be biased. First, our study population (participants having returned fully completed questionnaires) may differ from the overall population of in vivo researchers. For example, participants of the survey may be particularly conscious of risks of bias and the problem of poor reproducibility, which may have predisposed them to take part in this survey. This could explain better experimental conduct and better reporting, compared to the overall population. Alternatively, participants may have been prone to overestimate their own performance (e.g. [49]). We have only limited data to assess these two alternatives. However, when comparing the population of participants who returned fully completed questionnaires with the population of participants who started but

did not complete the questionnaire, we did not find any major differences in the characteristics of the participants (e.g., host institution, research animals, type of research), nor in the primary outcome variable of this study, the internal validity score for experimental conduct (IVS_{Exp}). Given that these two populations of participants together accounted for almost one third of the overall population of registered in vivo researchers in Switzerland, the difference in reporting rates between self-reports and systematic reviews are unlikely to be explained by a systematic bias towards better performers in our study sample. This is further supported by the fact that the participants performed rather poorly when asked about their knowledge of specific types of bias, and effective measures to avoid these. Overestimation of one's own performance tends to be the more pronounced, the less skilled and competent individuals are (i.e., the Kruger-Dunning Effect, [50]). Although researchers are generally highly skilled and competent in their field of research, the researchers' limited knowledge of types of bias and measures to avoid them renders their self-reports at risk for overestimation. We thus conclude that the difference between what researchers claimed to have reported in their latest paper and reporting rates found by systematic reviews are more likely explained by the researchers overestimating their own performance than a bias towards better performers in our study sample.

Subjective bias resulting in overestimation of their own performance may also have affected the researchers' self-reports on the actual use of measures against risks of bias. Thus, the true use of measures against risks of bias may lie anywhere between what has been found to be reported by systematic reviews, and the researchers' self-report presented here. Given the large difference between IVS_{Exp} and IVS_{Pub}, however, reporting rates found in the literature are likely to underestimate scientific rigor to a considerable extent.

Reasons for low reporting rates

The main reason for not reporting the use of measures against risks of bias in publications is that researchers do not find it necessary to report it. This was further corroborated by personal interviews. Thus, researchers argued, for example, that "certain things are self-evident and do not need to be reported", that "the journal did not request to describe it [e.g., randomization]", that "good scientific practice" actually implies that the criteria of good research practice are met without having to stress (i.e., report) this, or that "there is a threshold for what is relevant to the own laboratory and [what is relevant] to the research community outside the laboratory".

However, given the negative relationship between the reporting of measures against risks of bias and overstatement of treatment effect size (e.g. [28,29,30,40]), and the positive correlation between IVS_{Pub} and IVS_{Exp} found here, these statements appear questionable.

Although our findings suggest that scientific rigor in animal research may be considerably better than predicted by systematic reviews, there clearly is scope for improvement as, for example, only half of the participants self-reported using blinded outcome assessment (47%) or allocation concealment (52%). Blinding and allocation concealment, together with proper randomization procedures, are key measures to avoid selection bias and detection bias (cf. Table 1) and should be used in every study and reported in every publication (e.g. [5,31,51]).

Effect of knowledge of reporting guidelines on measures of scientific rigor

To assess the effects of specific characteristics of the researchers or their research on measures of scientific rigor, we calculated scores of experimental conduct (IVS_{Exp}) and reporting (IVS_{Pub}). Similar scores have previously been used to assess scientific rigor in systematic reviews and meta-analyses of reporting rates in the published literature (e.g., CAMARADES

checklist [24]). Variation in IVS_{Exp} was best explained by knowledge of the ARRIVE guidelines (yes vs. no) and type of research (applied vs. basic vs. other). Thus, researchers being familiar with the ARRIVE guidelines and researchers in applied research scored higher on IVS_{Exp}, and researchers knowing the ARRIVE guidelines also scored higher on IVS_{Pub}. These findings support the view that reporting guidelines may improve not only reporting but may actually improve the use of measures against risks of bias (e.g. [48]). The positive effect of applied research on IVS_{Exp} is more difficult to explain. It has previously been argued that the incentive for reliable results may be higher in applied research, for example in pharma research where also economic values are at stake (e.g. [19,52]). However, given the small size of this effect, and the fact that participants from academia and industry did not differ on both scores (IVS_{Exp}: academia = 0.73 vs. industry = 0.73, Wilcox test: W = 5243, p = 0.70; IVS_{Pub}: academia = 0.51 vs. industry = 0.46, Wilcox test: W = 3900, p = 0.40) suggests that it should be interpreted with caution.

