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 Dirnagl8,6 Heidrun Potschka,9 Ann-Marie Waldron,9 Kimberley Wever10,11 Thomas Steckler11,11 Tom van de Casteele11, Bruce Altevogt12, Annesha Sil13, Andrew SC Rice,1 The EQIPD WP3 study group members

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 13863 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. 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):

(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’).

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. 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.
5. Randomisation for analysis.
6. Randomised distribution of animals in the animal facilities.
7. Monitoring emergence of confounding characteristics in animals.
9. Addressing confounds associated with anaesthesia or analgesia.
10. Selection of appropriate control groups.
13. Use of multiple time points measuring outcomes.
15. Blinding of outcome assessment.
16. Establishment of primary and secondary end points.
17. Precision of effect size.
18. Management of conflicts of interest.
20. Recording of the flow of animals through the experiment.
21. A priori statements of hypothesis.
22. Choice of sample size.
24. Characterisation of animal properties at baseline.
25. Optimisation of complex treatment parameters.
27. Degree of characterisation and validity of outcome.
28. Treatment response along mechanistic pathway.
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.
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 13863 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.57-65

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). An additional 22 publications were consensus papers or proceedings of consensus meetings for journals or scientific or governmental organisations (category 1B). 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).

**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). There are also reasons why the consensus was not universal—many of the publications focus on single aspects (eg, statistics or sex differences) or...
specific medical fields or diseases). 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 societies 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 to reduce bias, 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, while other studies have not found such. 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. This is not attributed to blind spots in the ARRIVE guidelines. While enforcement by endorsing journals may be important, a recent randomised blinded controlled study suggests that even an insistence of completing an ARRIVE checklist has little or no impact on reporting quality. 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 13000 results, but we did not identify reports or systematic reviews of primary research.

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Absolute frequency</th>
<th>Relative frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management of conflicts of interest</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Treatment response along mechanistic pathway</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>Interstudy standardisation of experimental design</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>Assessment of multiple manifestations of disease phenotype</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Use of multiple time points measuring outcomes</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Definition of treatment</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Interstudy standardisation of endpoint choice</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Pharmacokinetics to support treatment decisions</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Randomised distribution of animals in the animal facilities</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Use of validated assay for molecular pathways assessment</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Faithful delivery of intended treatment</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Addressing treatment interactions with clinically relevant comorbidities</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Any additional elements that do not fit in the list above</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Comparability of control group characteristics to those of previous studies</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Critical appraisal of literature or systematic review during design phase</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Define programmatic purpose of research</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Replication at different ages</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Replication using variations in treatment</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Optimisation of complex treatment parameters</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Replication at different levels of disease severity</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Research within multicentre consortia</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Preregistration of study protocol and analysis procedures</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
showing the importance of specific recommendations, 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

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


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 access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

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/bmjqs-2018-000004.

Open peer review Prepublication and Review History is available online at http://dx.doi.org/10.1136/bmjqs-2019-100046.

ORCID ids

Jan Vollert http://orcid.org/0000-0003-0733-5201
Malcolm Macleod http://orcid.org/0000-0001-9187-9839
Ulrich Dimagl http://orcid.org/0000-0003-0755-6119
Kimberley Wever http://orcid.org/0000-0003-3635-3660
Thomas Stecklier 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.


Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *IAR J* 2002;43:244–58.


Hooijmans CR, Leenaars M, Ritkses-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate them into RCTs, and to make systematic reviews more feasible. *Atem L Ab L Anim* 2010;38:167–82.


75 Hair K, Macleod MR, Sena ES, et al. A randomised controlled trial of an intervention to improve compliance with the ARRIVE guidelines ([tICARUs]). *Res Integr Peer Rev* 2019;4:12.