## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | A systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. |
| **AUTHORS** | **Vollert, Jan; Schenker, Esther; Macleod, Malcolm; Bespalov, Anton; Wuerbel, Hanno; Michel, Martin; Dirnagl, Ulrich; Potschka, Heidrun; Waldron, Ann-Marie; Wever, Kimberley; Steckler, Thomas; van de Casteele, Tom; Altevogt, Bruce; Sil, Annesha; Rice, Andrew SC** |

## VERSION 1 – REVIEW

| | |
|---|---|
| **REVIEWER 1** | ***Kimmelman, Jonathan***<br>***McGill*** |
| **REVIEW RETURNED** | 15-10-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | This paper reports the results of a systematic review of recommendations contained in guidelines for design and conduct of preclinical efficacy studies. Briefly, the report finds 60 Guidelines, expressing 58 different recommendations. As the lead author for one of the first such systematic reviews (reference 5 in this paper), I endorse the spirit of this effort; much has changed since my team published our report. I also appreciate the tremendous practical as well as conceptual challenges in pulling off such a systematic review.<br><br>In addition to endorsing the spirit of this systematic review and the importance of the findings, there are a number of other features to commend the manuscript. For example, the pre-registration of a protocol (along with amendments in the methods section) is an improvement over our own work, as is the process used for screening articles and diligence determination. The observation that recommendations are not typically evidence based; that practices like randomization may be more of a surrogate of quality rather than quality itself, and the costs of excessive standards all seem very sensible.<br><br>Having said that (and again, drawing on my own experience in reference 5) there were a number of aspects about this report that, in my opinion, were a bit puzzling and should be addressed in a letter and/or revisions.<br><br>- The method of search for guidelines seems to have relied exclusively on academic literature databases. Such searching is likely to miss policies established by foundations, funding bodies, regulatory bodies, or scientific societies. For example, FDA has (somewhat lame) guidance policies for design of preclinical gene and cell therapy studies. ISSCR has standards for design of preclinical cell therapy studies (I know this because I wrote those |

| | |
|---|---|
| | standards). NIH has policies for reporting of preclinical studies (which may have been deemed relevant). Nature, as the authors know, now has a template for submission that obligates reporting of certain experimental details; this seems relevant as well. None of the above are captured- but could easily have been had the authors used a multi-pronged search strategy (e.g. citation analysis, Google searches, or soliciting opinion of outside (non-co-author, that is) experts. The authors note they initially planned Google searches but found it too laborious. But surely there are other search methods available, and other ways of fishing things out of Google. For example, in our own search, I believe we capped the number of hits screened from a Google search.<br><br>- Am also not clear why the authors used “Guideline or recommendation(s)” in the search string, but not terms like “Guidance,” “Standard,” or “Policy.” The challenge with doing SRs of this type is that language is not at all standardized as it is in medicine; as such, searches need to be operationalized in different ways.<br><br>- I was surprised the authors used single extraction. When we conducted our SR, we found extraction involved considerable judgment and interpretation. Recommendations and language in guidelines often bled into each other or were ambiguous, raising questions about lumping and splitting (e.g. when a policy says you should ‘blind’- does it mean blinded assignment or blinded outcome assessment or both? When a policy calls for a priori power calculations, does that mean it calls for a priori statement of hypothesis AND justification of sample size?). There was also lots of potential for extractors to overlook recommendations because they were in tables and not text, say. Double extraction is a huge amount of work and it is no guarantee for accuracy, but there needs to be some way of assuring the reader that the numbers and percentages in the manuscript are accurate, given the subjectivity and operational fallibility of extraction.<br><br>- I don’t completely understand the diligence classification (though it seems like a great idea). It seems as if each level has two variables determining it: method (e.g. systematic review or expert opinion) and group size (small group vs. “groups of individuals”). So if there is a large group (not level 1, since that pertains to small groups) that offers recommendation, but it doesn’t use Delphi (thus not level 2), how is the guideline classified? Also not clear how diligence was determined- for example, how endorsement was searched. I’m sure there is a clear explanation but right now this is opaque to me. Also for level 2 (and on page 11), why is Delphi process singled out as a method of opinion aggregation? Many people in Medicine seem to think Delphi represents a gold standard for aggregating expert opinion. What is the evidence supporting this? Delphi is at best controversial, and at worst discredited as a method of opinion aggregation by Decision Scientists, as the approach amplifies groupthink and bias. Due respect to the authors here, but the endorsement of Delphi in medicine (and here) seems sort of mindless.<br><br>- Small suggestion: since this article does not concern tox studies, I wonder if the title should be more specific, since “involving laboratory animals” is quite broad. Further to this, there are no doubt lots of lab animal experiment types that guidelines here are not directed to (e.g. more basic science studies). The introduction states the goal of assessing preclinical guidelines. |

| | |
|---|---|
| | Bottom line: important and timely update on guidelines for preclinical efficacy study design and conduct. Employs many compelling methodological features that commend this manuscript as compared with others (including my own). However, there are important methodological limitations that, in my view, need to be addressed.<br><br>-Jonathan Kimmelman |

| | |
|---|---|
| **REVIEWER 2** | ***Marc Avey***<br>***Public Health Agency of Canada, Global Health and Guidelines*** |
| **REVIEW RETURNED** | 31-10-2019 |

| | |
|---|---|
| **GENERAL COMMENTS** | The is an excellent systematic review (or scoping review) on a timely topic. The review follows the protocol and notes where amendments were made. The results are descriptive and of interest. The discussion (and design) focuses primarily on how these elements are published/endorsed and not whether they have conceptual validity or empirical evidence that support each supposed risk of bias element. Hopefully the authors will follow-up with de novo primary studies that explore these issues as they indicate in their closing sentences.<br><br>Pg 3 Line 26: The authors aim in the protocol was to: "The aim of this systematic review is to identify and harmonise existing experimental design, conduct and analysis guidelines relating to preclinical animal research." I do not see any discussion of harmonization in manuscript and it was dropped from the introduction.<br><br>Pg 3 Line 33: In the protocol the author indicate they will focus on both internal validity and reproducibility but that sentence is removed here. I do not see the results discussed in reference to reproducibility.<br><br>Pg 5 Line 26-27: Is the extraction form included here? [I see it is in box 1 but it's not linked here].<br><br>Pg 7 Line 42: Neither the protocol nor manuscript provide a rationale about why using a Delphi process is important . Is it simply because it is recommend for the development of clinical reporting guidelines (e.g. https://doi.org/10.1371/journal.pmed.1000217), or is it to reduce bias & noise from groups of experts (e.g. Dalkey 1967), or something else?<br><br>Page 7 Line 40: Editorial comment: Provenance and validity of recommendations may be inversely correlated. |

## VERSION 1 – AUTHOR RESPONSE

Dear Editor, dear Editor-in-Chief,

Thank you very much for your reply. We are grateful to you, the Section Editor and both reviewers for these very positive evaluations. We have revised the manuscript according to the reviewer’s comments. All changes are described en detail per comment below. The modifications of our manuscript are shaded

in yellow to allow easy recognition, these are only the changes made in this revision, or relevant to this revision, if introduced before.

Reviewer: 1

- Point1: regarding google searches issue raised by the reviewer, the authors mentioned that “added this to the limitations section on page 8”. However, I could not see this (anything related to google searches) on page 8.

Author’s response: The limitations section on page 8 of the original submission included the phrasing “Additionally, our plan to search the websites of professional organizations and funding bodies failed due to reasons of practicality.”

At the review stage, we added: “Although being aware of single recommendations outside of publication, we did not include those to keep methods reproducible.”

At this second review stage, we added the following statement: “Limiting the results included from a google search would have been a practical solution to overcome this issue, which we failed to decide at protocol generation”.

Point4: I feel that the authors have not addressed the comments by the reviewer properly: at least I could not see any changes made in the main text accordingly. Even the authors said that they altered the phrasing to “a Delphi process or other means of structured decision making”, I couldn’t see this as highlighted yellow in the main text.

Author’s response: We could have referenced this more clearly, apologies. The changes during the first review can be found in the methods under the section “Extraction, aggregation and diligence classification” on the pages 5 and 6. Additionally, in the discussion on page 7, we clarified that a Delphi process is suggested for clinical guidelines to reduce bias.

In this second revision, we now added the phrasing “or other means of structured decision making” on page 7 as well.

Reviewer: 2

The authors need to add the sentence “The harmonization process…will be published separately” in the revised manuscript.

Author’s response: We have added the following sentences to the deviation from the protocol section at the beginning of the methods section on pages 3-4: “In the protocol, we mention that the aim of this systematic review is an effort to harmonize guidelines and create a unified framework. This is still underway and will be published separately.”

On behalf of all authors
Yours sincerely

Jan Vollert, London, 10/12/2019

# A systematic review of guidelines for internal validity in the design, conduct and analysis of biomedical experiments involving laboratory animals.

Authors: Jan Vollert (1), Esther Schenker (2), Malcolm Macleod (3), Anton Bespalov (4,5), Hanno Wuerbel (6), Martin C Michel (4,7), Ulrich Dirnagl (8), Heidrun Potschka (9), Ann-Marie Waldron (9), Kimberley E Wever (10), Thomas Steckler (11), Tom Van de Casteele (11), Bruce Altevogt (12), Annesha Sil (13) and Andrew SC Rice (1) on behalf of the EQIPD consortium

1 Pain Research, Department of Surgery and Cancer, Imperial College London, London, UK
2 Institut de Recherches Servier, Croissy-sur- Seine, France
3 Centre for Clinical Brain Sciences, University of Edinburgh, UK
4 Partnership for Assessment and Accreditation of Scientific Practice, Heidelberg, Germany
5 Valdman Institute of Pharmacology, Pavlov Medical University, St. Petersburg, Russia
6 Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland
7 Universitätsmedizin Mainz, Johannes-Gutenberg-Universität Mainz, Mainz, Germany
8 Department of Experimental Neurology, Charité Universitätsmedizin Berlin, Germany
9 Institute of Pharmacology, Toxicology, and Pharmacy, Ludwig-Maximilians- University, Munich, Germany
10 Systematic Review Centre for Laboratory Animal Experimentation, Department for Health Evidence, Nijmegen Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands
11 Janssen Pharmaceutica NV, Beerse, Belgium
12 Pfizer Inc.
13 Institute of Medical Sciences, University of Aberdeen, UK

**Abstract**

Over the last two decades, awareness of the negative repercussions of flaws in the planning, conduct and reporting of preclinical research involving experimental animals has been growing. Several initiatives have set out to increase transparency and internal validity of preclinical studies, mostly publishing expert consensus and experience. While many of the points raised in these various guidelines are identical or similar, they differ in detail and rigour. Most of them focus on reporting, only few of them cover the planning and conduct of studies. The aim of this systematic review is to identify existing experimental design, conduct, analysis and reporting guidelines relating to preclinical animal research. A systematic search in Pubmed, EMBASE and Web of Science retrieved 13,863 unique results. After screening these on title and abstract, 613 papers entered the full-text assessment stage, from which 60 papers were retained. From these, we extracted unique 58 recommendations on the planning, conduct and reporting of preclinical animal studies. Sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment were recommended in more than half of the publications. While we consider these recommendations to be valuable, there is a striking lack of experimental evidence on their importance and relative effect on experiments and effect sizes.

