Planning experiments: updated guidance on experimental design and analysis and their reporting III

Article

Accepted Version

Planning experiments: updated guidance on experimental design and analysis and their reporting

Michael J. Curtis[1], Stephen P.H. Alexander[2], Giuseppe Cirino[3], Christopher H. George[4], David A. Kendall[2], Paul A. Insel[5], Angelo A. Izzo[3], Yong Ji[6], Reynold A. Panettieri[7], Hemal H. Patel[5], Christopher G. Sobey[8], S. Clare Stanford[9], Phil Stanley[10], Barbara Stefanska[11], Gary J. Stephens[12], Mauro M. Teixeira[13], Nathalie Vergnolle[14] & Amrita Ahluwalia[15]

Editorial Office
British Journal of Pharmacology

Correspondence:
A Ahluwalia
British Journal of Pharmacology
The Schild Plot
16 Angel Gate
City Road
London
EC1V 2PT | UK

Tel: +44 (0) 20 7239 0171
Fax: +44 (0) 20 7417 0114
E-mail: info@bps.ac.uk

Author affiliations:
[1]Kings College London, UK
[2] University of Nottingham, UK
[3] University of Naples Federico II, Italy
[4] Swansea University, UK
[5] University of California, San Diego, USA
[6] Nanjing Medical University, Nanjing, China
[7] Rutgers University, New Jersey, USA
[8] La Trobe University, Bundoora, VIC, Australia
[9] University College London, UK
[10] UCB Pharma:, Slough, Berkshire, GB
[11]The University of British Columbia, Vancouver, Canada
[12] University of Reading, UK
[13] Federal University of Minas Gerais, Belo Horizonte, Brazil
[14]INSERM, Toulouse, France
[15] William Harvey Research Institute, Queen Mary University of London, UK

Keywords
None applicable

Abbreviations
ANOVA, analysis of variance;
BJP, British Journal of Pharmacology;
s.e.mean, standard error of the mean

Contributions of authors
The article originated from discussions at the regular meetings of the Senior Editors of BJP during 2019 to 2021.

Acknowledgements We would like to thank Mick Bakhle, Andy Ramage, Jim Docherty and Brian King for their valuable contributions

Abstract

If you plan to publish in *British Journal of Pharmacology* (BJP) it is essential to read this article before undertaking a study. This editorial provides guidance that should be read prior to designing your experiments. We have published previously two guidance documents on experimental design and analysis (M. Curtis et al., 2018; M. J. Curtis et al., 2015). The present update clarifies and simplifies the requirements on design and analysis for *BJP* manuscripts. This editorial also details updated requirements following audit and discussion on best practice by the *BJP* editorial board. Explanations for the requirements are provided in the previous articles. Here, we address new issues that have arisen in the course of handling manuscripts and emphasize three aspects of design that continue to present the greatest challenge to authors: randomization, blinded analysis and balance of group sizes.

Introduction

BJP decided in 2015 to ***mandate*** standards for design and analysis for manuscripts published in BJP, abandoning the previous approach which was to ***advise*** on 'best practice'. This decision to change was driven from our assessments that a ***voluntary*** approach did not bring about effective change. It is good practice to review author guidelines on a regular basis, especially those that affect the reproducibility of findings. Regular review also enables incorporation of new good practice(s), but also reconsideration of our approach, based upon audit of outcome and benefit. As we noted previously, numerous journals, including *Plos One* and *Nature,* have reported that 'author compliance can be an issue' (Anonymous, 2017; Hair et al., 2019).

We have found that, since 2016, the quality of manuscripts published in BJP has improved with respect to design and analysis. The evidence for this is provided later when we share the outcome of audits of compliance with our requrements for design and analysis in papers published in BJP during 2016-2021. By reiterating and simplifying the requirements, we hope it will be easier for authors to ensure compliance, and for peer reviewers to confirm this has has occurred.

We include here a simplified list of requirements along with a discussion of new issues that have arisen.

Simplified list of requirements

The following methodological issues should be addressed and reported in the Methods. Please note that failure to fulfil certain data analysis requirements can be rectified after the study is complete. Other issues, such as blinding and randomization, cannot. Thus, these features must be incorporated into the experimental *design* and reported within the Methods.The list below is relevant for both in vivo and in vitro experiments.

