The EQIPD framework for rigor in the design, conduct, analysis and documentation of animal experiments

The EQIPD framework for rigor in animal experiments aims to unify current recommendations based on evidence behind their rationale and was prospectively tested for feasibility in multicenter animal experiments.

Jan Vollert, Malcolm Macleod, Ulrich Dirnagl, Martien J. Kas, Martin C. Michel, Heidrun Potschka, Gernot Riedel, Kimberley E. Wever, Hanno Würbel, Thomas Steckler, EQIPD Consortium and Andrew S. C. Rice

In the past decade, there has been growing awareness of the negative repercussions of poor standards of design, conduct and reporting of biomedical research, including animal experiments[1,2]. Several initiatives have set the aim of increasing the validity and replicability of scientific findings[3–11]. Although many of the recommendations are similar between various guidelines, they differ in detail, rigor and scope[12,13], and only a few cover rigorous planning and conduct of animal studies[5,14]. Consequently, it is difficult for researchers to decide which guidelines to follow, especially at the stage of planning studies.

This framework, created as part of the Enhancing Quality in Preclinical Data (EQIPD, https://quality-preclinical-data.eu/) project, is based on a systematic review of existing guidelines[12]. We conducted two rounds of Delphi consultations (an anonymized process for structured decision-making among groups[15]) among consortium members to rank recommendations based on their considered importance.

At a consensus meeting, participants agreed on a final list of recommendations that were collated into the five major domains described below, the first three focusing on design of the experiment, the fourth on conduct and the final domain on documentation and reporting. This framework was prospectively tested in ring tests, and an online survey was conducted among the experimenters to assess their rating of the importance of each domain along with subdomain examples. Based on this survey, at a final consensus meeting, the tested framework was finalized (Fig. 1; see Supplementary Methods for a detailed process description).

Domain 1: predefined hypotheses and how to use them

Be clear if an experiment aims to test a specific hypothesis or explores research questions that may generate new hypotheses or both. When planning an experiment, a decision must be made about whether it will formally test a specific statistical null hypothesis or less formally explore research questions that may generate new hypotheses.

'Hypothesis-testing' studies are also referred to as 'confirmatory' or 'knowledge-claiming' studies. For a formal hypothesis test, the sample size needs to be determined based on reasonably firm estimates of effect size and variability using power analysis; if more than one hypothesis will be tested based on the same study population, correction for multiple testing may be necessary.

Unless a specific statistical null hypothesis is formally tested as outlined above, an experiment will be 'exploratory' or 'hypothesis-generating' by default. Of note, any study may involve both hypothesis-testing and exploratory parts. For instance, a study may use a sample size with sufficient statistical power to test a specific hypothesis based on a predetermined primary outcome variable (hypothesis-testing part) and perform exploratory analyses on a range of additional outcome variables or separate subgroups (exploratory part). This approach has been standard practice in clinical research for decades, particularly in pivotal studies for regulatory approval of a new treatment, and may be applied to nonclinical research as well. Therefore, study design should consider the following points:

- Any hypothesis testing experiment will have to specify a testable statistical null hypothesis and a primary outcome variable to formally test the hypothesis based on an appropriate statistical test. The sample size, along with a justification based on the statistical power to detect a biologically relevant effect on the primary outcome variable, and the statistical analysis plan need to be defined before any data are collected.
- Any experiments not meeting the assumptions for formal hypothesis testing are exploratory by default. Therefore, probability statistics (*P* values) should not be used to assess findings, but rather effect sizes should be reported along with a measure of uncertainty (for example, confidence intervals). Keep in mind that performing multiple comparisons (by splitting the study population into subgroups or by analyzing multiple

Predefined hypotheses
Be clear if the experiment is aiming at testing a specific hypothesis or at exporting research questions that may generate new hypotheses.

Planning
Specify and document all methods and analyses, including standard operating procedures (SOPs), before conducting the experiment.

Statistics
Specify and document your statisical analysis plan before conducting the experiment and determine your sample size accordingly.

