Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers

Michael J. Curtis, Steve Alexander, Giuseppe Cirino, James R. Docherty, Christopher H. George, Mark A. Giembycz, Daniel Hoyer, Paul A. Insel, Angelo A. Izzo, Yong Ji, David J. MacEwan, Christopher G. Sobey, S. Clare Stanford, Mauro M. Teixeira, Sue Wonnacott & Amrita Ahluwalia

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:
MJC: Kings College London, UK
SA:
GC:
JRD:
CHG: Swansea University, UK
MAG: University of Calgary, Calgary, Alberta, Canada.
DH:
PAI: University of California, San Diego, La Jolla, California, USA
AAI:
DJM:
CGS: La Trobe University, Bundoora, Victoria, Australia.
SCS: University College London, UK
MMT: Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
SW:
AA: 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 2016 and 2017. Michael J Curtis coordinated the writing of the manuscript with contributions and edits from all of the other authors.

Acknowledgements We would like to thank Drs Mick Bakhle and Caroline Wedmore for their valuable contributions

Abstract

This article updates the guidance published in 2015 for authors submitting papers to *British Journal of Pharmacology* (BJP) (Curtis et al., 2015) and is intended to provide the rubric for peer review. Thus, it is directed towards authors, reviewers and editors. Explanations for many of the requirements were outlined previously and are not restated here. The new guidelines are intended to replace those published previously. The guidelines have been simplified for ease of understanding by authors, to make it more straightforward for peer reviewers to check compliance, and to facilitate the curation of the journal's efforts to improve standards.

Introduction

The aim of this update is to:

- share and apply lessons learnt during the 2 years since the implementation of our guidelines
- update aspects of design and analysis where our views have changed/advanced since 2015
- include advice and guidelines on additional areas of design and analysis pertinent to pharmacology research but not discussed in the previous version
- make the journal requirements clearer and easier to curate.

With respect to this final point, the two key changes/additions are a simplified list of requirements for authors, and a flow chart explaining how peer review may be accomplished efficiently. The latter is focused on addressing issues that have been journal requirements since 2015, but which internal audit has revealed are not being routinely addressed by authors and are being missed during the peer review process. This issue of 'non-compliance' is not unique to BJP and is a phenomenon experienced by many other journals. Indeed, *Nature* recently reported that when guidelines are introduced, 'author compliance can be an issue' (anonymous 2017).

Our intention is to improve our guidelines to make it clearer to authors what we require, easier for peer review to manage and easier for the British Pharmacological Society (BPS), which owns the journal, to curate. In doing this, we facilitate one of our key aims which is to support the improved reporting and 'transparency' of experimental work. This article also provides figures to illustrate key aspects of good and inappropriate practice in data acquisition and processing in order to further clarify what is acceptable and unacceptable to BJP.

We focus on issues that, in the experience of the *Senior Editorial* team, continue to be problematic (e.g. normalisation, transformation and inappropriate use of parametric statistics). Additionally, we provide new guidance on certain more nuanced issues, including the handling of outliers in datasets. We note, finally, that this new guidance is entirely focused on design and analysis. Previous attempts in our journal have combined requirements for design and analysis with other equally important, but distinct, issues concerning the use of animals and experimental ethics (e.g. ARRIVE). In hindsight, we feel that this combination has been detrimental and has contributed to a lack of clarity, which may have resulted in the inadequate compliance that we have found from our audits of research papers published in BJP in 2016 and 2017. Guidance on the publication in BJP of ethical animal experimentation and ARRIVE are published elsewhere (McGrath and Lilley 2015), and will be updated in 2018.

In order to achieve our objectives the BJP requires that every paper should contain a design and analysis sub-section within the Methods.