## Introduction

In recent years, there has been growing awareness of the negative repercussions of shortcomings in the planning, conduct and reporting of preclinical animal research [1 2]. Several initiatives involving academic groups, publishers and others have set out to increase the internal validity and reliability of primary research studies and the resulting publications. Additionally, several experts or groups of experts across the biomedical spectrum have published experience and opinion-based guidelines and guidance. While many of the points raised are broadly similar between these various guidelines (probably in part reflecting the observation that many experts in the field are part of more than one initiative), they differ in detail, rigour, and in particular whether they are broadly generalizable or specific to a single field. While all these guidelines cover the reporting of experiments, only a few specifically address rigorous planning and conduct of studies [3 4], which might increase validity from the earliest possible point [5]. Consequently, it is difficult for researchers to choose which guidelines to follow, especially at the stage of planning future studies.

We aimed to identify all existing guidelines and reporting standards relating to experimental design, conduct and analysis of preclinical animal research. We also sought to identify literature describing (either through primary research or systematic review) the prevalence and impact of perceived risks of bias pertaining to the design, conduct and analysis and reporting of preclinical biomedical research. While we focus on internal validity as influenced by experimental design, conduct and analysis we recognise that factors such as animal housing and welfare are highly relevant to the reproducibility and generalizability of experimental findings; however, these factors are not considered in this systematic review.

## Methods

The protocol for this systematic review has been published in [6]. The following amendments to the systematic review protocol were made: In addition to the systematic literature search, to capture standards set by funders or organisations that are not (or not yet) published, it was planned to conduct a google search for guidelines published on the websites of major funders and professional organisations using the systematic search string below [6]. This search, however, yielded either no returns, or, in the case of the National Institute of Health, identified over 193,000 results, which was an unfeasibly large number to screen. Therefore, for practical reasons this part of the search was excluded from the initial search strategy.

**Search strategy**

We systematically searched PubMed, Embase via Ovid and Web of Science to identify guidelines published in English language in peer-reviewed journals before January 10th, 2018 (the day the search was conducted), using appropriate terms for each database optimized from the following search string (as can be found in the protocol [6]):

(guideline OR recommendation OR recommendations) AND
("preclinical model" OR "preclinical models" OR "disease model" OR "disease models" OR "animal model" OR "animal models" OR "experimental model" OR "experimental models" OR "preclinical study" OR "preclinical studies" OR "animal study" OR "animal studies" OR "experimental study" OR "experimental studies") [6].

Furthermore, as many of the researchers participating in the EQIPD (European Quality in Preclinical Data) project (http://eqipd.org/) are experts in the field of experimental standardization, they were contacted personally to identify additional relevant publications.

**Inclusion and exclusion criteria**

We included all articles or systematic reviews in English which described or reviewed guidelines making recommendations intended to improve the validity or reliability (or both) of preclinical animal studies through optimising their design, conduct and analysis. Articles that focussed on toxicity studies or veterinary drug testing were not included. Although reporting standards were not the key primary objective of this systematic review these were also included, as they might contain useful relevant information.

**Screening and data management**

We combined the search results from all sources and identified duplicate search returns and the publication of identical guidelines by the same author group in several based on the PubMed ID, DOI, and the title, journal and author list. Unique references were then screened in two phases: 1) screening for eligibility based on title and abstract, followed by 2) screening for definitive inclusion based on full text. Screening was performed using the Systematic Review facility (SyRF) platform (http://syrf.org.uk). Ten reviewers contributed to the screening phase; each citation was presented to two independent reviewers with a real-time computer-generated random selection of the next citation to be reviewed. Citations remained available for screening until two reviewers agreed that is should be included or excluded. If the first two reviewers had disagreed the citation was offered to a third, but reviewers were not aware of previous screening decisions. A citation could not be offered to the same reviewer twice. Reviewers were not blinded to the authors of the presented record. In the first stage, two authors screened the title and abstract of the retrieved records for

eligibility based on predefined inclusion criteria (see below). The title/abstract screening stage aimed to maximise sensitivity rather than specificity – any paper considered to be of any possible interest was included.

Articles included after the title-abstract screening were retrieved as full-texts. Articles for which no full-text version could be obtained were excluded from the review. Full texts were then screened for definite inclusion and data extraction. At both screening stages, disagreements between reviewers were resolved by additional screening of the reference by a third adjudicating reviewer, who was unaware of the individual judgements of the first two reviewers. All data were stored on the SyRF platform.

**Extraction, aggregation and diligence classification**

From the publications identified, we extracted recommendations on the planning, conduct and reporting or preclinical animal studies as follows:
Elements of the included guidelines were identified using an extraction form inspired by the results from Henderson et al [5]. Across guidelines, the elements were ranked based on the number of guidelines in which that element appeared. Extraction was not done in duplicate, but only once. As the extracted results in this case are not quantitative, but qualitative, meta-analysis and risk of bias assessment are not appropriate for this review. Still, we applied a diligence classification of the guidelines based on the following system, improving level of evidence from 1 to 3 and support from a to b:

1 Recommendations of individuals or small groups of individuals based on individual experience only
    a. Published stand-alone
    b. Endorsed or initiated by at least one publisher or scientific society
2 Recommendations by groups of individuals, through a method which included a Delphi process
    a. Published stand-alone
    b. Endorsed or initiated by at least one publisher or scientific society
3 Recommendations based on a systematic review
    a. Published stand-alone
    b. Endorsed or initiated by at least one publisher or scientific society

## Results

### Search and study selection

A flow chart of the search results and screening process is in Figure 1. Our systematic search returned 13,863 results, with 3,573 papers from PubMed, 5,924 from Web of Science, and 5,982 from EMBASE. After first screening on title and abstract, 828 records were eligible for the full-text screening stage. After removing duplications (69), non-English resources (48), conference abstracts (25), book chapters (14), and announcements (4), 676 records remained. Of these, 62 publications were retained after full-text screening. We later identified two further duplicate publications of the same guidelines in different journals, giving a final list 60 publications. [5 7-65].

The project members did not identify any additional papers that had not been identified by the systematic search.

### Diligence classification

More than half of the included publications (32) were narrative reviews that fell under the 1a category of our rating system (recommendations of individuals or small groups of individuals based on individual experience only, published standalone) [7 9 10 14 15 18 20 25 27 29 30 33 35 36 39 41-43 45 47-55 57 60 61 65]. An additional 22 publications were consensus papers or proceedings of consensus meetings for journals or scientific or governmental organizations (category 1b) [3 4 8 12 13 17 19 24 26 28 32 34 37 38 44 46 56 59 62-64 66]. None of these reported the use of a Delphi process or systematic review of existing guidelines. The remaining six publications were systematic reviews of the literature (category 3a) [5 11 21 31 40 58].

### Extracting components of published guidance

From the 60 publications finally included, we extracted 58 unique recommendations on the planning, conduct and reporting of preclinical animal studies. The absolute and relative frequency for each of the extracted recommendations is provided in table 1. Sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment were recommended in more than half of the publications. Only a few publications (≤five) mentioned pre-registration of experimental protocols, research conducted in large consortia, replication at different levels of disease or by variation in treatment, and optimization of complex treatment parameters. The extraction form allowed the reviewers in free text fields to identify and extract additional recommendations not covered in the pre-specified list, but this facility was rarely used, with only “publication of negative results” and “clear specification of exclusion criteria” extracted in this way by more than one

reviewer. The full results table of this stage is published as csv file on figshare under the DOI 10.6084/m9.figshare.9815753.

**Discussion**

Based on our systematic literature search and screening using predefined inclusion and exclusion criteria, we identified 60 published guidelines for the planning, conduct or reporting of preclinical animal research. From these publications, we extracted a comprehensive list of 58 experimental rigour recommendations that the authors had proposed as being important to increase the internal validity of animal experiments. Most recommendations were repeated in a relevant proportion of the publications (sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment in more than half of the cases), showing that there is at least some consensus for those recommendations. In many cases this may be because authors are on more than one of the expert committees for these guidelines, and many of them build on the same principles and cite the same sources of inspiration (i.e., doing for the field what CONSORT did for clinical trials [66 67]). There are also reasons why the consensus was not universal – many of the publications focus on single aspects (e.g. statistics [21] or sex differences [60]) or specific medical fields or diseases [13 37 38 63]. In addition, the narrative review character of many of the publications may have led to authors focusing on elements they considered more important than others.