Randomization and blinding
Use appropriate randomization and blinding procedures and any other adequate measures to avoid risk of bias.

Documentation
Not all risks of bias can be avoided, but most can be uncovered: use comprehensive documentation.

Fig. 1 | The EQIPD framework. The EQIPD framework for rigor in animal experiments collates recommendations for five major domains.

outcome variables) will greatly increase the probability of detecting treatment effects by chance. Therefore, findings of exploratory experiments should be interpreted with caution. They may at best be treated as promising hypotheses that need independent confirmation by a formal hypothesis test.

- If findings are inconsistent with the predictions derived from the hypothesis, it should be stated that findings do not support rejection of the null hypothesis. However, data may always be further investigated from an exploratory perspective. Yet although a hypothesis-testing experiment may turn into an exploratory study, an exploratory experiment must never be reported as a hypothesis-testing study. This is not only a question of transparency, it is also to prevent scientific misconduct (generating hypotheses after the results are known (HARKing[16]) but reporting them as pre-specified hypotheses), which would prevent meaningful assessment of the false discovery rate (the chance that random findings appear significant) and distort evidence synthesis by meta-analysis.

Domain 2: planning of methods and analyses

Specify and document all methods and analyses, including standard operating procedures (SOPs), before conducting the experiment. Defining SOPs can facilitate harmonization of procedures across and within laboratories, thereby reducing uncontrolled inter-study heterogeneity in experimental conduct and improving the replicability of experimental results. It also facilitates replication studies in other labs and helps prevent unintended shifts in methods over time.

- Determine what is already known—consider performing a systematic search or systematic review, and search databases of registered protocols to avoid unnecessary duplication and to inform the study protocol.
- Consider preregistering study protocols in an open access repository (for example, https://www.preclinicaltrials.eu/, https://www.animalstudyregistry.org) before the experiment begins. This will increase transparency, avoid unnecessary duplication of work by others and make it easier to publish results irrespective of outcome. In addition, preregistration reduces opportunities for *P* hacking and HARKing. Results will thus gain credibility. Many public repositories offer delayed publication of registered protocols. If public registration even with delayed publication is not feasible, private, time-stamped protocols can still serve to increase a study's credibility. Increasing numbers of journals offer the 'registered report' format (https://cos.io/rr), in which the study protocol is reviewed and, if it is accepted, publication is guaranteed, irrespective of the results.
- Include meaningful negative and positive control treatments in the experiment: that is, include an experimental condition or group for which no effect on the outcome variable is expected (such as baseline measurements, vehicle treatment or sham treatment) and an experimental condition or group for which an effect on the outcome variable is expected (such as a gold-standard treatment). Include the expected outcomes in these groups in the statement of predictions derived from the scientific hypothesis and the link to the statistical null hypothesis.
- Define criteria for the inclusion and exclusion of subjects into the experiment and experimental groups. Any post hoc definition of such criteria has a high risk of being biased.
- Think about disclosure and a dissemination plan. Keep the original, full dataset and consider publishing raw datasets in a repository that is compliant with FAIR (findability, accessibility, interoperability, reuse) principles (https://www.force11.org/group/fairgroup/fairprinciples), for instance Figshare or Dryad, and/or making them available open access.
- Use calibrated instruments, define SOPs for animal housing and husbandry, sample collection and processing, and data recording, and train all experimenters on their procedures to minimize experimenter effects. If multiple experimenters share procedures, make sure to counterbalance treatment groups across experimenters. Document the sources of reagents.
- The use of electronic lab books can be an efficient first step in standardizing documentation. The EQIPD quality system[17] can help implement additional control measures.

Domain 3: statistics

Specify and document a statistical analysis plan before conducting the experiment and determine the sample size accordingly. *P* values have a role but are often used inappropriately in scientific publications[18]. Effect sizes and their confidence intervals are a more meaningful representation of results in most experiments and should always be reported. Bear in mind that often the data do not meet the assumptions of parametric tests such as *t*-tests and some forms of ANOVAs. Correction for multiplicity will be necessary when performing multiple analyses based on the same biological samples. This applies not only when comparing multiple groups for one parameter but also when multiple parameters are measured based on samples obtained from the same animals. Furthermore, in many exploratory analyses, it is sensible to not use *P* values at all.

