BMJ Open Science

Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals

Jan Vollert ,[1] Esther Schenker,[2] Malcolm Macleod ,[3] Anton Bespalov,[4,5] Hanno Wuerbel,[6] Martin Michel,[7] Ulrich Dirnagl ,[8] Heidrun Potschka,[9] Ann-Marie Waldron,[9] Kimberley Wever ,[10] Thomas Steckler ,[11] Tom van de Casteele,[11] Bruce Altevogt,[12] Annesha Sil,[13] Andrew S C Rice,[1] The EQIPD WP3 study group members

This article has received OSF badges for Open data, Open materials and Pre-registration.

To cite: Vollert J, Schenker E, Macleod M, *et al*. Systematic review of guidelines for internal validity in the design, conduct and analysis of preclinical biomedical experiments involving laboratory animals. *BMJ Open Science* 2020;**4**:e100046. doi:10.1136/bmjos-2019-100046

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjos-2019-100046)

Received 12 September 2019
Revised 10 December 2019
Accepted 15 January 2020

For numbered affiliations see end of article.

Correspondence to
Dr Jan Vollert;
j.vollert@imperial.ac.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 randomised 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 generalisable 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 generalisability of experimental findings; however, these factors are not considered in this systematic review.

METHODS

The protocol for this systematic review has been published in ref 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'. In the protocol, we mention that the aim of this systematic review is an effort to harmonise guidelines and create a unified framework. This is still under way and will be published separately.

Search strategy

We systematically searched PubMed, Embase via Ovid and Web of Science to identify the guidelines published in English language in peer-reviewed journals before 10 January 2018 (the day the search was conducted), using appropriate terms for each database optimised 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 European Quality in Preclinical Data project (http://eqipd.org/) are experts in the field of experimental standardisation, 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 focused 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 it 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 above). 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 of preclinical animal studies as follows:

Elements of the included guidelines were identified using an extraction form (box 1) 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.

Box 1 Extraction form

1. Matching or balancing treatment allocation of animals.
2. Matching or balancing sex of animals across groups.
3. Standardised handling of animals.
4. Randomised allocation of animals to treatment.
5. Randomisation for analysis.
6. Randomised 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. Characterisation of animal properties at baseline.
25. Optimisation of complex treatment parameters.
26. Faithful delivery of intended treatment.
27. Degree of characterisation 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 comorbidities.
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. Preregistration of study protocol and analysis procedures.
46. Pharmacokinetics to support treatment decisions.
47. Definition of treatment.
48. Interstudy standardisation of end point choice.
49. Define programmatic purpose of research.
50. Interstudy standardisation of experimental design.
51. Research within multicentre consortia.
52. Critical appraisal of literature or systematic review during design phase.
53. (Multiple) free text.

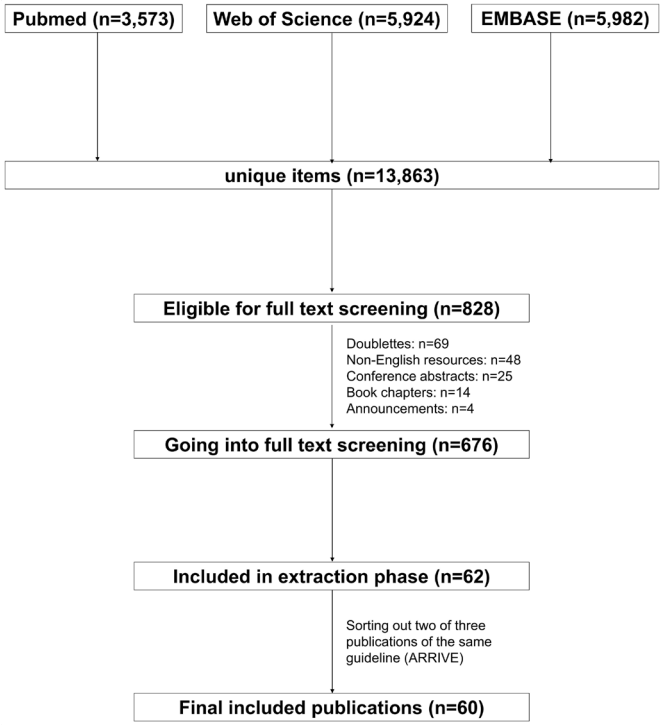

Figure 1 Search flow chart. ARRIVE, Animal Research: Reporting of In Vivo Experiments.