Despite loud calls for better reporting (e.g. [53]) and the widespread endorsement of reporting guidelines by many scientific journals (e.g. [34,54–56]), reporting has not yet improved much [35]. Thus, without active enforcement of reporting guidelines by journal editors and reviewers, the situation may not change [57]. This is also confirmed by results of this study: more than half of the participants having published their latest article in a journal that has endorsed the ARRIVE guidelines admitted that they had never heard of these guidelines. This ignorance is surprising given the wide coverage that the ARRIVE guidelines have received and we may only speculate about the reason for this. Most likely, researchers can still ignore them–and may continue to do so–as long as the journals do not enforce them more strictly.

This may reflect a general attitude we observed among the scientists we interviewed. While they agreed that guidelines for the design and conduct of experiments may be useful, they were skeptical towards reporting guidelines. As one interviewee put it, "introducing more checklists to tick boxes does not increase the quality of science". Thus, publication checklists are perceived as a sign of increasing over-regulation and bureaucracy and may therefore be ignored. Similarly, Begley and Ioannidis [39] warned that the burden of bureaucracy might lead to normative responses without measurable benefits for the quality of research and reproducibility. However, Minnerup and colleagues [48] recently showed that the quality of research published in the journal *Stroke* increased after the implementation of the 'Basic Science Checklist'. Thus, if enforced by reviewers and editors, adequate checklists may well be conductive to the quality of research.

Knowledge of risks of bias and measures to avoid them

Increasing evidence of bias associated with poor experimental conduct and reporting (e.g. [5,23,26,29,30,33]) is only partly mirrored by the participants' answers to the questionnaire. Thus, only about two thirds of the participants (58–68%) indicated being aware of selective reporting, selection bias, detection bias, and publication bias, and less than half of them were actually 'concerned' about such biases (35–42%) with respect to their own research. Furthermore, between 15% and 25% indicated that these biases were 'not relevant' to their own research. These results reflect a certain ignorance of risks of bias in experimental conduct, combined with a lack of knowledge about these risks and about effective measures to avoid them. Thus, when participants were asked about effective measures against specific types of bias from a list of 10 potential measures, there was no consistent preference of effective over ineffective measures, except for publication bias (and, to some extent, for selective reporting). In particular, participants performed poorly when asked for measures against attrition bias, detection bias, and performance bias, respectively. This lack of understanding may have

contributed to the participants overestimating the quality of their own experimental conduct. Therefore, besides the implementation of reporting guidelines (e.g. [34,48,56,58,59]), which will raise awareness of risks of bias, we conclude that researchers may need better training in scientific integrity and good research practice in view of minimizing risks of bias in future research.

Conclusions

Our findings indicate that reporting rates of measures against risks of bias may not be reliable measures of scientific rigor in animal research, and that better measures are needed. However, although the researchers' self-reports suggest that the actual use of measures against risks of bias may be considerably higher than predicted by the low reporting rates in the published literature, self-reports may overestimate their true use. Indeed, the results presented here indicate that there may be considerable scope for improvement of scientific rigor in experimental conduct of animal research, and that concepts and methods of good research practice should play a more important role in the education of young researchers (e.g. [11]). It is quite possible that lack of scientific rigor contributes to the so called "reproducibility crisis" (e.g. [3]). However, scientific rigor in experimental conduct is not the only factor affecting reproducibility, and perhaps not even the most important one; poor construct validity of animal models (e.g. [9,17]) and poor external validity due to highly standardized laboratory conditions (e.g. [8,60–63]) are important alternative causes. Further research is therefore needed on the effects of different aspects of scientific validity on reproducibility, to assess their scope for improvement and in view of prioritizing strategies towards improvement beyond reporting guidelines.