- Formulate a detailed statistical analysis plan based on the experimental design, before conducting the study. This should include key assumptions of the statistical approaches, including data distribution and strategies to deal with deviations from normal distribution. The statistical analysis plan should reflect the study design to maximize statistical power. Unless the primary experimenter has advanced training in statistical modeling and data analysis, a statistician or biometrician should be consulted. The EQIPD online training platform provides free webinars on the design and conduct of appropriate statistical analysis (https://quality-preclinical-data.eu/resources/eqipd-e-learning/).
- Define the unit of analysis and, if applicable, distinguish between technical replicates and independent replicates[19]. For example, multiple samples from each animal may be collected (for example, multiple cells in an in vitro experiment or multiple trials of a task in an in vivo experiment); however, such samples are not independent replicates and should therefore be averaged per animal before analysis. Thus, the sample size is determined by the number of independent replicate units of analysis. The unit of analysis is the unit that can be randomly allocated to the treatment groups. Often, this is the individual animal. However, when treatments are applied at cage level in group-housed animals (for example, a type of feed or a compound in the drinking water), cage, rather than animal, is the independent unit of analysis.
- When planning a hypothesis-testing experiment, define a quantifiable outcome measure and specify how the hypothesis will be statistically tested against the null hypothesis. Base the calculation of the required sample size on available data (either own or published data). If such data are not available, a sample size calculation can be based on the minimum effect size that would be considered to be of biological relevance in relation to stated hypothesis. When drop-outs are expected to occur, take this into account in the sample size calculation. Stick to the pre-calculated

sample size. A dynamic change of sample size is inappropriate in almost all experimental designs[20] unless an interim analysis is planned and total sample size and statistical analysis are adjusted accordingly. If the sample size is fixed for practical reasons or is outside the experimenter's control, it is possible to calculate the minimum effect that can be detected with a given sample size and consider if this would be relevant to the question. Refrain from using post hoc power calculations, which are meaningless.

- State all additional secondary hypotheses, and choose adequate statistics and outcome measures to test them.
- Consider using statistical measures of precision rather than probability: for example, present effect sizes along with confidence intervals rather than *P* values. If using *P* values appropriately, choose a method to correct for multiple testing. *P* values derived from exploratory analyses of outcome variables cannot be interpreted as hypothesis-testing.
- All statistical approaches follow certain assumptions: consider whether the data meet them. First, are there reasons to assume that the data and its residuals are not distributed normally? If they are not, consider data transformation—many types of biological data are, for example, distributed log-normally. If no normal distribution can be assumed, or other assumptions (for example, equal variances in each treatment group) are seriously violated, then parametric tests (*t*-test, ANOVA or similar methods) will produce unreliable results, and a non-parametric statistical method should be chosen instead. Consider analysis methods that take, for example, repeated measures and multiple variables (with or without collinearity) into account, where applicable. Using non-parametric tests is always the more conservative approach; they make fewer assumptions and therefore produce fewer false positives.
- Use appropriate granular means of informative data display that do not obscure outliers (for example, show dot plots rather than bar graphs, overlie box plots with violin or dot plots). Report statistical analysis in a comprehensive way, including specific analysis methods, degrees of freedom and test statistic.

Domain 4: randomization and blinding

Use appropriate randomization and blinding procedures and any other adequate measures (such as specification of inclusion and exclusion criteria) to avoid risk of bias. The difficulty of successfully replicating results from other groups, or even results of the same group from a month earlier, may often be the result of a substantial degree of bias, leading to systematic error that was unintentionally and/or unknowingly introduced into the experiment. To optimally assess the causal relationship between a treatment and the outcome, minimize risks of bias as much as possible.