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 found in figure 1. Our systematic search returned 13 863 results, with 3573 papers from PubMed, 5924 from Web of Science and 5982 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 of 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

stand-alone).[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 organisations (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 randomised 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 (≤5) mentioned preregistration of experimental protocols, research conducted in large consortia, replication at different levels of disease or by variation in treatment and optimisation of complex treatment parameters. The extraction form allowed the reviewers in free-text fields to identify and extract additional recommendations not covered in the prespecified 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 randomised 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 (ie, doing for the field what the Consolidated Standards of Reporting Trials did for clinical trials).[66 67] There are also reasons why the consensus was not universal—many of the publications focus on single aspects (eg, statistics[21] or sex differences[60] or

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
Randomised 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
Characterisation 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 characterisation 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
Randomisation for analysis	14	23
Replication in different species or strains	14	23
Standardised 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

Continued

Table 1 Continued

Recommendation	Absolute frequency	Relative frequency (%)
Management of conflicts of interest	11	18
Treatment response along mechanistic pathway	11	18
Interstudy standardisation 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
Interstudy standardisation of end point choice	8	13
Pharmacokinetics to support treatment decisions	8	13
Randomised 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 comorbidities	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
Optimisation of complex treatment parameters	4	7
Replication at different levels of disease severity	4	7
Research within multicentre consortia	4	7
Preregistration of study protocol and analysis procedures	3	5

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 organisations. It is noteworthy that none of these publications provide any rationale or justification for the validity of their recommendations. None used a Delphi process or other means of structured 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. Second, each procedure that is added to experimental set-up can in itself lead to sources of variation, so these should be minimised 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 Animal Research: Reporting of In Vivo Experiments (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 randomised 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. The 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 organisations and funding bodies failed due to reasons of practicality. 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. 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, for example, '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.

Author affiliations
[1]Pain Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK
[2]Institut de Recherches Internationales Servier, Suresnes, Île-de-France, France
[3]Centre for Clinical Brain Sciences, Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK
[4]Partnership for Assessment and Accreditation of Scientific Practice, Heidelberg, Germany
[5]Valdman Institute of Pharmacology, Pavlov First State Medical University of Saint Petersburg, Sankt Petersburg, Russian Federation
[6]Division of Animal Welfare, Vetsuisse Faculty, VPH Institute, University of Bern, Bern, Switzerland
[7]Universitätsmedizin Mainz, Johannes Gutenberg Universität Mainz, Mainz, Rheinland-Pfalz, Germany
[8]Department of Experimental Neurology, Charité–Universitätsmedizin Berlin, Berlin, Germany
[9]Institute of Pharmacology, Toxicology, and Pharmacy, Ludwig-Maximilians-Universitat Munchen, Munchen, Bayern, Germany
[10]Systematic Review Centre for Laboratory Animal Experimentation, Department for Health Evidence, Nijmegen Institute for Health Sciences, Radboud Universiteit, Nijmegen, Gelderland, Netherlands
[11]Janssen Pharmaceutica, Beerse, Antwerpen, Belgium
[12]Pfizer, New York, New York, USA
[13]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

Correction notice This article has been corrected since it was published Online First. ORCIDs have been added for authors.

Acknowledgements We thank Alice Tillema of Radboud University, Nijmegen, The Netherlands, for her help in constructing and optimising the systematic search strings.

Collaborators 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 S C Rice.