1. Experimental design
    a. We ask for sample size ('power') analysis, with a clearly justified expectation of effect size, to help determine what the required group sizes should be for each experimental approach. The norm for such estimations is to take a conservative position. If *a priori* sample size estimation has not been conducted, an explanation for how sample sizes were chosen must be provided.

b. In a prospective study (e.g., planned to test the activity of a drug), randomisation (e.g., of test substance to subject, tissue or cell) is required, and a statement on how randomisation was achieved should be provided.

c. Blinded data collection (with full justification if this is unavoidably impossible, such as in certain types of behavioural study) and blinded analysis of data should be used.

2. Group sizes.

a. Group size (n) is defined as the number of **independent** experimental 'units' in a group. If technical replicates are generated, authors must declare that these are averaged to generate a single independent value that contributes to the n value used for comparative statistical analysis.

b. Data sets with fewer than n=5 should be described as 'exploratory' or 'preliminary', and the data should not be subjected to comparative statistical analysis. Inclusion of data with group sizes of fewer than n=5 is not permitted unless there is an unavoidable shortage of sample availability (which must be explained). An exception to the n=5 requirement would be the ***preliminary*** data that emanate from large drug screens which may be included (but without comparative statistics). For further elaboration on this issue see section below for how to avoid designing a study that will be rejected without peer review

c. Sample sizes for each dataset presented in the results should be clearly stated in the Methods section and these should match the sample sizes reported in the Results for each outcome measured. If samples are lost, an explanation must be given in the Results along with the reason for the missing samples, such as a technical failure during experimentation.

d. Statistical Plan.When datasets are compared in a study, *a priori* stipulated *post hoc* comparison should be conducted only if necessary conditions are met. For example, for a *post hoc* test multiple comparisons should be run only if the data are normally distributed, and there is no inhomogeneity of variance. In the data and statistical analysis subsection of the Methods, please include a clear statement that *post hoc* tests were run only when the criteria listed above were met. If individual F values (from ANOVA) are considered important enough to be reported, these should be provided in the form of a table in a supplement.

4. Data and Statistical analysis

a. The presentation and processing of a dataset should map to its mathematical distribution (see section that discusses means) .

b. Normalization of individual test group values emanating from *in vitro* studies to the assay-matched control group (or baseline value) can be a useful means of controlling for inter-assay variability (e.g. if there are day-to-day variations in environmental factors that affect the baseline for the assay) provided the study is randomised and blinded. Any normalization of data should be justified. If normalizing values to an assay-matched control generates a control mean value of 1 with no error bar, the data cannot be analysed using tests that rely upon homogeneity of variance between the groups (e.g., ANOVA). In this instance, alternative statistical methods should be used that do not include the control data (e.g. non-parametric analysis) and the Y axis appropriately labelled ('fold assay-matched control'). If groups are not assay-matched or unequal, this type of analysis is not permissible.

c. The presence of variability when conducting experiments with multiple groups can be controlled by using analysis of covariance (ANCOVA) wherever possible, e.g. by including baseline or pre-dose values in *in vivo* studies.

5. Level of probability
    a. P values are commonly used to denote whether two or more means differ. The value, in general, is used as a binary decision making tool. Authors must state the P value they have deemed to constitute the threshold for statistical significance in the data and analysis subsection (the significance level, α). In general a single P value threshold is chosen and if this is stated authors must display only a single P value throughout.
    b. Authors may elect to report the full value of P, but they must explain the meaning fully in the Methods section. (see section on P values).
6. Outliers/exclusion criteria
    a. Outliers should be *included* in data analysis and presentation *unless* they meet predefined and acceptable exclusion criteria. The exclusion criteria must be stated clearly in Methods, as well as the statistical analysis used to test that those criteria are satisfied. Unbiased statistical approaches for identifying outliers (such as Rout's or Grubbs test, or more preferably Tukey's rule) should be used only with large data sets (>12) and only ***once*** on any dataset. The number of exclusions and the reasons should be reported.