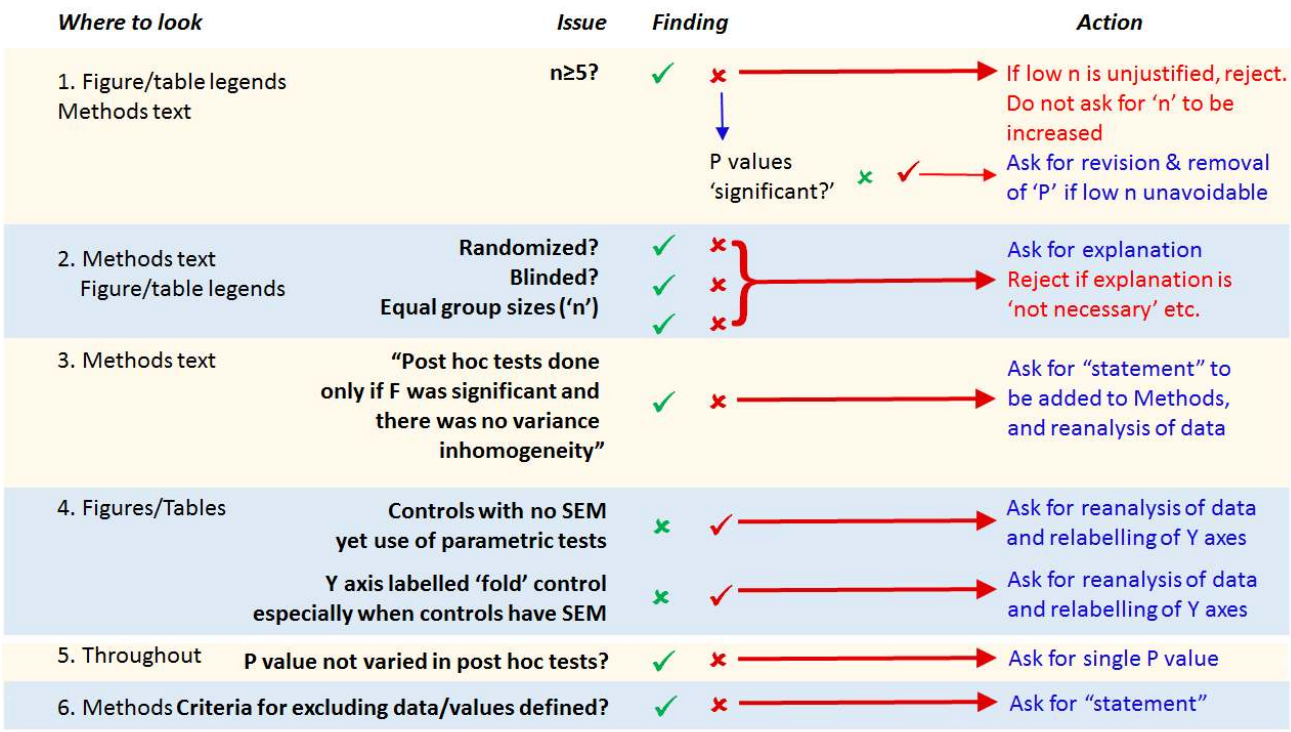
Key points of the updated guidance

1. Group sizes should be sufficient to permit any statistical analysis to be meaningful. At the present time, BJP has set 5 as the minimum "n" required for datasets subjected to statistical analysis. Inclusion of datasets with a smaller group size (without statistical analysis) is permissible but only if carefully justified (e.g. shortage of sample availability). Such data should be labelled as 'exploratory' or 'preliminary', and should constitute only a small proportion of the data in the paper. We strongly advise investigators to undertake power analysis and then increase the calculated minimum group size by at least 50% (explained below). Group size is the number of independent variables, so one sample run 5 times is **n=1, not n=5**. We note that it is common for authors to run 3 samples for instance in quintuplicate, then analyse the data, with statistical analysis, as if it were n=15 rather than n=3. This is not acceptable for publication in BJP.
2. Studies should be designed to generate groups of equal size, using randomisation and blinded analysis where possible (with a credible justification if this is not possible). Even a well-designed study may give rise to unequal group sizes owing to loss of animals/samples etc. due, for instance, to technical failure. If the latter has occurred, this must be explained in the Results. Clear statements on the use of randomisation, blinding and establishment of *a priori* equal group sizes (and whether and how any lost values were replaced) must be provided in Methods.
3. It is important to emphasise that after ANOVA, post hoc tests may be run *only* if F achieves the necessary level of statistical significance (i.e. $P<0.05$) and there is no significant variance inhomogeneity. Adherence must be stated in the Methods ('data were analysed by ANOVA followed by Tukey's test' is not sufficient). If these criteria are not met, a post hoc test should not be run (even if the software permits this, *which it may*).
4. The Methods should explain the reason for any data normalisation (which means the correction of test values to baseline or control group values). Approaches used to reduce unwanted sources of variation, or to transform log-normal data to normal (Gaussian) is encouraged, but the details and reasons must be explained in Methods. Such transformations of the experimental data can affect the appropriateness of the chosen statistical method. For example, normalisation to matched controls will generate a control mean of 1 and no s.e.mean, meaning that parametric tests (ANOVA, etc.) cannot be used (only non-parametric analysis is acceptable). Any dataset where one group has no s.e.mean (common in western blot analysis) must be analysed by non-parametric statistics. Following data transformation, the Y axis is often labelled incorrectly (it should be 'fold matched control values', or 'fold mean of the controls', and not 'fold control').
5. When comparing groups, a level of probability (P) deemed to constitute the threshold for statistical significance (typically in pharmacology this is $P<0.05$) should be defined in Methods, and not varied later in Results (by presentation of multiple levels of significance).
6. Outliers are data values that digress from the central tendency. A range of issues may need to be considered before an investigator can decide how to deal with outliers. It is possible to define an outlier in a control population, but only if the control group has been generated numerous times previously so that a large number of control data are available for inspection. Outliers should therefore be *included* in data analysis and presentation *unless* a predefined and defensible set of exclusion criteria can be generated and applied.