Contributors JV wrote the manuscript, mainly designed and conducted the systematic review and organised the process. ES and ASCR supervised the process and the designing and conduction of the systematic review and helped in writing the manuscript. All other authors helped in designing and conducting the systematic review and corrected the manuscript.

Funding 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 number 777364. This joint undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Competing interests None declared.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on Figshare doi:10.6084/m9.figshare.9815753

Open Practices

Open data The data are available at https://figshare.com/articles/SyR_results_guidelines_preclinical_research_quality_EQIPD_csv/9815753.

Open materials The materials used are widely available.

Pre-registration The systematic review and meta-analysis reported in this article was formally preregistered and the protocol published in BMJ Open Science doi:10.1136/ bmjos-2018-000004.

Open peer review Prepublication and Review History is available online at http://dx.doi.org/10.1136/bmjos-2019-1000046.

ORCID iDs
Jan Vollert http://orcid.org/0000-0003-0733-5201
Malcolm Macleod http://orcid.org/0000-0001-9187-9839
Ulrich Dirnagl http://orcid.org/0000-0003-0755-6119
Kimberley Wever http://orcid.org/0000-0003-3635-3660
Thomas Steckler http://orcid.org/0000-0002-9927-5121

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;10:712.
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:e7824.
3 Smith AJ, Clutton RE, Lilley E, *et al*. Prepare: guidelines for planning animal research and testing. *Lab Anim* 2018;52:135–41.
4 du Sert NP, Bamsey I, Bate ST, *et al*. The experimental design assistant. *Nat Methods* 2017;14:1024–5.

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:e1001489.
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:e000004.
7 Anders H-J, Vielhauer V. Identifying and validating novel targets with in vivo disease models: guidelines for study design. *Drug Discov Today* 2007;12:446–51.
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.
9 Baker D, Amor S. Publication guidelines for refereeing and reporting on animal use in experimental autoimmune encephalomyelitis. *J Neuroimmunol* 2012;242:78–83.
10 Bordage G, Dawson B. Experimental study design and grant writing in eight steps and 28 questions. *Med Educ* 2003;37:376–85.
11 Chang C-F, Cai L, Wang J. Translational intracerebral hemorrhage: a need for transparent descriptions of fresh tissue sampling and preclinical model quality. *Transl Stroke Res* 2015;6:384–9.
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;139:213–48.
13 Daugherty A, Tall AR, Daemen MJAP, *et al*. Recommendation on design, execution, and reporting of animal atherosclerosis studies: a scientific statement from the American heart association. *Circ Res* 2017;121:e53–79.
14 de Caestecker M, Humphreys BD, Liu KD, *et al*. Bridging translation by improving preclinical study design in AKI. *J Am Soc Nephrol* 2015;26:2905–16.
15 Festing MFW. Design and statistical methods in studies using animal models of development. *Ilar J* 2006;47:5–14.
16 Festing MFW, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *Ilar J* 2002;43:244–58.
17 García-Bonilla L, Rosell A, Torregrosa G, *et al*. Recommendations guide for experimental animal models in stroke research. *Neurologia* 2011;26:105–10.
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.
19 Grundy D. Principles and standards for reporting animal experiments in the Journal of physiology and experimental physiology. *Exp Physiol* 2015;100:755–8.
20 Gulinello M, Mitchell HA, Chang Q, *et al*. Rigor and reproducibility in rodent behavioral research. *Neurobiol Learn Mem* 2019;165:106780.
21 Hawkins D, Gallacher E, Gammell M. Statistical power, effect size and animal welfare: recommendations for good practice. *Anim Welf* 2013;22:339–44.
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:e98856.
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;38:167–82.
24 Hooijmans CR, Rovers MM, de Vries RBM, *et al*. SYRCLE's risk of bias tool for animal studies. *BMC Med Res Methodol* 2014;14:43.
25 Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. *Nat Rev Neurol* 2014;10:37–43.
26 Hsu CY. Criteria for valid preclinical trials using animal stroke models. *Stroke* 1993;24:633–6.
27 Jones JB. Research fundamentals: statistical considerations in research design: a simple person's approach. *Acad Emerg Med* 2000;7:194–9.
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;5:733–45.
29 Kimmelman J, Henderson V. Assessing risk/benefit for trials using preclinical evidence: a proposal. *J Med Ethics* 2016;42:50–3.
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;7:58–70.
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;121:985–92.
32 Landis SC, Amara SG, Asadullah K, *et al*. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* 2012;490:187–91.
33 Lara-Pezzi E, Menasché P, Trouvin J-H, *et al*. Guidelines for translational research in heart failure. *J Cardiovasc Transl Res* 2015;8:3–22.
34 Lecour S, Bøtker 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;104:399–411.
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;2:2–27.
36 Llovera G, Liesz A. The next step in translational research: lessons learned from the first preclinical randomized controlled trial. *J Neurochem* 2016;139:271–9.
37 Ludolph AC, Bendotti C, Blaugrund E, *et al*. Guidelines for preclinical animal research in ALS/MND: a consensus meeting. *Amyotroph Lateral Scler* 2010;11:38–45.
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;8:217–23.
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;4:3–5.
40 Martić-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:e0137235.
41 Menalled L, Brunner D. Animal models of Huntington's disease for translation to the clinic: best practices. *Mov Disord* 2014;29:1375–90.
42 Muhlhausler BS, Bloomfield FH, Gillman MW. Whole animal experiments should be more like human randomized controlled trials. *PLoS Biol* 2013;11:e1001481.
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;2:391–3.
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. doi:10.15252/embr.201846069
45 Perrin S. Preclinical research: make mouse studies work. *Nature* 2014;507:423–5.
46 Pitkänen A, Nehlig A, Brooks-Kayal AR, *et al*. Issues related to development of antiepileptogenic therapies. *Epilepsia* 2013;54:35–43.
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;58:40–52.
48 Regenberg A, Mathews DJH, 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;29:1–9.
49 Rice ASC, 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;139:243–7.
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.
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;30:433–9.
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;3:28.
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;29:551–61.
54 Sjoberg EA. Logical fallacies in animal model research. *Behav Brain Funct* 2017;13:3.
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;25:354–63.
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;12:1177–85.