How to avoid designing a study that will be rejected without peer review

Of the items listed above, several cannot be resolved after the experiment has been completed. Thus, such items must be incorporated into the experimental design: i.e. *a priori*. The publication of *a priori* design by 'preregistering' the protocol prior to conducting the experiment is an initiative for which support is growing. This is exemplified by advocation for universal adoption in the recent ARRIVE 2.0 guidance for animal experimentation (Percie du Sert, Ahluwalia, et al., 2020; Percie du Sert, Hurst, et al., 2020). Whilst we acknowledge the utility of such preregistration databases we also accept that not all experimental protocols are suitable for preregistration and thus we do not mandate this at present. In previous articles, we provided a flow chart to help authors and referees with the design process and we refer readers to this (M. Curtis et al., 2018). Most papers submitted to BJP that contain un-correctable design violations are now rejected without scientific peer review ('triage reject'). Here are the essentials to avoid such an outcome:

I. Randomise the study.
   If this is not possible (because the study is not a prospective designed study), include an explanation for the lack of randomisation in the Methods.
II. Conduct *a priori* sample size estimation.
   Explain in the Methods if data loss affected the sample size and the reason why replacements were not possible. Provide details of the sample size estimation, clearly indicating the expected effect size and the expected variance. Note that group size should be increased to at least n=5 if authors wish to undertake prospective statistical analysis and their sample size estimate tells them n<5 is 'sufficient'.
III. All studies should be undertaken using a blinded design for intervention and for data analysis. We recognise that the practicalities of conducting particular types of experiment blinded to experimental group may be difficult if the sample source is unavoidably disclosed.

However, blinding the analysis is always achievable. Details of the approach must be given in methodology.

We do not rehearse here our arguments for a minimum n=5, please read Curtis et al, 2018 for a full explanation. But we do note that use of heterologous expression systems to test the activity of large numbers of compounds to identify agents that interact with a particular target is a common screening approach employed in pharmacology. Often these data are n=3 and BJP will accept this for compound screens (but please do not conduct comparative statistics). Once this screen has been conducted, the 'lead' compounds are taken through further assays to confirm activity and characterise the ligand:target interaction. For this second step of analysis, the minimum n=5 should be applied to permit statistical analyis.

We also note that it may be necessary to conduct an experiment where the control group is larger than the test groups (Bate & Karp, 2014); this may arise when there is a limited availability of test material (novel drug, for example) resulting in small test group sizes, and a larger control group gives more statistical power. Care must be taken with such a design to ensure that randomization is properly achieved. This approach should not be done in order to reduce the test group sizes. For most studies in pharmacology group sizes should be equal *by design*.

How to avoid rejection by eliminating uncorrectable errors

Several of BJP's requirements can be incorporated into a revision of a manuscript. If a study is designed adequately, choice of the appropriate statistical test is determined by the design and by the data and can be corrected if deemed suboptimal. A common issue related to this is found in many submissions where non-Gaussian data have been analysed using methods that assume normally distributed data. We recommend that authors check whether their dataset fits a normal distribution, or not, and then with this information apply the correct statistical test. With small groups this is best done by examining the residuals (the difference between the actual and the estimated value) from the analysis. If the data set is non-Gaussian then the use of transformations to generate Gaussian data, discussed previously (M. Curtis et al., 2018), is a technique that, if applied appropriately, can render a data set fit for parametric statistical analysis. There is a simple rule of thumb: if all the control values have been normalized to 1 (or 100%) with no standard error, then a nonparametric statistical test should be applied. BJP recommends that wherever possible raw data should be shared within the manuscript.

Analysis errors accompanied by design errors increase the risk of false conclusions, and if an author is asked to re-analyse his/her data, interpretation of the findings may change. It is best to minimise risk of all types of design and analysis error before starting a study.

Areas of concern that require renewed vigilance

The data below represent a survey of the reporting of compliance with *BJP's* requirements after the publication of the second BJP design and analysis guidance document (M. Curtis et al., 2018). The volumes examined were selected at random and only include evaluation of original research reports. For comparison, two volumes (30 manuscripts in total) from 2014, a year before the publication of the first set of guidance was published (M. J. Curtis et al., 2015) are included. Compliance is expressed as the % of manuscripts in the (anonymised) volume that clearly report adherence to the design and analysis requirement listed at the head of each column. The trajectory

of progress from poor compliance (≤40% of papers) to better (41-80%) and to good (81-100%) is colour coded and illustrates a time-related improvement and an improving trajectory.