- Randomize the allocation of animals to experimental groups, and counterbalance animal housing and husbandry (for example, cages of animals in animal racks), order of experimental procedures and order of sample processing across treatment groups. When animals are selected from within cohorts (for instance, for sacrifice at a given time or for behavioral measurement), these should also be selected at random. Randomization should be maintained until the end of the study.
- Use validated methods or software to randomize rather than using chance or pseudo-random procedures. Most haphazard methods may seem random but hold a significant chance of introducing bias. Document the method used.
- Conceal the allocation sequence, if feasible. This is the first level of blinding (or masking) and helps to reduce the risk of bias caused by expectations of the experimenters from the start.
- Keep as many people as possible (especially experimenters and animal caregivers) blind to treatment, and the allocation masked at all stages of the experiment: that is, during experimental conduct, outcome assessment and data analysis. Keeping experimenters blind to treatment at all stages will reduce the risk of all kinds of performance and detection biases.
- Independent variables such as age, sex, cage, cage rack position and other variables that may affect outcome variables need to be counterbalanced across treatment groups to avoid confounding of results by such variables. Methods for stratified randomization (that maintain balance of key aspects during randomization) can be helpful. In nearly all experiments, balancing or matching at least age group and sex is important, but other aspects will be specific for your setting. Draw an informed opinion from the literature about which additional factors need to be addressed as potential confounders. Usual suspects include type of anesthesia, treatment, (experimental) setting or comorbidities.

Domain 5: documentation

Not all risks of bias can be avoided, but most can be uncovered: use comprehensive documentation. It will not be possible to foresee everything that may affect the outcome of your experiment, and you may not be able to control each variable[21]. Therefore, it is of highest importance that all potential confounders and risks of bias be documented, as well as any deviations from the planned study protocol, such as unexpected events during the experiment.

- Keep track of the characteristics of the animals at the beginning of the experiment (baseline) and, where meaningful, during the experiment. The list of characteristics will differ depending on the goal of the experiment, although some variables are universal (for example, health status, weight or age and sex). Genetic background and breeding scheme will be of high importance for transgenic animal lines and relevant in other cases. Physiological variables are important to keep track of during the experiments in many cases. Animal housing and environmental conditions (for example, temperature, humidity, handling) need to be documented as well.
- Keep track of the flow of each animal and sample through the experiment(s). In case of exclusions or drop-outs, specify which experimental group these animals belonged to and what pre-specified exclusion criteria applied to each of them. If none of the predefined exclusion criteria apply, document the alternative reason for exclusion or drop-out/loss to follow-up.
- Keep track of accidental unblinding and other unexpected events observed during the experimental conduct, as well as any deviations from the protocol. Unforeseen circumstances may require deviations from the planned protocol. However, if these are clearly documented and communicated, transparency is maintained, enabling others to interpret study results in the light of these deviations. This can also help inform future experimental design.
- Accessibility of the documentation is important—use of electronic lab books or the EQIPD quality system can help in standardizing documentation and guaranteeing its sustainability.
- Comprehensive reporting is key to increase methods reproducibility. This includes granular reporting of many of the aspects from domains 1–4.
- Reporting guidelines provide essential and recommended items to be considered. The EQUATOR network

(https://www.equator-network.org/) provides a comprehensive list; for animal experiments, the ARRIVE guidelines are of most relevance[22].

Closing remarks

There are many reasons for low replicability in animal studies, many of which may fall under the umbrella of 'rigor', as addressed here. Financial restrictions, especially in smaller academic laboratories, may lead to insufficient sample sizes[2]. Working under highly standardized laboratory conditions with very low diversity can contribute to results being specific to a single laboratory context[23]. A replicability study confirmed a bias for 'positive' results: 80% of hypothesis-rejecting but only 40% of hypothesis-confirming studies could be replicated[24].

Although we consider the domains listed above as generically applicable to all preclinical experiments, specific recommendations may be challenging or impractical to implement in specific settings. Increasing costs of sophisticated but more and more common methods (such as single-cell sequencing) put additional pressure on laboratories, especially academic laboratories, to prioritize restricted funding.