Indeed, more than half (32 out of 60) of the publications reviewed here were topical reviews by a small group of authors (usually fewer than five). Another 22 (37%) were proceedings of consensus meetings or consensus papers set in motion by professional scientific or governmental organizations. It is noteworthy that none of these publications provide any rationale or justification for the validity or provenance of their recommendations. None used a Delphi process to structure decision making, and none reported using a systematic review of existing guidelines to inform themselves about literature. Of course, many of these expert groups will have been informed by pre-existing reviews (the remaining six included here were systematic literature reviews). However, there is a consistent feature across recommendations – that the steps recommended to increase validity are considered to be self-evident, and a basis in experiments and evidence is seldom linked or provided. There are hints that applying these principles does contribute to internal validity, as it has been shown that the reporting of measures to reduce risks of bias is associated with smaller outcome effect sizes [68], while other studies have not found such [69]. However, it is unclear if these measures taken are the perfect ones to reduce bias, or if they are merely surrogate markers for more awareness and thus more thorough research conduct. We consider this to be problematic for

at least two reasons: first, to increase compliance with guidelines it is crucial to keep them as simple and as easy to implement as possible. An endless checklist can easily lead to fatalistic thinking in researchers desperately wanting to publish, and it could be debated whether guidelines are seen by some researchers as hindering their progression rather than being an aide to conducting the best possible science, still, there is a difference between an 'endless' list and a 'minimal set of rules' that guarantees good research reproducibility. Secondly, each procedure that is added to experimental setup can in itself lead to sources of variation, so these should be minimized unless it can be shown that they add value to experiments.

Compliance is a significant problem for guidelines, as recently reported with the widely adopted ARRIVE guidelines of the UK's National Centre for the 3Rs [66 70]. This is not attributed to blind spots in the ARRIVE guidelines. While enforcement by endorsing journals may be important [71 72], a recent randomized blinded controlled study suggests that even an insistence of completing an ARRIVE checklist has little or no impact on reporting quality [73]. We believe that training and availability of tools to improve research quality will facilitate implementation of guidelines over time, as they become more prominent in researchers' mindset.

This systematic review has important limitations. Protocol wise, we only included publications in English language, reflecting the limited language pool of our team. Our broad search strategy identified more than 13,000 results, but we did not identify reports or systematic reviews of primary research showing the importance of specific recommendations [74], which must reflect a weakness in our search strategy. Additionally, our plan to search the websites of professional organizations and funding bodies failed due to reasons of practicality. Hence, we cannot ascertain whether we have included all important sources of literature. As hinted above, the results presented here also only paint an overview of the literature consensus, which should by no means be mistaken for an absolute ground truth of which steps need to be taken to improve internal validity in animal experiments. Indeed, literature debating the quality of these measures is sparse, and many of them have been borrowed from the clinical trials community or been considered self-evident from the literature. There is an urgent need for experimental testing of the importance of most of these measures, to provide better evidence of their effect.

**Acknowledgment**

We thank Alice Tillema of Radboud University, Nijmegen, The Netherlands, for her help in constructing and optimising the systematic search strings. This work is part of the European Quality

In Preclinical Data (EQIPD) consortium. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The EQIPD WP3 study group members are: Jan Vollert, Esther Schenker, Malcolm Macleod, Judi Clark, Emily Sena, Anton Bespalov, Bruno Boulanger, Gernot Riedel, Bettina Platt, Annesha Sil, Martien J Kas, Hanno Wuerbel, Bernhard Voelkl, Martin C Michel, Mathias Jucker, Bettina M Wegenast-Braun, Ulrich Dirnagl, René Bernard, Esmeralda Heiden, Heidrun Potschka, Maarten Loos, Kimberley E Wever, Merel Ritskes-Hoitinga, Tom Van De Casteele, Thomas Steckler, Pim Drinkenburg, Juan Diego Pita Almenar, David Gallacher, Henk Van Der Linde, Anja Gilis, Greet Teuns, Karsten Wicke, Sabine Grote, Bernd Sommer, Janet Nicholson, Sanna Janhunen, Sami Virtanen, Bruce Altevogt, Kristin Cheng, Sylvie Ramboz, Emer Leahy, Isabel A Lefevre, Fiona Ducrey, Javier Guillen, Patri Vergara, Ann-Marie Waldron, Isabel Seiffert and Andrew SC Rice.

**Box 1 – Extraction form**

1. Matching or balancing treatment allocation of animals
2. Matching or balancing sex of animals across groups
3. Standardized handling of animals
4. Randomized allocation of animals to treatment
5. Randomization for analysis
6. Randomized distribution of animals in the animal facilities
7. Monitoring emergence of confounding characteristics in animals
8. Specification of unit of analysis
9. Addressing confounds associated with anaesthesia or analgesia
10. Selection of appropriate control groups
11. Concealed allocation of treatment
12. Study of dose-response relationships
13. Use of multiple time points measuring outcomes
14. Consistency of outcome measurement
15. Blinding of outcome assessment
16. Establishment of primary and secondary end points
17. Precision of effect size
18. Management of conflicts of interest
19. Choice of statistical methods for inferential analysis

20. Recording of the flow of animals through the experiment
21. A priori statements of hypothesis
22. Choice of sample size
23. Addressing confounds associated with treatment
24. Characterization of animal properties at baseline
25. Optimization of complex treatment parameters
26. Faithful delivery of intended treatment
27. Degree of characterization and validity of outcome
28. Treatment response along mechanistic pathway
29. Assessment of multiple manifestations of disease phenotype
30. Assessment of outcome at late/relevant time points
31. Addressing treatment interactions with clinically relevant co-morbidities
32. Use of validated assay for molecular pathways assessment
33. Definition of outcome measurement criteria
34. Comparability of control group characteristics to those of previous studies
35. Reporting on breeding scheme
36. Reporting on genetic background
37. Replication in different models of the same disease
38. Replication in different species or strains
39. Replication at different ages
40. Replication at different levels of disease severity
41. Replication using variations in treatment
42. Independent replication
43. Addressing confounds associated with experimental setting
44. Addressing confounds associated with setting
45. Pre-registration of study protocol and analysis procedures
46. Pharmacokinetics to support treatment decisions
47. Definition of treatment
48. Inter-study standardization of end point choice
49. Define programmatic purpose of research
50. Inter-study standardization of experimental design
51. Research within multicentre consortia
52. Critical appraisal of literature or systematic review during design phase
53. (multiple) free text

**Figure 1: search flow chart.**

**Table 1 – extraction results**

| **Recommendation** | **absolute frequency** | **relative frequency** |
|---|---|---|
| Adequate choice of sample size | 41 | 68% |
| Blinding of outcome assessment | 41 | 68% |
| Choice of statistical methods for inferential analysis | 38 | 63% |
| Randomized allocation of animals to treatment | 38 | 63% |
| Concealed allocation of treatment | 31 | 52% |
| Recording of the flow of animals through the experiment | 31 | 52% |
| A priori statements of hypothesis | 30 | 50% |
| Selection of appropriate control groups | 29 | 48% |
| Characterization of animal properties at baseline | 28 | 47% |
| Addressing confounds associated with setting | 23 | 38% |
| Definition of outcome measurement criteria | 23 | 38% |
| Reporting on genetic background | 23 | 38% |
| Matching or balancing sex of animals across groups | 20 | 33% |
| Degree of characterization and validity of outcome | 19 | 32% |
| Consistency of outcome measurement | 18 | 30% |
| Monitoring emergence of confounding characteristics in animals | 18 | 30% |
| Precision of effect size | 18 | 30% |
| Study of dose-response relationships | 18 | 30% |
| Addressing confounds associated with experimental setting | 17 | 28% |
| Establishment of primary and secondary end points | 17 | 28% |
| Reporting on breeding scheme | 16 | 27% |
| Assessment of outcome at late/relevant time points | 15 | 25% |
| Independent replication | 15 | 25% |
| Matching or balancing treatment allocation of animals | 15 | 25% |
| Specification of unit of analysis | 15 | 25% |
| Randomization for analysis | 14 | 23% |
| Replication in different species or strains | 14 | 23% |
| Standardized handling of animals | 14 | 23% |

| | | |
|---|---|---|
| Addressing confounds associated with anaesthesia or analgesia | 13 | 22% |
| Replication in different models of the same disease | 13 | 22% |
| Addressing confounds associated with treatment | 12 | 20% |
| Management of conflicts of interest | 11 | 18% |
| Treatment response along mechanistic pathway | 11 | 18% |
| Inter-study standardization of experimental design | 10 | 17% |
| Assessment of multiple manifestations of disease phenotype | 9 | 15% |
| Use of multiple time points measuring outcomes | 9 | 15% |
| Definition of treatment | 8 | 13% |
| Inter-study standardization of end point choice | 8 | 13% |
| Pharmacokinetics to support treatment decisions | 8 | 13% |
| Randomized distribution of animals in the animal facilities | 8 | 13% |
| Use of validated assay for molecular pathways assessment | 8 | 13% |
| Faithful delivery of intended treatment | 7 | 12% |
| Addressing treatment interactions with clinically relevant co-morbidities | 6 | 10% |
| Any additional elements that do not fit in the list above | 6 | 10% |
| Comparability of control group characteristics to those of previous studies | 6 | 10% |
| Critical appraisal of literature or systematic review during design phrase | 6 | 10% |
| Define programmatic purpose of research | 6 | 10% |
| Replication at different ages | 6 | 10% |
| Replication using variations in treatment | 5 | 8% |
| Optimization of complex treatment parameters | 4 | 7% |
| Replication at different levels of disease severity | 4 | 7% |
| Research within multicentre consortia | 4 | 7% |
| Pre-registration of study protocol and analysis procedures | 3 | 5% |