	Year	No Papers sampled	Comparative statistics on N=5 or greater	Experiments designed with equal N	Randomisation reported	Blind analysis reported	use of single defined P value	Correct Y-axis label	Correct use of post-hoc test
	2014	15	13	27	0	0	27	14	0
	2014	15	27	20	13	7	7	40	0
	2019	9	78	67	33	11	78	22	60
	2019	8	88	25	38	38	63	88	88
	2020	13	92	54	62	46	100	31	62
	2020	13	100	77	62	31	92	85	54
	2021	14	93	64	64	64	86	93	79
	2021	8	100	75	88	75	100	50	100
	2021	13	100	62	38	15	92	46	62
	2021	6	83	50	50	50	100	33	50
	Progress								
	Poor	Better	Good						

Table: Reporting of compliance with BJP guidelines. Issues were selected at random (M. Curtis) from each year with more selected in 2021 to provide greater scrutiny in the most recent complete year. BJP publishes reviews in addition to original articles and often these reviews are contained within 'Themed Issues'. Where this was the case the issue was excluded and the next available issue with greater than 5 original articles chosen. Data are shown as per cent of articles (papers) that met the desired requirements.

The data above confirm that two of the 'unfixable issues' noted earlier, lack of randomization or blinding, have improved compared with the poor compliance (less than 45% overall in volumes published between 2016 and 2017) reported in our previous survey (M. Curtis et al., 2018). The number of papers published annually by BJP has remained steady: 329, 340, 338 and 330 in consecutive years from 2018 to 2021, indicating that although editors and referees are being more diligent in implementing acceptance criteria, and authors are planning studies with a closer eye on the expectations of the journal, this has not impacted negatively on submission or acceptances.

BJP can also take some comfort from the independent evaluation of 'transparency and reproducibility criteria' by the SciScore tool (sciscore.com/rti). Sciscore is a tool established to assess transparency and reporting standards in research. Using its own criteria for good design and analysis (which overlap to a degree with BJP's), (Menke et al., 2020) it can be seen that after the introduction of BJP's last guidance in 2018 there was a leap in the journal's ratings. The improvements may reflect growing awareness of the issues within the biomedical research community and invigorated efforts by the BJP editorial board in addressing the 'fixable' errors mentioned above. The most recent analysis conducted by SciScore indicates that BJP is one of the most highly scoring journals of all those assessed (https://sciscore.com/rti/).

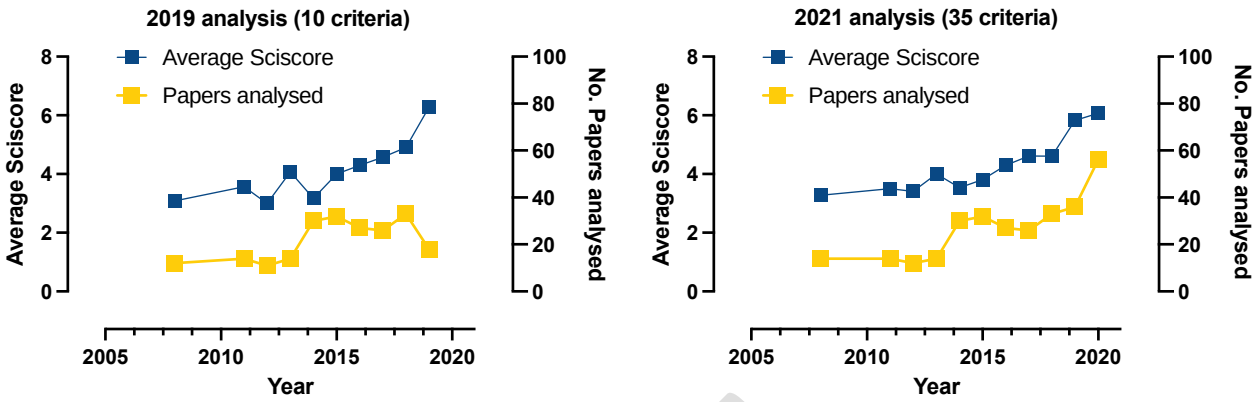

Table 2 SciScore estimate of compliance to experimental design and analysis and transparency best practice of BJP articles (data provided by Dr Anita Bandowski and available at https://sciscore.com/rti/). The analysis up to 2019 included 5 rigour criteria and 5 biological resource criteria. In 2021 these criteria were expanded to include a total of 35 distinct reproducibility criteria.

Practical considerations for sample sizes: biological versus technical replicates

Conventionally, n experimental units give a group size of n independent samples. Unfortunately, there is disagreement across the specialities about what represents an independent sample. In this section, we provide clarity on the current consensus approach taken by BJP and illustrate the agreed approach using scenarios relating to group sizes encountered recently during the review processes in BJP.