Supporting Information

S1 Table. Overview of Certifications of Participants' Institutions.
(DOCX)

S2 Table. Full Results of Use and Reporting of Measures to Avoid Risk of Bias.
(DOCX)

S3 Table. A) Overview of Candidate Models and B) Model Outputs of Best Performing Models.
(DOCX)

S1 Text. Online Survey.
(DOCX)

S2 Text. Personal Interviews.
(DOCX)

Acknowledgments

The authors are grateful to Markus Zürcher for helpful input at the conception of this study and for comments on earlier drafts of this manuscript; to Isabel Lechner for the introduction to and help with Limesurvey; to Heinrich Binder, Sven Süptitz, and Corinne Grandjean from the Federal Food Safety and Veterinary Office (FSVO) for distributing the survey among animal researchers in Switzerland; and to the five interviewees, who wish to stay anonymous, for participation in this study and for their open and honest answers.

Author Contributions

Conceptualization: HW TSR.

Data curation: TSR.

Formal analysis: TSR.

Funding acquisition: HW.

Investigation: TSR.

Methodology: TSR HW.

Project administration: HW.

Supervision: HW.

Visualization: TSR HW.

Writing – original draft: TSR.

Writing – review & editing: TSR LV HW.

References

1. Prinz F, Schlange T, Asadullah K. Believe it or not: How much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. [Internet]. Nature Publishing Group; 2011 [cited 2013 Feb 27]; 10:712. Available: http://www.ncbi.nlm.nih.gov/pubmed/21892149 doi: 10.1038/nrd3439-c1 PMID: 21892149
2. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012; 483:531–3. doi: 10.1038/483531a PMID: 22460880
3. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat. Rev. Neurol. [Internet]. Nature Publishing Group; 2014 [cited 2014 Jan 29]; 10:37–43. Available: http://www.ncbi.nlm.nih.gov/pubmed/24247324 doi: 10.1038/nrneurol.2013.232 PMID: 24247324
4. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. [Internet]. 2004; 3:711–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/15286737 doi: 10.1038/nrd1470 PMID: 15286737
5. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med. [Internet]. 2010; 7:e1000245. Available: http://www.plosmedicine.org/article/fetchObject.action?uri=info%253Adoi%252F10.1371%252Fjournal.pmed.1000245&representation=PDF doi: 10.1371/journal.pmed.1000245 PMID: 20361020
6. Editor. Further confirmation needed. Nat. Biotechnol. 2012; 30:806. doi: 10.1038/nbt.2335 PMID: 22965029
7. McNutt M. Reproducibility. Science (80-.). 2014; 343:229.
8. Richter SH, Garner JP, Würbel H. Environmental standardization: Cure or cause of poor reproducibility in animal experiments? Nat. Methods. 2009; 6:257–61. doi: 10.1038/nmeth.1312 PMID: 19333241
9. Bailoo JD, Reichlin TS, Würbel H. Refinement of experimental design and conduct in laboratory animal research. ILAR J. [Internet]. 2014 [cited 2014 Dec 29]; 55:383–91. Available: http://ilarjournal.oxfordjournals.org/cgi/doi/10.1093/ilar/ilu037 doi: 10.1093/ilar/ilu037 PMID: 25541540
10. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature [Internet]. 2014 [cited 2014 Jun 3]; 505:612–3. Available: http://europepmc.org/abstract/MED/24482835 PMID: 24482835
11. Festing MFW. We are not born knowing how to design and analyse scientific experiments. Altern. to Lab. Anim. ATLA. 2013; 41:19–21.
12. Fanelli D. Do pressures to publish increase scientists' bias? An empirical support from US States data. PLoS One. 2010; 5:e10271. doi: 10.1371/journal.pone.0010271 PMID: 20422014
13. Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications. Proc. Natl. Acad. Sci. [Internet]. 2012 [cited 2013 Jun 7]; 110:1136–7. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1220833110