The peer review process

We expect BJP papers to be written in such a way that the basic requirements of design and analysis are described clearly by authors, and can be checked in peer review. To improve this process, we have prepared a summary flow chart of how peer reviewers may quickly triage the key areas and check for compliance with BJP's core requirements. At the same time this flow chart explains to authors what BJP expects from them. The triage scheme (Figure 1) should be used by authors and those involved in peer review to complement the updated guidance.

Figure 1. This flow chart describes how triage of the design and analysis aspects of a paper may be checked by authors and by peer reviewers.



Areas of particular concern that require renewed vigilance

Many of the key issues in the flowchart (above) will be simple to address. Methods must make statements about the items listed above (Figure 1). In studies that do not comply with our requirements, examination of figure legends generally reveals unequal group sizes, statistical analysis often applied to group sizes of n<5 and use of inappropriate application of specific statistical tests. Together, these render a paper fundamentally flawed with respect to BJP guidance. We accept that there are instances during an experiment where a sample may be lost due to a technical issue, or where blinding the data analysis was not possible owing to a very large i.e. obvious, drug effect. In such scenarios authors can and should easily *explain* why their study falls short in these aspects, and we expect peer reviewers to *accept* a good explanation. Indeed, we also encourage the inclusion of pilot data (that may or may not be blinded, randomized or fit for statistical analysis) so long as it is clearly identified as such. In view of this we conducted an audit of

the general compliance with our guidance introduced in 2015. Table 1 illustrates it has not succeeded as we had hoped.

Table 1. Many papers published in British Journal of Pharmacology ***do not*** adhere to all the journal's guidance

	Cite Guidance 2015	all N=5+	equal N	Randomized statement	blind statement	P constant	correct Y axis label	correct post hoc test
	1	0	0	0	0	0	0	1
	1	0	0	0	0	1	1	1
	1	0	0	0	0	1	0	1
	1	1	0	0	0	1	1	1
	1	0	1	0	0	1	1	1
	1	1	0	1	1	1	0	1
	1	1	1	1	1	1	1	1
	1	1	0	1	1	1	1	1
	1	1	0	1	0	1	0	1
	1	1	1	1	0	1	1	1
	1	0	0	0	0	1	1	1
	1	1	0	0	0	1	1	1
	1	1	0	0	1	0	1	1
	1	1	0	0	0	0	1	0
	1	1	1	0	0	0	1	1
	1	1	1	1	0	1	1	1
	1	1	0	1	1	1	0	1
	1	1	0	1	1	1	1	1
	1	1	1	0	1	1	0	0
	1	1	0	1	1	1	0	0
	1	1	0	0	0	1	0	0
	1	1	1	0	0	1	0	0
	1	0	0	1	1	1	0	0
	1	1	0	0	0	1	0	0
	1	1	0	0	0	1	0	0
	1	0	1	0	0	1	0	1
% compliance	**100**	**73**	**31**	**38**	**35**	**85**	**50**	**69**

The data above summarizes papers in three issues of British Journal of Pharmacology published at least 6 months after the publication of the journal's original guidance on design and analysis. The issues were selected at random and are representative and the entire set of original research in each of the three issues sampled was assessed. Papers were judged compliant (1) or not (0). Papers not stating whether the study was blinded or randomized or not were assumed to be non-compliant. 'Correct Y axis label' means that for normalized data, if the control mean is 1 with no SEM the label should be 'fold matched controls', whereas if the control mean has an SEM the Y axis label should be 'fold control mean' (especially relevant to Western blot and qRT-PCR data). 'Correct post hoc test' means that parametric tests were used only when the control mean has a variance, or there is no obvious violation of ANOVA (e.g., post hoc tests done when the controls have sequence of groups' SEMs that are proportional to the mean). Given that all the papers were found to cite the previous guidance document, and in doing so made a declaration of concordance with the guidelines. We would expect the % compliance for each item to be 100%. The fact it is not explains why we have written this article.

Below we discuss specific issues illustrated in Table 1.