**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* 2011;10(9):712. doi: 10.1038/nrd3439-c1

2. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 2009;4(11):e7824. doi: 10.1371/journal.pone.0007824
3. Smith AJ, Clutton RE, Lilley E, et al. PREPARE: guidelines for planning animal research and testing. *Lab Anim* 2018;52(2):135-41. doi: 10.1177/0023677217724823
4. du Sert NP, Bamsey I, Bate ST, et al. The Experimental Design Assistant. *Nat Methods* 2017;14(11):1024-25. doi: 10.1038/nmeth.4462
5. Henderson VC, Kimmelman J, Fergusson D, et al. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med* 2013;10(7):e1001489. doi: 10.1371/journal.pmed.1001489
6. Vollert J, Schenker E, Macleod M, et al. Protocol for a systematic review of guidelines for rigour in the design, conduct and analysis of biomedical experiments involving laboratory animals. *BMJ Open Science* 2018;2(1):e000004. doi: 10.1136/bmjos-2018-000004
7. Anders HJ, Vielhauer V. Identifying and validating novel targets with in vivo disease models: guidelines for study design. *Drug Discov Today* 2007;12(11-12):446-51. doi: 10.1016/j.drudis.2007.04.001
8. Auer JA, Goodship A, Arnoczky S, et al. Refining animal models in fracture research: seeking consensus in optimising both animal welfare and scientific validity for appropriate biomedical use. *BMC Musculoskelet Disord* 2007;8:72. doi: 10.1186/1471-2474-8-72
9. Baker D, Amor S. Publication guidelines for refereeing and reporting on animal use in experimental autoimmune encephalomyelitis. *J Neuroimmunol* 2012;242(1-2):78-83. doi: 10.1016/j.jneuroim.2011.11.003
10. Bordage G, Dawson B. Experimental study design and grant writing in eight steps and 28 questions. *Medical Education* 2003;37(4):376-85. doi: doi:10.1046/j.1365-2923.2003.01468.x
11. Chang CF, Cai L, Wang J. Translational intracerebral hemorrhage: a need for transparent descriptions of fresh tissue sampling and preclinical model quality. *Transl Stroke Res* 2015;6(5):384-9. doi: 10.1007/s12975-015-0399-5
12. Curtis MJ, Hancox JC, Farkas A, et al. The Lambeth Conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias. *Pharmacol Ther* 2013;139(2):213-48. doi: 10.1016/j.pharmthera.2013.04.008
13. Daugherty A, Tall AR, Daemen M, et al. Recommendation on Design, Execution, and Reporting of Animal Atherosclerosis Studies: A Scientific Statement From the American Heart Association. *Circ Res* 2017;121(6):e53-e79. doi: 10.1161/RES.0000000000000169
14. de Caestecker M, Humphreys BD, Liu KD, et al. Bridging Translation by Improving Preclinical Study Design in AKI. *J Am Soc Nephrol* 2015;26(12):2905-16. doi: 10.1681/ASN.2015070832
15. Festing MF. Design and statistical methods in studies using animal models of development. *ILAR J* 2006;47(1):5-14.
16. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* 2002;43(4):244-58.
17. Garcia-Bonilla L, Rosell A, Torregrosa G, et al. Recommendations guide for experimental animal models in stroke research. *Neurologia* 2011;26(2):105-10. doi: 10.1016/j.nrl.2010.09.001
18. Green SB. Can animal data translate to innovations necessary for a new era of patient-centred and individualised healthcare? Bias in preclinical animal research. *BMC Med Ethics* 2015;16:53. doi: 10.1186/s12910-015-0043-7
19. Grundy D. Principles and standards for reporting animal experiments in The Journal of Physiology and Experimental Physiology. *Exp Physiol* 2015;100(7):755-8. doi: 10.1113/EP085299
20. Gulinello M, Mitchell HA, Chang Q, et al. Rigor and reproducibility in rodent behavioral research. *Neurobiol Learn Mem* 2018 doi: 10.1016/j.nlm.2018.01.001

21. Hawkins D, Gallacher E, Gammell M. Statistical power, effect size and animal welfare: recommendations for good practice. *Animal Welfare* 2013;22(3):339-44. doi: 10.7120/09627286.22.3.339
22. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an overview of systematic reviews. *PLoS One* 2014;9(6):e98856. doi: 10.1371/journal.pone.0098856
23. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *Altern Lab Anim* 2010;38(2):167-82.
24. Hooijmans CR, Rovers MM, de Vries RB, et al. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol* 2014;14:43. doi: 10.1186/1471-2288-14-43
25. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. *Nat Rev Neurol* 2014;10(1):37-43. doi: 10.1038/nrneurol.2013.232
26. Hsu CY. Criteria for valid preclinical trials using animal stroke models. *Stroke* 1993;24(5):633-6.
27. Jones JB. Research Fundamentals: Statistical Considerations in Research Design: A Simple Person's Approach. *Academic Emergency Medicine* 2000;7(2):194-99. doi: doi:10.1111/j.1553-2712.2000.tb00529.x
28. Katz DM, Berger-Sweeney JE, Eubanks JH, et al. Preclinical research in Rett syndrome: setting the foundation for translational success. *Dis Model Mech* 2012;5(6):733-45. doi: 10.1242/dmm.011007
29. Kimmelman J, Henderson V. Assessing risk/benefit for trials using preclinical evidence: a proposal. *J Med Ethics* 2016;42(1):50-3. doi: 10.1136/medethics-2015-102882
30. Knopp KL, Stenfors C, Baastrup C, et al. Experimental design and reporting standards for improving the internal validity of pre-clinical studies in the field of pain: Consensus of the IMI-Europain consortium. *Scand J Pain* 2015;7(1):58-70. doi: 10.1016/j.sjpain.2015.01.006
31. Krauth D, Woodruff TJ, Bero L. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environ Health Perspect* 2013;121(9):985-92. doi: 10.1289/ehp.1206389
32. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 2012;490(7419):187-91. doi: 10.1038/nature11556
33. Lara-Pezzi E, Menasche P, Trouvin JH, et al. Guidelines for translational research in heart failure. *J Cardiovasc Transl Res* 2015;8(1):3-22. doi: 10.1007/s12265-015-9606-8
34. Lecour S, Botker HE, Condorelli G, et al. ESC working group cellular biology of the heart: position paper: improving the preclinical assessment of novel cardioprotective therapies. *Cardiovasc Res* 2014;104(3):399-411. doi: 10.1093/cvr/cvu225
35. Liu S, Zhen G, Meloni BP, et al. Rodent Stroke Model Guidelines for Preclinical Stroke Trials (1st Edition). *J Exp Stroke Transl Med* 2009;2(2):2-27.
36. Llovera G, Liesz A. The next step in translational research: lessons learned from the first preclinical randomized controlled trial. *J Neurochem* 2016;139 Suppl 2:271-79. doi: 10.1111/jnc.13516
37. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in ALS/MND: A consensus meeting. *Amyotroph Lateral Scler* 2010;11(1-2):38-45. doi: 10.3109/17482960903545334
38. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for the preclinical in vivo evaluation of pharmacological active drugs for ALS/MND: report on the 142nd ENMC international workshop. *Amyotroph Lateral Scler* 2007;8(4):217-23. doi: 10.1080/17482960701292837
39. Macleod MR, Fisher M, O'Collins V, et al. Reprint: Good laboratory practice: preventing introduction of bias at the bench. *Int J Stroke* 2009;4(1):3-5. doi: 10.1111/j.1747-4949.2009.00241.x
40. Martic-Kehl MI, Wernery J, Folkers G, et al. Quality of Animal Experiments in Anti-Angiogenic Cancer Drug Development--A Systematic Review. *PLoS One* 2015;10(9):e0137235. doi: 10.1371/journal.pone.0137235

41. Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic: best practices. *Mov Disord* 2014;29(11):1375-90. doi: 10.1002/mds.26006
42. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol* 2013;11(2):e1001481. doi: 10.1371/journal.pbio.1001481
43. Omary MB, Cohen DE, El-Omar EM, et al. Not All Mice Are the Same: Standardization of Animal Research Data Presentation. *Cell Mol Gastroenterol Hepatol* 2016;2(4):391-93. doi: 10.1016/j.jcmgh.2016.04.001
44. Osborne N, Avey MT, Anestidou L, et al. Improving animal research reporting standards: HARRP, the first step of a unified approach by ICLAS to improve animal research reporting standards worldwide. *EMBO Rep* 2018;19(5) doi: 10.15252/embr.201846069
45. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;507(7493):423-5. doi: 10.1038/507423a
46. Pitkanen A, Nehlig A, Brooks-Kayal AR, et al. Issues related to development of antiepileptogenic therapies. *Epilepsia* 2013;54 Suppl 4:35-43. doi: 10.1111/epi.12297
47. Raimondo JV, Heinemann U, de Curtis M, et al. Methodological standards for in vitro models of epilepsy and epileptic seizures. A TASK1-WG4 report of the AES/ILAE Translational Task Force of the ILAE. *Epilepsia* 2017;58 Suppl 4:40-52. doi: 10.1111/epi.13901
48. Regenberg A, Mathews DJ, Blass DM, et al. The role of animal models in evaluating reasonable safety and efficacy for human trials of cell-based interventions for neurologic conditions. *J Cereb Blood Flow Metab* 2009;29(1):1-9. doi: 10.1038/jcbfm.2008.98
49. Rice AS, Cimino-Brown D, Eisenach JC, et al. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. *Pain* 2008;139(2):243-7. doi: 10.1016/j.pain.2008.08.017
50. Rostedt Punga A, Kaminski HJ, Richman DP, et al. How clinical trials of myasthenia gravis can inform pre-clinical drug development. *Exp Neurol* 2015;270:78-81. doi: 10.1016/j.expneurol.2014.12.022
51. Sena E, van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 2007;30(9):433-9. doi: 10.1016/j.tins.2007.06.009
52. Shineman DW, Basi GS, Bizon JL, et al. Accelerating drug discovery for Alzheimer's disease: best practices for preclinical animal studies. *Alzheimers Res Ther* 2011;3(5):28. doi: 10.1186/alzrt90
53. Singh VP, Pratap K, Sinha J, et al. Critical evaluation of challenges and future use of animals in experimentation for biomedical research. *Int J Immunopathol Pharmacol* 2016;29(4):551-61. doi: 10.1177/0394632016671728
54. Sjoberg EA. Logical fallacies in animal model research. *Behav Brain Funct* 2017;13(1):3. doi: 10.1186/s12993-017-0121-8
55. Smith MM, Clarke EC, Little CB. Considerations for the design and execution of protocols for animal research and treatment to improve reproducibility and standardization: "DEPART well-prepared and ARRIVE safely". *Osteoarthritis Cartilage* 2017;25(3):354-63. doi: 10.1016/j.joca.2016.10.016
56. Snyder HM, Shineman DW, Friedman LG, et al. Guidelines to improve animal study design and reproducibility for Alzheimer's disease and related dementias: For funders and researchers. *Alzheimers Dement* 2016;12(11):1177-85. doi: 10.1016/j.jalz.2016.07.001
57. Steward O, Balice-Gordon R. Rigor or mortis: best practices for preclinical research in neuroscience. *Neuron* 2014;84(3):572-81. doi: 10.1016/j.neuron.2014.10.042
58. Stone HB, Bernhard EJ, Coleman CN, et al. Preclinical Data on Efficacy of 10 Drug-Radiation Combinations: Evaluations, Concerns, and Recommendations. *Transl Oncol* 2016;9(1):46-56. doi: 10.1016/j.tranon.2016.01.002
59. Stroke Therapy Academic Industry R. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999;30(12):2752-8.