14. Ioannidis JPA. Why most published research findings are false. PLoS Med. [Internet]. 2005 [cited 2013 May 21]; 2:e124. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1182327&tool=pmcentrez&rendertype=abstract

15. Macleod MR. Why animal research needs to improve. Nature. 2011; 477:511. doi: 10.1038/477511a PMID: 21956292

16. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat. Rev. Cancer. 2005; 5:142–9. doi: 10.1038/nrc1550 PMID: 15685197

17. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. PLoS Med. [Internet]. 2013 [cited 2014 Feb 20]; 10:e1001489. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3720257&tool=pmcentrez&rendertype=abstract doi: 10.1371/journal.pmed.1001489 PMID: 23935460

18. Steckler T. Editorial: preclinical data reproducibility for R&D—the challenge for neuroscience. Springerplus [Internet]. 2015; 4:1. Available: http://www.ncbi.nlm.nih.gov/pubmed/25674489\nhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320139/pdf/40064_2014_Article_1534.pdf doi: 10.1186/2193-1801-4-1 PMID: 25674489

19. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. [Internet]. 2015; 13:e1002165. Available: http://dx.plos.org/10.1371/journal.pbio.1002165 doi: 10.1371/journal.pbio.1002165 PMID: 26057340

20. Freedman LP, Gibson MC. The impact of pharmacogenomics research on drug development. Clin. Pharmacol. Ther. 2015; 97:16–8. doi: 10.1002/cpt.9 PMID: 25670378

21. O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp HB, Howells DW. 1,026 experimental treatments in acute stroke. Ann. Neurol. [Internet]. 2006 [cited 2013 Jun 6]; 59:467–77. Available: http://www.ncbi.nlm.nih.gov/pubmed/16453316 doi: 10.1002/ana.20741 PMID: 16453316

22. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat. Rev. Drug Discov. [Internet]. Nature Publishing Group; 2012; 11:191–200. Available: http://www.ncbi.nlm.nih.gov/pubmed/22378269 doi: 10.1038/nrd3681 PMID: 22378269

23. Henderson VC, Demko N, Hakala A, MacKinnon N, Federico CA, Fergusson D, et al. A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib. Elife [Internet]. 2015; 4:1–13. Available: http://elifesciences.org/lookup/doi/10.7554/eLife.08351

24. Sena ES, van der Worp HB, Howells DW, Macleod MR. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. [Internet]. 2007 [cited 2013 May 15]; 30:433–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/17765332 doi: 10.1016/j.tins.2007.06.009 PMID: 17765332

25. McCann SK, Irvine C, Mead GE, Sena ES, Currie GL, Egan KE, et al. Efficacy of antidepressants in animal models of ischemic stroke: A systematic review and meta-analysis. Stroke [Internet]. 2014; 45:3055–63. Available: http://stroke.ahajournals.org/cgi/doi/10.1161/STROKEAHA.114.006304 doi: 10.1161/STROKEAHA.114.006304 PMID: 25184357

26. Macleod MR, Lawson McLean A, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, et al. Risk of bias in reports of in vivo research: A focus for improvement. PLoS Biol. [Internet]. 2015; 13:e1002273. Available: http://dx.plos.org/10.1371/journal.pbio.1002273 doi: 10.1371/journal.pbio.1002273 PMID: 26460723

27. Frantzias J, Sena ES, Macleod MR, Al-Shahi Salman R. Treatment of intracerebral hemorrhage in animal models: meta-analysis. Ann. Neurol. [Internet]. 2011 [cited 2013 Jun 28]; 69:389–99. Available: http://www.ncbi.nlm.nih.gov/pubmed/21387381 doi: 10.1002/ana.22243 PMID: 21387381