If a single biological sample (e.g. tissue or cell culture homogenate) is divided into three aliquots so that three adjacent wells/tubes/lanes measure enzyme activity or receptor response, this methodology generates three values that are almost identical as they are technical replicates; a process often referred to as 'pseudoreplication'. The pseudoreplicates serve also to disguise unwanted sources of variation caused by the experimental procedure. Some investigators infer this means that sources of variation are minimised, which is not the case. Independent repeats unmask technical and biological errors and allow the mean response to emerge. Since technical replicates do not represent independent samples of the population, they must be averaged to provide n=1 for any further comparative statistical analysis.

Organ bath pharmacology is a common method reported in BJP. Indeed, the journal is known for publishing ground breaking and seminal papers incorporating this technique, including the discovery of prostacyclin and the first functional description of an inhibitor of NO synthesis (Rees et al., 1989; Vane, 1964). Blood vessels may be prepared as segments (often rings), with each ring used to generate a concentration response curve. Papers published recently in BJP have arisen where twelve rings from two animals (six rings from each animal) are included in each test group and described as n=6 for each. The current agreed position in BJP is that for blood vessel ring preparations, multiple rings from one animal subject to the same 'test' (e.g. concentration response curve to acetylcholine) usually equate to n=1. For any particular independent n value each time we do the concentration response curve, we should have fresh drug solution, fresh Krebs or Tyrode, and freshly calibrated equipment.

A similar position is taken when characterising ion channel profiles using the same drug preparation in cells/neurons/slices collected from the same tissue from the same animal. BJP considers, as above with blood vessel rings, that repeating the same test on multiple neurons from multiple slices is n=1. For cell culture work, immunoblotting, etc., multiple aliquots of one sample is n=1. For statistical analysis, a minimum of five independent samples are required by BJP (albeit n=5 is an arbitrary number, as explained in Curtis et al., 2018).

A further distinct scenario we wish to highlight occurs when using molecular techniques such as immunoblotting and RT-qPCR. Sample values are typically 'normalized' to levels of a 'housekeeping' protein or gene, then 'normalized' again to the control group. The normalisations are not normally reported transparently. BJP requires blinded, randomized, $n \geq 5$ independent biological replicates, combined with the application of non-parametric or parametric testing, dependent upon how the pair-match is normalized to the control group. Digression from these processes introduces the risk of false positives (Type I error). This is especially the case if the analysis is not blinded.

A reminder about means

For most datasets published in BJP it is appropriate to present the values as means with a standard deviation to show the variability or standard error of the mean to depict the precision of an estimated value. This is the most common practice and is correct for data with a Gaussian distribution. However, in many manuscripts authors present concentration or dose-response curves that utilise a logarithmic scale for the x-axis of concentration or dose. From these curves it is usual to calculate an EC_{50}/IC_{50} or pEC_{50}/pIC_{50} to enable comparison between drugs/treatments. As with most experiments to provide confidence in the finding, the experiment must be repeated at least five times (according to BJP guidance) and so an average is appropriate to present this data. If authors choose to show the mean of the EC_{50}/IC_{50} then, if this value has been derived from a logarithmic scale this should be presented as a geometric mean with 95 CI. If the authors however convert this value to the pEC_{50} and average these then it is the arithmetic mean with standard deviation that should be presented, as per BJP guidelines issued in 2005 (Anonymous, 2005).

A consideration of P values

Stakeholders (in particular, academic colleagues) continue to petition BJP to change its rubric on P values to allow a manuscript to report multiple P values (e.g., $P<0.05$, $P<0.01$, $P<0.001$) for different data sets. It is better understood today that $P<0.001$ does not mean that a difference is bigger or perceived as more 'important' than one that is only $p<0.05$. The P value and the effect size are unrelated. A P-value is calculated to help decide whether the null hypothesis (no effect) can be rejected in favour of the alternative hypothesis (the difference is real)). It is also better understood that if a paper states that for instance $P<0.01$ was taken to indicate statistical significance in binary ('effect' versus 'no effect' tests), then this defines how P values should be presented in the Results section: *denotes statistical significance ($P<0.01$).

However, there are arguments, in some branches of research, where the typically used $P<0.05$ is not 'sufficient' to support persuasive evidence of a difference. In genome-wide association studies (GWAS), for example, $P<0.00000001$ has been taken as a criterion for the threshold of statistical significance. The challenge here is to minimise the risk of a type II error where an extreme number of multiple comparisons are not accommodated by the statistical analysis method commonly employed. Best practice in statistical analysis in this context, although complex and different from

what is typically used to assess effects of drugs, does not advocate 'raising the P bar' to accommodate multiple comparisons (Wang & Xu, 2019). In typical pharmacological experiments one can find ways of accommodating for multiple comparisons such as Dunnett or Tukey post-tests.