60. Tannenbaum C, Day D, Matera A. Age and sex in drug development and testing for adults. *Pharmacol Res* 2017;121:83-93. doi: 10.1016/j.phrs.2017.04.027
61. Tuzun E, Berrih-Aknin S, Brenner T, et al. Guidelines for standard preclinical experiments in the mouse model of myasthenia gravis induced by acetylcholine receptor immunization. *Exp Neurol* 2015;270:11-7. doi: 10.1016/j.expneurol.2015.02.009
62. Verhagen H, Aruoma OI, van Delft JHM, et al. The 10 basic requirements for a scientific paper reporting antioxidant, antimutagenic or anticarcinogenic potential of test substances in in vitro experiments and animal studies in vivo. *Food and Chemical Toxicology* 2003;41(5 ER - ):603Б–⌈10. doi: 10.1016/s0278-6915(03)00025-5
63. Webster JD, Dennis MM, Dervisis N, et al. Recommended guidelines for the conduct and evaluation of prognostic studies in veterinary oncology. *Vet Pathol* 2011;48(1):7-18. doi: 10.1177/0300985810377187
64. Willmann R, De Luca A, Benatar M, et al. Enhancing translation: guidelines for standard pre-clinical experiments in mdx mice. *Neuromuscul Disord* 2012;22(1):43-9. doi: 10.1016/j.nmd.2011.04.012
65. Willmann R, Luca A, Nagaraju K, et al. Best Practices and Standard Protocols as a Tool to Enhance Translation for Neuromuscular Disorders. *J Neuromuscul Dis* 2015;2(2):113-17. doi: 10.3233/JND-140067
66. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010;8(6):e1000412. doi: 10.1371/journal.pbio.1000412
67. Rennie D. CONSORT revised--improving the reporting of randomized trials. *JAMA* 2001;285(15):2006-7. [published Online First: 2001/04/20]
68. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 2008;39(10):2824-9. doi: 10.1161/STROKEAHA.108.515957 [published Online First: 2008/07/19]
69. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke* 2008;39(3):929-34. doi: 10.1161/STROKEAHA.107.498725
70. Leung V, Rousseau-Blass F, Beauchamp G, et al. ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One* 2018;13(5):e0197882. doi: 10.1371/journal.pone.0197882 [published Online First: 2018/05/26]
71. Avey MT, Moher D, Sullivan KJ, et al. The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. *PLoS One* 2016;11(11):e0166733. doi: 10.1371/journal.pone.0166733 [published Online First: 2016/11/18]
72. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol* 2014;12(1):e1001756. doi: 10.1371/journal.pbio.1001756 [published Online First: 2014/01/11]
73. Hair K, Macleod MR, Sena ES, et al. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev* 2019;4:12. doi: 10.1186/s41073-019-0069-3
74. Bello S, Krogsboll LT, Gruber J, et al. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *J Clin Epidemiol* 2014;67(9):973-83. doi: 10.1016/j.jclinepi.2014.04.008

# A systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals.

Authors: Jan Vollert (1), Esther Schenker (2), Malcolm Macleod (3), Anton Bespalov (4,5), Hanno Würbel (6), Martin C Michel (4,7), Ulrich Dirnagl (8), Heidrun Potschka (9), Ann-Marie Waldron (9), Kimberley E Wever (10), Thomas Steckler (11), Tom Van de Casteele (11), Bruce Altevogt (12), Annesha Sil (13) and Andrew SC Rice (1) on behalf of the EQIPD consortium

1 Pain Research, Department of Surgery and Cancer, Imperial College London, London, UK
2 Institut de Recherches Servier, Croissy-sur- Seine, France
3 Centre for Clinical Brain Sciences, University of Edinburgh, UK
4 Partnership for Assessment and Accreditation of Scientific Practice, Heidelberg, Germany
5 Valdman Institute of Pharmacology, Pavlov Medical University, St. Petersburg, Russia
6 Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland
7 Universitätsmedizin Mainz, Johannes-Gutenberg-Universität Mainz, Mainz, Germany
8 Department of Experimental Neurology, Charité Universitätsmedizin Berlin, Germany
9 Institute of Pharmacology, Toxicology, and Pharmacy, Ludwig-Maximilians- University, Munich, Germany
10 Systematic Review Centre for Laboratory Animal Experimentation, Department for Health Evidence, Nijmegen Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands
11 Janssen Pharmaceutica NV, Beerse, Belgium
12 Pfizer Inc.
13 Institute of Medical Sciences, University of Aberdeen, UK

**Abstract**

Over the last two decades, awareness of the negative repercussions of flaws in the planning, conduct and reporting of preclinical research involving experimental animals has been growing. Several initiatives have set out to increase transparency and internal validity of preclinical studies, mostly publishing expert consensus and experience. While many of the points raised in these various guidelines are identical or similar, they differ in detail and rigour. Most of them focus on reporting, only few of them cover the planning and conduct of studies. The aim of this systematic review is to identify existing experimental design, conduct, analysis and reporting guidelines relating to preclinical animal research. A systematic search in Pubmed, EMBASE and Web of Science retrieved 13,863 unique results. After screening these on title and abstract, 613 papers entered the full-text assessment stage, from which 60 papers were retained. From these, we extracted unique 58 recommendations on the planning, conduct and reporting of preclinical animal studies. Sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment were recommended in more than half of the publications. While we consider these recommendations to be valuable, there is a striking lack of experimental evidence on their importance and relative effect on experiments and effect sizes.

**Introduction**

In recent years, there has been growing awareness of the negative repercussions of shortcomings in the planning, conduct and reporting of preclinical animal research [1 2]. Several initiatives involving academic groups, publishers and others have set out to increase the internal validity and reliability of primary research studies and the resulting publications. Additionally, several experts or groups of experts across the biomedical spectrum have published experience and opinion-based guidelines and guidance. While many of the points raised are broadly similar between these various guidelines (probably in part reflecting the observation that many experts in the field are part of more than one initiative), they differ in detail, rigour, and in particular whether they are broadly generalizable or specific to a single field. While all these guidelines cover the reporting of experiments, only a few specifically address rigorous planning and conduct of studies [3 4], which might increase validity from the earliest possible point [5]. Consequently, it is difficult for researchers to choose which guidelines to follow, especially at the stage of planning future studies.

We aimed to identify all existing guidelines and reporting standards relating to experimental design, conduct and analysis of preclinical animal research. We also sought to identify literature describing (either through primary research or systematic review) the prevalence and impact of perceived risks of bias pertaining to the design, conduct and analysis and reporting of preclinical biomedical research. While we focus on internal validity as influenced by experimental design, conduct and analysis we recognise that factors such as animal housing and welfare are highly relevant to the reproducibility and generalizability of experimental findings; however, these factors are not considered in this systematic review.

**Methods**

The protocol for this systematic review has been published in [6]. The following amendments to the systematic review protocol were made: In addition to the systematic literature search, to capture standards set by funders or organisations that are not (or not yet) published, it was planned to conduct a google search for guidelines published on the websites of major funders and professional organisations using the systematic search string below [6]. This search, however, yielded either no returns, or, in the case of the National Institute of Health, identified over 193,000 results, which was an unfeasibly large number to screen. Therefore, for practical reasons this part of the search was excluded from the initial search strategy. Reassessing the goals of this review, we decided to focus on internal validity, in the protocol we used the term “internal validity and reproducibility”.

**Search strategy**

We systematically searched PubMed, Embase via Ovid and Web of Science to identify guidelines published in English language in peer-reviewed journals before January 10th, 2018 (the day the search was conducted), using appropriate terms for each database optimized from the following search string (as can be found in the protocol [6]):

(guideline OR recommendation OR recommendations) AND
("preclinical model" OR "preclinical models" OR "disease model" OR "disease models" OR "animal model" OR "animal models" OR "experimental model" OR "experimental models" OR "preclinical study" OR "preclinical studies" OR "animal study" OR "animal studies" OR "experimental study" OR "experimental studies") [6].

Furthermore, as many of the researchers participating in the EQIPD (European Quality in Preclinical Data) project (http://eqipd.org/) are experts in the field of experimental standardization, they were contacted personally to identify additional relevant publications.

**Inclusion and exclusion criteria**

We included all articles or systematic reviews in English which described or reviewed guidelines making recommendations intended to improve the validity or reliability (or both) of preclinical animal studies through optimising their design, conduct and analysis. Articles that focussed on toxicity studies or veterinary drug testing were not included. Although reporting standards were not the key primary objective of this systematic review these were also included, as they might contain useful relevant information.

**Screening and data management**

We combined the search results from all sources and identified duplicate search returns and the publication of identical guidelines by the same author group in several based on the PubMed ID, DOI, and the title, journal and author list. Unique references were then screened in two phases: 1) screening for eligibility based on title and abstract, followed by 2) screening for definitive inclusion based on full text. Screening was performed using the Systematic Review facility (SyRF) platform (http://syrf.org.uk). Ten reviewers contributed to the screening phase; each citation was presented to two independent reviewers with a real-time computer-generated random selection of the next citation to be reviewed. Citations remained available for screening until two reviewers agreed that is should be included or excluded. If the first two reviewers had disagreed the citation was offered to a third, but reviewers were not aware of previous screening decisions. A citation could not be offered to the same reviewer twice. Reviewers were not blinded to the authors of the presented record. In the first stage, two authors screened the title and abstract of the retrieved records for

eligibility based on predefined inclusion criteria (see below). The title/abstract screening stage aimed to maximise sensitivity rather than specificity – any paper considered to be of any possible interest was included.