We hope that this framework will help researchers to keep an emphasis on value over shininess of their work, and that funders will take practical steps to acknowledge the importance of replicable research. In addition, we encourage researchers to seek wider collaborations: multicentric studies, standard in clinical studies, reduce costs per laboratory through sharing of resources, while increasing replicability as a result of increased external validity[23].

On the practical side, we are aware that some of these suggestions can be difficult to implement in practice. For example, preregistration of exploratory research, however feasible[25], might face strong resistance, whereas it might become standard for confirmatory research. Therefore, rather than as a comprehensive checklist, these recommendations are suggested as examples to illustrate the spirit of the domains, while recognizing that these are neither universally applicable or mandatory for all experimental settings, nor a comprehensive list. Instead, they should be used as food for thought to adapt to any experimental situation. In addition, if any of the recommendations above do not seem appropriate and are not followed in a given instance, transparency should be provided by documenting this, including a rationale. ❐

Jan Vollert[1,2,3,4] ✉, Malcolm Macleod[5], Ulrich Dirnagl[6,7], Martien J. Kas[8], Martin C. Michel[9,10], Heidrun Potschka[11], Gernot Riedel[12], Kimberley E. Wever[13], Hanno Würbel[14], Thomas Steckler[15], EQIPD Consortium[*,**] and Andrew S. C. Rice[1]

*[1]Pain Research, Department of Surgery and Cancer, Imperial College London, London, UK. [2]Division of Neurological Pain Research and Therapy, Department of Neurology, University Hospital of Schleswig-Holstein, Campus Kiel, Kiel, Germany. [3]Department of Anaesthesiology, Intensive Care and Pain Medicine, University Hospital Muenster, Muenster, Germany. [4]Neurophysiology, Mannheim Center of Translational Neuroscience (MCTN), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany. [5]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. [6]Department of Experimental Neurology, Charité Universitätsmedizin Berlin, Berlin, Germany. [7]QUEST Center for Responsible Research, Berlin Institute of Health, Berlin, Germany. [8]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands. [9]Partnership for Assessment and Accreditation of Scientific Practice, Heidelberg, Germany. [10]Universitätsmedizin Mainz, Johannes-Gutenberg-Universität Mainz, Mainz, Germany. [11]Institute of Pharmacology, Toxicology, and Pharmacy, Ludwig-Maximilians-Universität, Munich, Germany. [12]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK. [13]Systematic Review Centre for Laboratory Animal Experimentation, Department for Health Evidence, Nijmegen Institute for Health Sciences, Radboud University Medical Centre, Nijmegen, the Netherlands. [14]Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland. [15]Janssen Pharmaceutica NV, Beerse, Belgium. *A list of authors and their affiliations appears at the end of the paper.*