57 Steward O, Balice-Gordon R. Rigor or mortis: best practices for preclinical research in neuroscience. *Neuron* 2014;84:572–81.
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;9:46–56.
59 Stroke Therapy Academic Industry Roundtable (STAIR). Recommendations for standards regarding preclinical neuroprotective and restorative drug development. *Stroke* 1999;30:2752–8.
60 Tannenbaum C, Day D. Age and sex in drug development and testing for adults. *Pharmacol Res* 2017;121:83–93.
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;270:11–17.
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 Chem Toxicol* 2003;41:603–10.
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;48:7–18.
64 Willmann R, De Luca A, Benatar M, *et al*. Enhancing translation: guidelines for standard pre-clinical experiments in mdx mice. *Neuromuscul Disord* 2012;22:43–9.
65 Willmann R, Luca AD, Nagaraju K, *et al*. Best practices and standard protocols as a tool to enhance translation for neuromuscular disorders. *J Neuromuscul Dis* 2015;2:113–7.
66 Kilkenny C, Browne WJ, Cuthill IC, *et al*. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol* 2010;8:e1000412.
67 Rennie D. CONSORT revised--improving the reporting of randomized trials. *JAMA* 2001;285:2006–7.
68 Moher D, Schulz KF, Simera I, *et al*. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7:e1000217.
69 Dalkey N. An experimental study of group opinion: the Delphi method. *Futures* 1969;1:408–26 https://doi.org/
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;39:2824–9.
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;39:929–34.
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:e0197882.
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:e0166733.
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:e1001756.
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.
76 Bello S, Krogsbøll LT, Gruber J, *et al*. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *J Clin Epidemiol* 2014;67:973–83.