Articles included after the title-abstract screening were retrieved as full-texts. Articles for which no full-text version could be obtained were excluded from the review. Full texts were then screened for definite inclusion and data extraction. At both screening stages, disagreements between reviewers were resolved by additional screening of the reference by a third adjudicating reviewer, who was unaware of the individual judgements of the first two reviewers. All data were stored on the SyRF platform.

**Extraction, aggregation and diligence classification**

From the publications identified, we extracted recommendations on the planning, conduct and reporting or preclinical animal studies as follows:
Elements of the included guidelines were identified using an extraction form inspired by the results from Henderson et al [5]. Across guidelines, the elements were ranked based on the number of guidelines in which that element appeared. Extraction was not done in duplicate, but only once. As the extracted results in this case are not quantitative, but qualitative, meta-analysis and risk of bias assessment are not appropriate for this review. Still, we applied a diligence classification of the guidelines based on the following system, improving level of evidence from 1 to 3 and support from a to b:

1 Recommendations of individuals or small groups of individuals based on individual experience only
    a. Published stand-alone
    b. Endorsed or initiated by at least one publisher or scientific society as stated in the publication
2 Recommendations by groups of individuals, through a method which included a Delphi process or other means of structured decision making
    a. Published stand-alone
    b. Endorsed or initiated by at least one publisher or scientific society as stated in the publication
3 Recommendations based on a systematic review
    a. Published stand-alone

b. Endorsed or initiated by at least one publisher or scientific society as stated in the publication

## Results

### Search and study selection

A flow chart of the search results and screening process is in Figure 1. Our systematic search returned 13,863 results, with 3,573 papers from PubMed, 5,924 from Web of Science, and 5,982 from EMBASE. After first screening on title and abstract, 828 records were eligible for the full-text screening stage. After removing duplications (69), non-English resources (48), conference abstracts (25), book chapters (14), and announcements (4), 676 records remained. Of these, 62 publications were retained after full-text screening. We later identified two further duplicate publications of the same guidelines in different journals, giving a final list 60 publications. [5 7-65].

The project members did not identify any additional papers that had not been identified by the systematic search.

### Diligence classification

More than half of the included publications (32) were narrative reviews that fell under the 1a category of our rating system (recommendations of individuals or small groups of individuals based on individual experience only, published standalone) [7 9 10 14 15 18 20 25 27 29 30 33 35 36 39 41-43 45 47-55 57 60 61 65]. An additional 22 publications were consensus papers or proceedings of consensus meetings for journals or scientific or governmental organizations (category 1b) [3 4 8 12 13 17 19 24 26 28 32 34 37 38 44 46 56 59 62-64 66]. None of these reported the use of a Delphi process or systematic review of existing guidelines. The remaining six publications were systematic reviews of the literature (category 3a) [5 11 21 31 40 58].

### Extracting components of published guidance

From the 60 publications finally included, we extracted 58 unique recommendations on the planning, conduct and reporting of preclinical animal studies. The absolute and relative frequency for each of the extracted recommendations is provided in table 1. Sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment were recommended in more than half of the publications. Only a few publications (≤five) mentioned pre-registration of experimental protocols, research conducted in large consortia, replication at different levels of disease or by variation in treatment, and optimization of complex treatment parameters. The extraction form

allowed the reviewers in free text fields to identify and extract additional recommendations not covered in the pre-specified list, but this facility was rarely used, with only "publication of negative results" and "clear specification of exclusion criteria" extracted in this way by more than one reviewer. The full results table of this stage is published as csv file on figshare under the DOI 10.6084/m9.figshare.9815753.

**Discussion**

Based on our systematic literature search and screening using predefined inclusion and exclusion criteria, we identified 60 published guidelines for the planning, conduct or reporting of preclinical animal research. From these publications, we extracted a comprehensive list of 58 experimental rigour recommendations that the authors had proposed as being important to increase the internal validity of animal experiments. Most recommendations were repeated in a relevant proportion of the publications (sample size calculations, adequate statistical methods, concealed and randomized allocation of animals to treatment, blinded outcome assessment and recording of animal flow through the experiment in more than half of the cases), showing that there is at least some consensus for those recommendations. In many cases this may be because authors are on more than one of the expert committees for these guidelines, and many of them build on the same principles and cite the same sources of inspiration (i.e., doing for the field what CONSORT did for clinical trials [66 67]). There are also reasons why the consensus was not universal – many of the publications focus on single aspects (e.g. statistics [21] or sex differences [60]) or specific medical fields or diseases [13 37 38 63]. In addition, the narrative review character of many of the publications may have led to authors focusing on elements they considered more important than others.

Indeed, more than half (32 out of 60) of the publications reviewed here were topical reviews by a small group of authors (usually fewer than five). Another 22 (37%) were proceedings of consensus meetings or consensus papers set in motion by professional scientific or governmental organizations. It is noteworthy that none of these publications provide any rationale or justification for the validity of their recommendations. None used a Delphi process to structure decision making as suggested for clinical guidelines[68] to reduce bias[69], and none reported using a systematic review of existing guidelines to inform themselves about literature. Of course, many of these expert groups will have been informed by pre-existing reviews (the remaining six included here were systematic literature reviews). However, there is a consistent feature across recommendations – that the steps recommended to increase validity are considered to be self-evident, and a basis in experiments and evidence is seldom linked or provided. There are hints that applying these principles does contribute

to internal validity, as it has been shown that the reporting of measures to reduce risks of bias is associated with smaller outcome effect sizes [70], while other studies have not found such [71]. However, it is unclear if these measures taken are the perfect ones to reduce bias, or if they are merely surrogate markers for more awareness and thus more thorough research conduct. We consider this to be problematic for at least two reasons: first, to increase compliance with guidelines it is crucial to keep them as simple and as easy to implement as possible. An endless checklist can easily lead to fatalistic thinking in researchers desperately wanting to publish, and it could be debated whether guidelines are seen by some researchers as hindering their progression rather than being an aide to conducting the best possible science, still, there is a difference between an 'endless' list and a 'minimal set of rules' that guarantees good research reproducibility. Secondly, each procedure that is added to experimental setup can in itself lead to sources of variation, so these should be minimized unless it can be shown that they add value to experiments.

Compliance is a significant problem for guidelines, as recently reported with the widely adopted ARRIVE guidelines of the UK's National Centre for the 3Rs [66 72]. This is not attributed to blind spots in the ARRIVE guidelines. While enforcement by endorsing journals may be important [73 74], a recent randomized blinded controlled study suggests that even an insistence of completing an ARRIVE checklist has little or no impact on reporting quality [75]. We believe that training and availability of tools to improve research quality will facilitate implementation of guidelines over time, as they become more prominent in researchers' mindset.

This systematic review has important limitations. Main limitation is that we used single extraction only, which was due to feasibility, but creates a source of uncertainty that we cannot rule out. We decided so as we think the bias created here is significantly lower than in a quantitative extraction that includes meta-analysis. Protocol wise, we only included publications in English language, reflecting the limited language pool of our team. Our broad search strategy identified more than 13,000 results, but we did not identify reports or systematic reviews of primary research showing the importance of specific recommendations [76], which must reflect a weakness in our search strategy. Additionally, our plan to search the websites of professional organizations and funding bodies failed due to reasons of practicality. Although being aware of single recommendations outside of publication, we did not include those to keep methods reproducible. In addition, we focused the search on "guidelines", instead of a broader focus on adding, e.g., "guidance", "standard" or "policy", as we feared these terms would inflate the search results by magnitude (particularly "standard" is a broadly used word). Hence, we cannot ascertain whether we have included all important sources of literature. As hinted above, the results presented here also only paint an overview of the literature consensus, which should by no means be mistaken for an

absolute ground truth of which steps need to be taken to improve internal validity in animal experiments. Indeed, literature debating the quality of these measures is sparse, and many of them have been borrowed from the clinical trials community or been considered self-evident from the literature. There is an urgent need for experimental testing of the importance of most of these measures, to provide better evidence of their effect.

**Acknowledgment**

We thank Alice Tillema of Radboud University, Nijmegen, The Netherlands, for her help in constructing and optimising the systematic search strings. This work is part of the European Quality In Preclinical Data (EQIPD) consortium. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777364. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The EQIPD WP3 study group members are: Jan Vollert, Esther Schenker, Malcolm Macleod, Judi Clark, Emily Sena, Anton Bespalov, Bruno Boulanger, Gernot Riedel, Bettina Platt, Annesha Sil, Martien J Kas, Hanno Wuerbel, Bernhard Voelkl, Martin C Michel, Mathias Jucker, Bettina M Wegenast-Braun, Ulrich Dirnagl, René Bernard, Esmeralda Heiden, Heidrun Potschka, Maarten Loos, Kimberley E Wever, Merel Ritskes-Hoitinga, Tom Van De Casteele, Thomas Steckler, Pim Drinkenburg, Juan Diego Pita Almenar, David Gallacher, Henk Van Der Linde, Anja Gilis, Greet Teuns, Karsten Wicke, Sabine Grote, Bernd Sommer, Janet Nicholson, Sanna Janhunen, Sami Virtanen, Bruce Altevogt, Kristin Cheng, Sylvie Ramboz, Emer Leahy, Isabel A Lefevre, Fiona Ducrey, Javier Guillen, Patri Vergara, Ann-Marie Waldron, Isabel Seiffert and Andrew SC Rice.