It is now widely commented that there is an over-reliance on P values in research and that they can often be mis-interpreted(Lew, 2012). There is a growing interest in using Bayesian statistical techniques where posterior distributions can be used to make probability statements about effects of scientific interest. Whilst we acknowledege this development these methods are not well understood or used by the pharmacological researcher. Most pharmacologists understand that when examining differences between multiple groups treated with different drugs or controls, one compares data with a Gaussian distribution by ANOVA followed by an appropriate *post hoc* test. The probability of a false positive increases exponentially as more post hoc tests are performed. To protect against the increased likelihood of false positives we recommend the use of either Dunnett's or Tukey's post hoc tests. These raise the 't bar' so that a larger t value is required for between-group comparisons to reach statistical significance, based on the number of intended comparisons (often called 'multiplicity'). Dunnett's test can be used when comparing treated groups with the same control group (i.e. the maximum number of tests is the number of groups minus one). Tukey's test should be used in all other post hoc situations. With just three groups, a non-significant F test from the ANOVA suggests that there are no differences amongst the groups and no post hoc tests should be performed. However, with increasing numbers of groups incorporated into a study design the F test becomes increasingly unreliable in its power to detect differences, and thus may not function well as a "gatekeeper" (Bate ST and Clark RA 2014). If the author feels this is the case the author must undertake an alternative form of analysis and justify this fully in the Methods.

Authors may choose to set a single P value and apply this one value throughout the manuscript using symbols such as the asterisk (*) to represent the threshold P value according to the crierion described in the Methods. If authors choose to present the calculated (exact) P value, the authors must explain the meaning of the P value within the "Data and Statistical Analysis" section.

Conclusions

We continue to simplify and respond to changes in best practice in addition to auditing author uptake of key design and analysis principles highlighted in BJP's requirements. Some issues have a greater impact on the reproducibility of data than others. Our primary objective is to ensure that randomization, blinding and adequate group sizes become the norm, in pharmacology discovery and translational research.

Acknowledgements

We would like to thank Anita Bandrowski of Sciscore for providing the data shown within the manuscript.

List of references

Anonymous. (2005). Nomenclature Guidelines for Authors. *British Journal of Pharmacology, 144*(1), 155-157.

Anonymous. (2017). Announcement: Transparency upgrade for Nature journals. *Nature, 543*(7645), 288-288.

Bate ST and Clark RA 2014). The design and statistical analysis of animal experiments. Cambridge University Press, p. 224.

Bate, S., & Karp, N. A. (2014). A Common Control Group - Optimising the Experiment Design to Maximise Sensitivity. *PLOS ONE, 9*(12), e114872.

Curtis, M., Alexander, S., Cirino, G., Docherty, J. R., George, C. H., Giembycz, M. A., … Ahluwalia, A. (2018). Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers. *British Journal of Pharmacology, 175*(7), 987-993.

Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P., Giembycz, M. A., … McGrath, J. C. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *Br J Pharmacol, 172*(14), 3461-3471.

Hair, K., Macleod, M. R., Sena, E. S., Sena, E. S., Hair, K., Macleod, M. R., … on behalf of the, I. C. (2019). A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). *Research Integrity and Peer Review, 4*(1), 12.

Menke, J., Roelandse, M., Ozyurt, B., Martone, M., & Bandrowski, A. (2020). The Rigor and Transparency Index Quality Metric for Assessing Biological and Medical Science Methods. *iScience, 23*(11), 101698.

Percie du Sert, N., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., … Würbel, H. (2020). Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. *PLOS Biology, 18*(7), e3000411.

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., … Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology, 18*(7), e3000410.

Rees, D. D., Palmer, R. M., Hodson, H. F., & Moncada, S. (1989). A specific inhibitor of nitric oxide formation from L-arginine attenuates endothelium-dependent relaxation. *Br.J.Pharmacol., 96*(2), 418-424.

Vane, J. R. (1964). The Use of Isolated Organs for Detecting Active Substances in the Circulating Blood. *Br J Pharmacol Chemother, 23*, 360-373.

Wang, M., & Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity, 123*(3), 287-306.