28. Rooke EDM, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: A systematic review and meta-analysis. Parkinsonism Relat. Disord. [Internet]. Elsevier Ltd; 2011 [cited 2013 Jun 26]; 17:313–20. Available: http://www.ncbi.nlm.nih.gov/pubmed/21376651 doi: 10.1016/j.parkreldis.2011.02.010 PMID: 21376651

29. Vesterinen HM, Sena ES, Ffrench-Constant C, Williams A, Chandran S, Macleod MR. Improving the translational hit of experimental treatments in multiple sclerosis. Mult. Scler. 2010; 16:1044–55. doi: 10.1177/1352458510379612 PMID: 20685763

30. Currie GL, Delaney A, Bennett MI, Dickenson AH, Egan KJ, Vesterinen HM, et al. Animal models of bone cancer pain: Systematic review and meta-analyses. Pain [Internet]. International Association for the Study of Pain; 2013 [cited 2013 Aug 7]; 154:917–26. Available: http://www.ncbi.nlm.nih.gov/pubmed/23582155 doi: 10.1016/j.pain.2013.02.033 PMID: 23582155

31. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, et al. The need for randomization in animal trials: An overview of systematic reviews. PLoS One. 2014; 9.

32. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke [Internet]. 2004 [cited 2013 Apr 29]; 35:1203–8.

Available: http://www.ncbi.nlm.nih.gov/pubmed/15060322 doi: 10.1161/01.STR.0000125719.25853.20 PMID: 15060322

33. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One [Internet]. 2009 [cited 2013 May 22]; 4:e7824. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2779358&tool=pmcentrez&rendertype=abstract doi: 10.1371/journal.pone.0007824 PMID: 19956596

34. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. PLoS Biol. [Internet]. 2010 [cited 2013 May 24]; 8:e1000412. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2893951&tool=pmcentrez&rendertype=abstract doi: 10.1371/journal.pbio.1000412 PMID: 20613859

35. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: Journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. Eisen JA, editor. PLoS Biol. [Internet]. Public Library of Science; 2014 [cited 2014 May 26]; 12:e1001756. Available: http://dx.plos.org/10.1371/journal.pbio.1001756 doi: 10.1371/journal.pbio.1001756 PMID: 24409096

36. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: Why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. [Internet]. Nature Publishing Group; 2013 [cited 2013 Nov 6]; 14:365–76. Available: http://www.ncbi.nlm.nih.gov/pubmed/23571845 doi: 10.1038/nrn3475 PMID: 23571845

37. Nuzzo R. Statistical errors. Nature. 2014; 506:150–2. doi: 10.1038/506150a PMID: 24522584

38. Vaux DL. Know when your numbers are significant. Nature. 2012; 492:180–1. doi: 10.1038/492180a PMID: 23235861

39. Begley CG, Ioannidis JPA. Reproducibility in science: Improving the standard for basic and preclinical research. Circ. Res. [Internet]. 2015; 116:116–26. Available: http://circres.ahajournals.org/cgi/doi/10.1161/CIRCRESAHA.114.303819 doi: 10.1161/CIRCRESAHA.114.303819 PMID: 25552691

40. Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke [Internet]. 2008 [cited 2013 Mar 11]; 39:2824–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/18635842 doi: 10.1161/STROKEAHA.108.515957 PMID: 18635842

41. Pound P, Bracken MB. Is animal research sufficiently evidence based to be a cornerstone of biomedical research? BMJ Br. Med. J. [Internet]. 2014 [cited 2014 Sep 4]; 348:g3387. Available: http://www.ncbi.nlm.nih.gov/pubmed/24879816

42. McNutt M. Journals unite for reproducibility. Nature. 2014; 515:7.

43. LimeSurvey Project Team. LimeSurvey: An open source survey tool [Internet]. Schmitz C, editor. Hamburg, Germany: LimeSurvey Project; 2012. Available: http://www.limesurvey.org

44. Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Methods Res. 2004; 33:261–304.