✉*e-mail: j.vollert@imperial.ac.uk*

Published online: 05 September 2022
https://doi.org/10.1038/s41592-022-01615-y

References

1. Kilkenny, C. et al. *PLoS One* **4**, e7824 (2009).
2. Prinz, F., Schlange, T. & Asadullah, K. *Nat. Rev. Drug Discov.* **10**, 712 (2011).
3. Hirst, J. A. et al. *PLoS One* **9**, e98856 (2014).
4. Sansone, S. A. et al. *Nat. Biotechnol.* **37**, 358–367 (2019).
5. du Sert, N. P. et al. *Nat. Methods* **14**, 1024–1025 (2017).
6. Hooijmans, C. R. et al. *BMC Med. Res. Methodol.* **14**, 43 (2014).
7. Sil, A. et al. *Front. Behav. Neurosci.* **15**, 755812 (2021).
8. International Brain Laboratory. *Neuron* **96**, 1213–1218 (2017).
9. Stone, K. *Ann. Neurol.* **68**, A11–A13 (2010).
10. Simera, I. et al. *BMC Med.* **8**, 24 (2010).
11. Macleod, M. R. et al. *Lancet* **383**, 101–104 (2014).
12. Vollert, J. et al. *BMJ Open Sci.* **4**, e100046 (2020).
13. Henderson, V. C., Kimmelman, J., Fergusson, D., Grimshaw, J. M. & Hackam, D. G. *PLoS Med.* **10**, e1001489 (2013).
14. Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A. & Brattelid, T. *Lab. Anim.* **52**, 135–141 (2018).
15. Murphy, E., Black, N., Lamping, D., Mckee, C. & Sanderson, C. *Health Technol. Assess.* **2**, 1–88 (1998).
16. Kerr, N. L. *Pers. Soc. Psychol. Rev.* **2**, 196–217 (1998).
17. Bespalov, A. et al. *eLife* **10**, e63294 (2021).
18. Amrhein, V., Greenland, S. & McShane, B. *Nature* **567**, 305–307 (2019).
19. Eisner, D. A. J. Gen. *J. Gen. Physiol.* **153**, e202012826 (2021).
20. Motulsky, H. J. Naunyn Schmiedebergs Arch. *Pharmacol.* **387**, 1017–1023 (2014).
21. Sorge, R. E. et al. *Nat. Methods* **11**, 629–632 (2014).
22. Percie du Sert, N. et al. *PLoS Biol.* **18**, e3000410 (2020).
23. Alegre, M. L. *Genome Biol.* **20**, 108 (2019).
24. Errington, T. M. et al. *eLife* **10**, e71601 (2021).
25. Dirnagl, U. *PLoS Biol.* **18**, e3000690 (2020).

Acknowledgements

This work has been conducted as part of the European Quality in Preclinical Data (EQIPD) IMI project, which 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 program and EFPIA. We thank Esther Schenker for her contributions to this work. For field testing and surveying the framework, we thank EQIPD work package 4 members as well as Keith Geoffrey Phillips at Eli Lilly and Tony Blockeel at the University of Bristol as part of our collaboration with IMI PainCare (an Innovative Medicines Initiative 2 Joint undertaking under grant agreement no. 777500). This publication reflects only the authors' view, and the Innovative Medicines Initiative 2 Joint Undertaking is not responsible for any use that may be made of the information it contains.

Author contributions

J.V. is the corresponding author and has coordinated this undertaking, mainly organized the meetings and the Delphi process, written the first draft of this manuscript and coordinated author feedback. All authors have contributed to the Delphi process preceding this manuscript. J.V., M.M., U.D., M.C.M., H.P., G.R., K.E.W., H.W., T.S. and A.S.C.R. have participated in the two consensus meetings during which the text and items for this framework were agreed on. M.M. and T.S. have led the EQIPD consortium, of which this work was a part. A.S.C.R. has chaired the work package focused on this undertaking. M.J.K. has coordinated the feedback from field testing of the framework. All authors have read and agreed on the final version of the manuscript.

Competing interests

J.V. has received consulting fees from Casquar, Embody Orthopaedics and Vertex Pharmaceuticals. M.C.M. has received honorary and/or travel support in relationship to being a consultant and/or lecturer for the following pharmaceutical companies: Apogepha, Astellas, Boehringer Ingelheim, Dr. Willmar Schwabe, GSK and Sanofi-Aventis. H.P. has received funding for consulting, talks and research collaborations from Eisai, Zogenix, Bayer, Elanco, Roche, Exeed Epidarex, Lario, Angelini, Jazz Pharmaceuticals, Galapagos and MSD. A.S.C.R. declares remunerated consultancy work for Imperial College Consultants in last 24 months including for Confo, Vertex, Novartis, CombiGene, Orion and Shanghai SIMR Biotech.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01615-y.

Peer review information *Nature Methods* thanks Julia Menon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

EQIPD Consortium

Jan Vollert[1,2,3,4], Malcolm Macleod[5], Ulrich Dirnagl[6,7], Martien J. Kas[8], Martin C. Michel[9,10], Heidrun Potschka[11], Gernot Riedel[12], Kimberley E. Wever[13], Hanno Würbel[14], Thomas Steckler[15] and Andrew S. C. Rice[1]

A list of members and their affiliations appears in the Supplementary Information.