**Box 1 – Extraction form**

1. Matching or balancing treatment allocation of animals
2. Matching or balancing sex of animals across groups
3. Standardized handling of animals
4. Randomized allocation of animals to treatment
5. Randomization for analysis
6. Randomized distribution of animals in the animal facilities
7. Monitoring emergence of confounding characteristics in animals
8. Specification of unit of analysis
9. Addressing confounds associated with anaesthesia or analgesia

10. Selection of appropriate control groups
11. Concealed allocation of treatment
12. Study of dose-response relationships
13. Use of multiple time points measuring outcomes
14. Consistency of outcome measurement
15. Blinding of outcome assessment
16. Establishment of primary and secondary end points
17. Precision of effect size
18. Management of conflicts of interest
19. Choice of statistical methods for inferential analysis
20. Recording of the flow of animals through the experiment
21. A priori statements of hypothesis
22. Choice of sample size
23. Addressing confounds associated with treatment
24. Characterization of animal properties at baseline
25. Optimization of complex treatment parameters
26. Faithful delivery of intended treatment
27. Degree of characterization and validity of outcome
28. Treatment response along mechanistic pathway
29. Assessment of multiple manifestations of disease phenotype
30. Assessment of outcome at late/relevant time points
31. Addressing treatment interactions with clinically relevant co-morbidities
32. Use of validated assay for molecular pathways assessment
33. Definition of outcome measurement criteria
34. Comparability of control group characteristics to those of previous studies
35. Reporting on breeding scheme
36. Reporting on genetic background
37. Replication in different models of the same disease
38. Replication in different species or strains
39. Replication at different ages
40. Replication at different levels of disease severity
41. Replication using variations in treatment
42. Independent replication
43. Addressing confounds associated with experimental setting

44. Addressing confounds associated with setting
45. Pre-registration of study protocol and analysis procedures
46. Pharmacokinetics to support treatment decisions
47. Definition of treatment
48. Inter-study standardization of end point choice
49. Define programmatic purpose of research
50. Inter-study standardization of experimental design
51. Research within multicentre consortia
52. Critical appraisal of literature or systematic review during design phase
53. (multiple) free text

**Figure 1: search flow chart.**

**Table 1 – extraction results**

| **Recommendation** | **absolute frequency** | **relative frequency** |
|---|---|---|
| Adequate choice of sample size | 41 | 68% |
| Blinding of outcome assessment | 41 | 68% |
| Choice of statistical methods for inferential analysis | 38 | 63% |
| Randomized allocation of animals to treatment | 38 | 63% |
| Concealed allocation of treatment | 31 | 52% |
| Recording of the flow of animals through the experiment | 31 | 52% |
| A priori statements of hypothesis | 30 | 50% |
| Selection of appropriate control groups | 29 | 48% |
| Characterization of animal properties at baseline | 28 | 47% |
| Addressing confounds associated with setting | 23 | 38% |
| Definition of outcome measurement criteria | 23 | 38% |
| Reporting on genetic background | 23 | 38% |
| Matching or balancing sex of animals across groups | 20 | 33% |
| Degree of characterization and validity of outcome | 19 | 32% |
| Consistency of outcome measurement | 18 | 30% |
| Monitoring emergence of confounding characteristics in animals | 18 | 30% |
| Precision of effect size | 18 | 30% |
| Study of dose-response relationships | 18 | 30% |
| Addressing confounds associated with experimental setting | 17 | 28% |

| | | |
|---|---|---|
| Establishment of primary and secondary end points | 17 | 28% |
| Reporting on breeding scheme | 16 | 27% |
| Assessment of outcome at late/relevant time points | 15 | 25% |
| Independent replication | 15 | 25% |
| Matching or balancing treatment allocation of animals | 15 | 25% |
| Specification of unit of analysis | 15 | 25% |
| Randomization for analysis | 14 | 23% |
| Replication in different species or strains | 14 | 23% |
| Standardized handling of animals | 14 | 23% |
| Addressing confounds associated with anaesthesia or analgesia | 13 | 22% |
| Replication in different models of the same disease | 13 | 22% |
| Addressing confounds associated with treatment | 12 | 20% |
| Management of conflicts of interest | 11 | 18% |
| Treatment response along mechanistic pathway | 11 | 18% |
| Inter-study standardization of experimental design | 10 | 17% |
| Assessment of multiple manifestations of disease phenotype | 9 | 15% |
| Use of multiple time points measuring outcomes | 9 | 15% |
| Definition of treatment | 8 | 13% |
| Inter-study standardization of end point choice | 8 | 13% |
| Pharmacokinetics to support treatment decisions | 8 | 13% |
| Randomized distribution of animals in the animal facilities | 8 | 13% |
| Use of validated assay for molecular pathways assessment | 8 | 13% |
| Faithful delivery of intended treatment | 7 | 12% |
| Addressing treatment interactions with clinically relevant co-morbidities | 6 | 10% |
| Any additional elements that do not fit in the list above | 6 | 10% |
| Comparability of control group characteristics to those of previous studies | 6 | 10% |
| Critical appraisal of literature or systematic review during design phrase | 6 | 10% |
| Define programmatic purpose of research | 6 | 10% |
| Replication at different ages | 6 | 10% |
| Replication using variations in treatment | 5 | 8% |

| | | |
|---|---|---|
| Optimization of complex treatment parameters | 4 | 7% |
| Replication at different levels of disease severity | 4 | 7% |
| Research within multicentre consortia | 4 | 7% |
| Pre-registration of study protocol and analysis procedures | 3 | 5% |

**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* 2011;10(9):712. doi: 10.1038/nrd3439-c1
2. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 2009;4(11):e7824. doi: 10.1371/journal.pone.0007824
3. Smith AJ, Clutton RE, Lilley E, et al. PREPARE: guidelines for planning animal research and testing. *Lab Anim* 2018;52(2):135-41. doi: 10.1177/0023677217724823
4. du Sert NP, Bamsey I, Bate ST, et al. The Experimental Design Assistant. *Nat Methods* 2017;14(11):1024-25. doi: 10.1038/nmeth.4462
5. Henderson VC, Kimmelman J, Fergusson D, et al. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLoS Med* 2013;10(7):e1001489. doi: 10.1371/journal.pmed.1001489
6. Vollert J, Schenker E, Macleod M, et al. Protocol for a systematic review of guidelines for rigour in the design, conduct and analysis of biomedical experiments involving laboratory animals. *BMJ Open Science* 2018;2(1):e000004. doi: 10.1136/bmjos-2018-000004
7. Anders HJ, Vielhauer V. Identifying and validating novel targets with in vivo disease models: guidelines for study design. *Drug Discov Today* 2007;12(11-12):446-51. doi: 10.1016/j.drudis.2007.04.001
8. Auer JA, Goodship A, Arnoczky S, et al. Refining animal models in fracture research: seeking consensus in optimising both animal welfare and scientific validity for appropriate biomedical use. *BMC Musculoskelet Disord* 2007;8:72. doi: 10.1186/1471-2474-8-72
9. Baker D, Amor S. Publication guidelines for refereeing and reporting on animal use in experimental autoimmune encephalomyelitis. *J Neuroimmunol* 2012;242(1-2):78-83. doi: 10.1016/j.jneuroim.2011.11.003
10. Bordage G, Dawson B. Experimental study design and grant writing in eight steps and 28 questions. *Medical Education* 2003;37(4):376-85. doi: doi:10.1046/j.1365-2923.2003.01468.x
11. Chang CF, Cai L, Wang J. Translational intracerebral hemorrhage: a need for transparent descriptions of fresh tissue sampling and preclinical model quality. *Transl Stroke Res* 2015;6(5):384-9. doi: 10.1007/s12975-015-0399-5
12. Curtis MJ, Hancox JC, Farkas A, et al. The Lambeth Conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias. *Pharmacol Ther* 2013;139(2):213-48. doi: 10.1016/j.pharmthera.2013.04.008
13. Daugherty A, Tall AR, Daemen M, et al. Recommendation on Design, Execution, and Reporting of Animal Atherosclerosis Studies: A Scientific Statement From the American Heart Association. *Circ Res* 2017;121(6):e53-e79. doi: 10.1161/RES.0000000000000169
14. de Caestecker M, Humphreys BD, Liu KD, et al. Bridging Translation by Improving Preclinical Study Design in AKI. *J Am Soc Nephrol* 2015;26(12):2905-16. doi: 10.1681/ASN.2015070832
15. Festing MF. Design and statistical methods in studies using animal models of development. *ILAR J* 2006;47(1):5-14.

16. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J* 2002;43(4):244-58.
17. Garcia-Bonilla L, Rosell A, Torregrosa G, et al. Recommendations guide for experimental animal models in stroke research. *Neurologia* 2011;26(2):105-10. doi: 10.1016/j.nrl.2010.09.001
18. Green SB. Can animal data translate to innovations necessary for a new era of patient-centred and individualised healthcare? Bias in preclinical animal research. *BMC Med Ethics* 2015;16:53. doi: 10.1186/s12910-015-0043-7
19. Grundy D. Principles and standards for reporting animal experiments in The Journal of Physiology and Experimental Physiology. *Exp Physiol* 2015;100(7):755-8. doi: 10.1113/EP085299
20. Gulinello M, Mitchell HA, Chang Q, et al. Rigor and reproducibility in rodent behavioral research. *Neurobiol Learn Mem* 2018 doi: 10.1016/j.nlm.2018.01.001
21. Hawkins D, Gallacher E, Gammell M. Statistical power, effect size and animal welfare: recommendations for good practice. *Animal Welfare* 2013;22(3):339-44. doi: 10.7120/09627286.22.3.339
22. Hirst JA, Howick J, Aronson JK, et al. The need for randomization in animal trials: an overview of systematic reviews. *PLoS One* 2014;9(6):e98856. doi: 10.1371/journal.pone.0098856
23. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *Altern Lab Anim* 2010;38(2):167-82.
24. Hooijmans CR, Rovers MM, de Vries RB, et al. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol* 2014;14:43. doi: 10.1186/1471-2288-14-43
25. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. *Nat Rev Neurol* 2014;10(1):37-43. doi: 10.1038/nrneurol.2013.232
26. Hsu CY. Criteria for valid preclinical trials using animal stroke models. *Stroke* 1993;24(5):633-6.
27. Jones JB. Research Fundamentals: Statistical Considerations in Research Design: A Simple Person's Approach. *Academic Emergency Medicine* 2000;7(2):194-99. doi: doi:10.1111/j.1553-2712.2000.tb00529.x
28. Katz DM, Berger-Sweeney JE, Eubanks JH, et al. Preclinical research in Rett syndrome: setting the foundation for translational success. *Dis Model Mech* 2012;5(6):733-45. doi: 10.1242/dmm.011007
29. Kimmelman J, Henderson V. Assessing risk/benefit for trials using preclinical evidence: a proposal. *J Med Ethics* 2016;42(1):50-3. doi: 10.1136/medethics-2015-102882
30. Knopp KL, Stenfors C, Baastrup C, et al. Experimental design and reporting standards for improving the internal validity of pre-clinical studies in the field of pain: Consensus of the IMI-Europain consortium. *Scand J Pain* 2015;7(1):58-70. doi: 10.1016/j.sjpain.2015.01.006
31. Krauth D, Woodruff TJ, Bero L. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environ Health Perspect* 2013;121(9):985-92. doi: 10.1289/ehp.1206389
32. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 2012;490(7419):187-91. doi: 10.1038/nature11556
33. Lara-Pezzi E, Menasche P, Trouvin JH, et al. Guidelines for translational research in heart failure. *J Cardiovasc Transl Res* 2015;8(1):3-22. doi: 10.1007/s12265-015-9606-8
34. Lecour S, Botker HE, Condorelli G, et al. ESC working group cellular biology of the heart: position paper: improving the preclinical assessment of novel cardioprotective therapies. *Cardiovasc Res* 2014;104(3):399-411. doi: 10.1093/cvr/cvu225
35. Liu S, Zhen G, Meloni BP, et al. Rodent Stroke Model Guidelines for Preclinical Stroke Trials (1st Edition). *J Exp Stroke Transl Med* 2009;2(2):2-27.
36. Llovera G, Liesz A. The next step in translational research: lessons learned from the first preclinical randomized controlled trial. *J Neurochem* 2016;139 Suppl 2:271-79. doi: 10.1111/jnc.13516

37. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for preclinical animal research in ALS/MND: A consensus meeting. *Amyotroph Lateral Scler* 2010;11(1-2):38-45. doi: 10.3109/17482960903545334
38. Ludolph AC, Bendotti C, Blaugrund E, et al. Guidelines for the preclinical in vivo evaluation of pharmacological active drugs for ALS/MND: report on the 142nd ENMC international workshop. *Amyotroph Lateral Scler* 2007;8(4):217-23. doi: 10.1080/17482960701292837
39. Macleod MR, Fisher M, O'Collins V, et al. Reprint: Good laboratory practice: preventing introduction of bias at the bench. *Int J Stroke* 2009;4(1):3-5. doi: 10.1111/j.1747-4949.2009.00241.x
40. Martic-Kehl MI, Wernery J, Folkers G, et al. Quality of Animal Experiments in Anti-Angiogenic Cancer Drug Development--A Systematic Review. *PLoS One* 2015;10(9):e0137235. doi: 10.1371/journal.pone.0137235
41. Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic: best practices. *Mov Disord* 2014;29(11):1375-90. doi: 10.1002/mds.26006
42. Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol* 2013;11(2):e1001481. doi: 10.1371/journal.pbio.1001481
43. Omary MB, Cohen DE, El-Omar EM, et al. Not All Mice Are the Same: Standardization of Animal Research Data Presentation. *Cell Mol Gastroenterol Hepatol* 2016;2(4):391-93. doi: 10.1016/j.jcmgh.2016.04.001
44. Osborne N, Avey MT, Anestidou L, et al. Improving animal research reporting standards: HARRP, the first step of a unified approach by ICLAS to improve animal research reporting standards worldwide. *EMBO Rep* 2018;19(5) doi: 10.15252/embr.201846069
45. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;507(7493):423-5. doi: 10.1038/507423a
46. Pitkanen A, Nehlig A, Brooks-Kayal AR, et al. Issues related to development of antiepileptogenic therapies. *Epilepsia* 2013;54 Suppl 4:35-43. doi: 10.1111/epi.12297
47. Raimondo JV, Heinemann U, de Curtis M, et al. Methodological standards for in vitro models of epilepsy and epileptic seizures. A TASK1-WG4 report of the AES/ILAE Translational Task Force of the ILAE. *Epilepsia* 2017;58 Suppl 4:40-52. doi: 10.1111/epi.13901
48. Regenberg A, Mathews DJ, Blass DM, et al. The role of animal models in evaluating reasonable safety and efficacy for human trials of cell-based interventions for neurologic conditions. *J Cereb Blood Flow Metab* 2009;29(1):1-9. doi: 10.1038/jcbfm.2008.98
49. Rice AS, Cimino-Brown D, Eisenach JC, et al. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. *Pain* 2008;139(2):243-7. doi: 10.1016/j.pain.2008.08.017
50. Rostedt Punga A, Kaminski HJ, Richman DP, et al. How clinical trials of myasthenia gravis can inform pre-clinical drug development. *Exp Neurol* 2015;270:78-81. doi: 10.1016/j.expneurol.2014.12.022
51. Sena E, van der Worp HB, Howells D, et al. How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* 2007;30(9):433-9. doi: 10.1016/j.tins.2007.06.009
52. Shineman DW, Basi GS, Bizon JL, et al. Accelerating drug discovery for Alzheimer's disease: best practices for preclinical animal studies. *Alzheimers Res Ther* 2011;3(5):28. doi: 10.1186/alzrt90
53. Singh VP, Pratap K, Sinha J, et al. Critical evaluation of challenges and future use of animals in experimentation for biomedical research. *Int J Immunopathol Pharmacol* 2016;29(4):551-61. doi: 10.1177/0394632016671728
54. Sjoberg EA. Logical fallacies in animal model research. *Behav Brain Funct* 2017;13(1):3. doi: 10.1186/s12993-017-0121-8
55. Smith MM, Clarke EC, Little CB. Considerations for the design and execution of protocols for animal research and treatment to improve reproducibility and standardization: "DEPART

well-prepared and ARRIVE safely". *Osteoarthritis Cartilage* 2017;25(3):354-63. doi: 10.1016/j.joca.2016.10.016
56. Snyder HM, Shineman DW, Friedman LG, et al. Guidelines to improve animal study design and reproducibility for Alzheimer's disease and related dementias: For funders and researchers. *Alzheimers Dement* 2016;12(11):1177-85. doi: 10.1016/j.jalz.2016.07.001
57. Steward O, Balice-Gordon R. Rigor or mortis: best practices for preclinical research in neuroscience. *Neuron* 2014;84(3):572-81. doi: 10.1016/j.neuron.2014.10.042
58. Stone HB, Bernhard EJ, Coleman CN, et al. Preclinical Data on Efficacy of 10 Drug-Radiation Combinations: Evaluations, Concerns, and Recommendations. *Transl Oncol* 2016;9(1):46-56. doi: 10.1016/j.tranon.2016.01.002
59. Stroke Therapy Academic Industry R. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999;30(12):2752-8.
60. Tannenbaum C, Day D, Matera A. Age and sex in drug development and testing for adults. *Pharmacol Res* 2017;121:83-93. doi: 10.1016/j.phrs.2017.04.027
61. Tuzun E, Berrih-Aknin S, Brenner T, et al. Guidelines for standard preclinical experiments in the mouse model of myasthenia gravis induced by acetylcholine receptor immunization. *Exp Neurol* 2015;270:11-7. doi: 10.1016/j.expneurol.2015.02.009
62. Verhagen H, Aruoma OI, van Delft JHM, et al. The 10 basic requirements for a scientific paper reporting antioxidant, antimutagenic or anticarcinogenic potential of test substances in in vitro experiments and animal studies in vivo. *Food and Chemical Toxicology* 2003;41(5 ER - ):603Б–⌈10. doi: 10.1016/s0278-6915(03)00025-5
63. Webster JD, Dennis MM, Dervisis N, et al. Recommended guidelines for the conduct and evaluation of prognostic studies in veterinary oncology. *Vet Pathol* 2011;48(1):7-18. doi: 10.1177/0300985810377187
64. Willmann R, De Luca A, Benatar M, et al. Enhancing translation: guidelines for standard pre-clinical experiments in mdx mice. *Neuromuscul Disord* 2012;22(1):43-9. doi: 10.1016/j.nmd.2011.04.012
65. Willmann R, Luca A, Nagaraju K, et al. Best Practices and Standard Protocols as a Tool to Enhance Translation for Neuromuscular Disorders. *J Neuromuscul Dis* 2015;2(2):113-17. doi: 10.3233/JND-140067
66. Kilkenny C, Browne WJ, Cuthill IC, et al. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010;8(6):e1000412. doi: 10.1371/journal.pbio.1000412
67. Rennie D. CONSORT revised--improving the reporting of randomized trials. *JAMA* 2001;285(15):2006-7. [published Online First: 2001/04/20]
68. Moher D, Schulz KF, Simera I, et al. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7(2):e1000217. doi: 10.1371/journal.pmed.1000217
69. Dalkey N. An experimental study of group opinion: The Delphi method. *Futures* 1969;1(5):408-26. doi: https://doi.org/10.1016/S0016-3287(69)80025-X
70. Macleod MR, van der Worp HB, Sena ES, et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 2008;39(10):2824-9. doi: 10.1161/STROKEAHA.108.515957 [published Online First: 2008/07/19]
71. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke* 2008;39(3):929-34. doi: 10.1161/STROKEAHA.107.498725
72. Leung V, Rousseau-Blass F, Beauchamp G, et al. ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One* 2018;13(5):e0197882. doi: 10.1371/journal.pone.0197882 [published Online First: 2018/05/26]

73. Avey MT, Moher D, Sullivan KJ, et al. The Devil Is in the Details: Incomplete Reporting in Preclinical Animal Research. *PLoS One* 2016;11(11):e0166733. doi: 10.1371/journal.pone.0166733 [published Online First: 2016/11/18]
74. Baker D, Lidster K, Sottomayor A, et al. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol* 2014;12(1):e1001756. doi: 10.1371/journal.pbio.1001756 [published Online First: 2014/01/11]
75. Hair K, Macleod MR, Sena ES, et al. A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev* 2019;4:12. doi: 10.1186/s41073-019-0069-3
76. Bello S, Krogsboll LT, Gruber J, et al. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *J Clin Epidemiol* 2014;67(9):973-83. doi: 10.1016/j.jclinepi.2014.04.008