45. Aho K, Derryberry D, Peterson T. Model selection for ecologists: The worldviews of AIC and BIC. Ecology. 2014; 95:631–6. PMID: 24804445

46. Barton K. MuMIn: Multi-model inference [Internet]. 2014. Available: http://cran.r-project.org/package=MuMIn

47. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available: http://www.r-project.org/

48. Minnerup J, Zentsch V, Schmidt A, Fisher M, Schäbitz W- R. Methodological quality of experimental stroke studies published in the Stroke Journal. Stroke [Internet]. 2016; 47:267–72. Available: http://stroke.ahajournals.org/lookup/doi/10.1161/STROKEAHA.115.011695 doi: 10.1161/STROKEAHA.115.011695 PMID: 26658439

49. Ehrlinger J, Dunning D. How chronic self-views influence (and potentially mislead) estimates of performance. J. Pers. Soc. Psychol. 2003; 84:5–17. PMID: 12518967

50. Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. 1999; 7:1121–34.

51. Bello S, Krogsbøll LT, Gruber J, Zhao ZJ, Fischer D, Hróbjartsson A. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. J. Clin. Epidemiol. [Internet]. 2014 [cited 2014 Aug 25]; 67:973–83. Available: http://www.ncbi.nlm.nih.gov/pubmed/24972762 doi: 10.1016/j.jclinepi.2014.04.008 PMID: 24972762

52. Howells DW, Sena ES, O'Collins V, Macleod MR. Improving the efficiency of the development of drugs for stroke. Int. J. Stroke [Internet]. 2012 [cited 2013 Jun 4]; 7:371–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/22712738 doi: 10.1111/j.1747-4949.2012.00805.x PMID: 22712738

53. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature [Internet]. Nature Publishing Group; 2012 [cited 2013 Feb 27]; 490:187–91. Available: http://www.ncbi.nlm.nih.gov/pubmed/23060188 doi: 10.1038/nature11556 PMID: 23060188

54. Cressey D. Surge in support for animal-research guidelines. Nature. News [Internet]. 2016; Available: http://www.nature.com/doifinder/10.1038/nature.2016.19274

55. Kontinen VK. From clear reporting to better research models. Scand. J. Pain [Internet]. Elsevier B.V.; 2013 [cited 2013 Oct 9]; 4:57. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877886013000074

56. Editor. Reducing our irreproducibility. Nature. 2013; 496:198.

57. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet [Internet]. Elsevier Ltd; 2014 [cited 2014 Jan 20]; 383:267–76. Available: http://www.ncbi.nlm.nih.gov/pubmed/24411647 doi: 10.1016/S0140-6736(13)62228-X PMID: 24411647

58. Kontinen V. Raising the standards of preclinical pain studies. Scand. J. Pain [Internet]. Elsevier B.V.; 2015; 7:38–9. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877886015000063

59. Mullane K, Enna SJ, Piette J, Williams M. Guidelines for manuscript submission in the peer-reviewed pharmacological literature. Biochem. Pharmacol. [Internet]. 2015; 97:225–35. Available: http://www.sciencedirect.com/science/article/pii/S0006295215003585 doi: 10.1016/j.bcp.2015.06.023 PMID: 26208784

60. Würbel H. Behaviour and the standardization fallacy. Nat. Genet. [Internet]. 2000; 26:263. Available: http://www.ncbi.nlm.nih.gov/pubmed/11062457 doi: 10.1038/81541 PMID: 11062457

61. Würbel H. Behavioral phenotyping enhanced–beyond (environmental) standardization. Genes, Brain Behav. [Internet]. 2002; 1:3–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/12886944

62. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nat. Methods [Internet]. Nature Publishing Group; 2010 [cited 2013 Apr 4]; 7:167–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/20195246 doi: 10.1038/nmeth0310-167 PMID: 20195246

63. Voelkl B, Würbel H. Reproducibility crisis: Are we ignoring reaction norms? Trends Pharmacol. Sci. 2016; 37:509–10. doi: 10.1016/j.tips.2016.05.003 PMID: 27211784