Parametric versus non-parametric

One important issue concerns normalisation and the generation of ‘normal’ (Gaussian) data necessary for the application of statistics that depend upon a Gaussian distribution. Figure 2, as an example, shows two different acceptable ways of analysing and presenting data that, for reasons the author should explain, are normalised to control (a common procedure in western blot analysis and electrophysiology studies). It is interesting, and not well recognised, that the data

normalisation approach shown in the right-hand part of the figure is not actually a true normalisation and is, in fact, simply a rescaling of the Y axis of the raw data (see legend). The important difference between the two, however, is that the control on the right-hand side has a standard error whilst the control on the left-hand side does not. Both are acceptable forms of data presentation for BJP, however the statistical tests that should be applied are different.

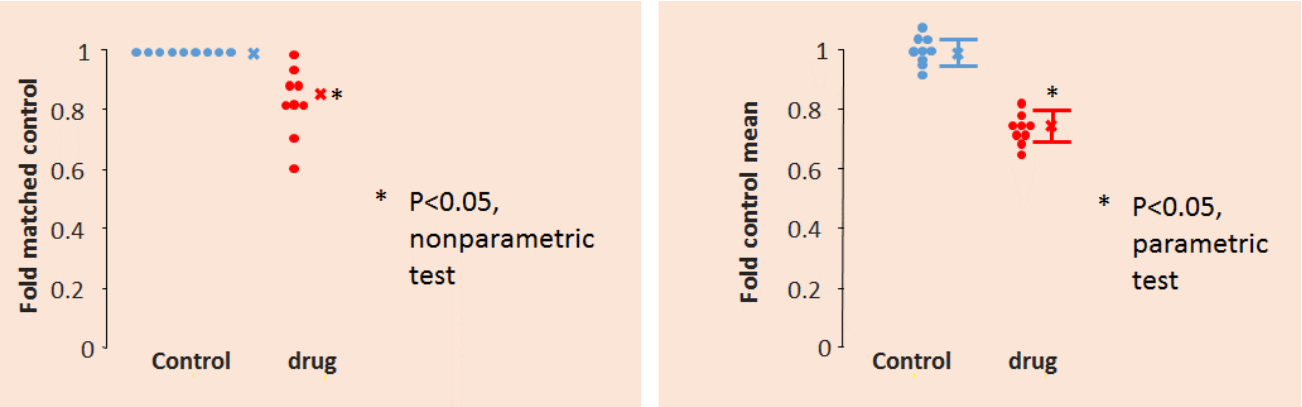


Figure 2. Individual data points (circles), mean values (x) and s.e.mean values are shown. On the left are two datasets derived from an analysis (e.g. western blotting) where each drug experiment has a *matched* control. Each drug value is normalised to each matched control value. This means that the control mean is 1, and there is no variance in the control group (s.e.mean is zero). The correct way to analyse these data is using a non-parametric statistical test, and the correct label for the Y axis is 'fold matched control'. Because analysis is non-parametric, it is misleading to show the parameter s.e.mean. In the figure on the right, each control and each drug value has been 'normalised' to the *mean* value of the control group (mean values shown as x). In other words, each raw value has been divided by the value of the mean of the control values. This generates a dataset that ***can*** be analysed by parametric statistics (provided the variance is similar in the two groups – a t test may falsely identify a nonsignificant difference if the two s.e.mean values differ greatly – see Figure 3) and, if so, it is appropriate to show the s.e.mean (error bars in the figure). However, there is actually no benefit in making this transformation since the ratio between each mean and each s.e.mean is the same as the equivalent ratios for the raw data. In other words, this 'transformation' is identical to a relabelling or rescaling of the Y axis from the absolute raw values to new values for which the control mean value is relabelled as '1'. This has no effect on the ability of the parametric statistical test to detect a significant difference. However, readers have a tendency to make 'eyeball' comparisons between normalised datasets in the same paper or indeed from one paper to the next, and this may lead to false inferences. Thus, use of either the normalisation shown in the left-hand figure, or presentation of the raw data (in units that may be arbitrary if this is the case) is preferred, although we acknowledge that for quantitative PCR the 'transformation' depicted in the right-hand panel represents common practice at the present time. We also remind authors to ensure that whatever normalisation is chosen, the Y axis is labelled correctly (fold control is not correct) and the appropriate type of statistical test is used.

Data transformation is another important issue for transparency. Our audit has revealed two key issues:

- The first is that data that are not normally distributed must not be subjected to parametric statistical analysis (t tests, ANOVA and post hoc tests that account for multiple groups such as Tukey). They should be subjected to non-parametric tests such as the Mann Whitney U test.
- The second key point is that a data transformation can be helpful for analysis and can often convert the data to fit a normal distribution. In this context, in pharmacology, we are most familiar

with log-normal datasets and recommend log transformation when it can be justified. This is much easier to identify than one might imagine: if the s.e.mean increases in size in proportion to the size of the mean, this is typically log normal, and the benefits (and indeed the necessity) of log transformation are shown in Figure 3.

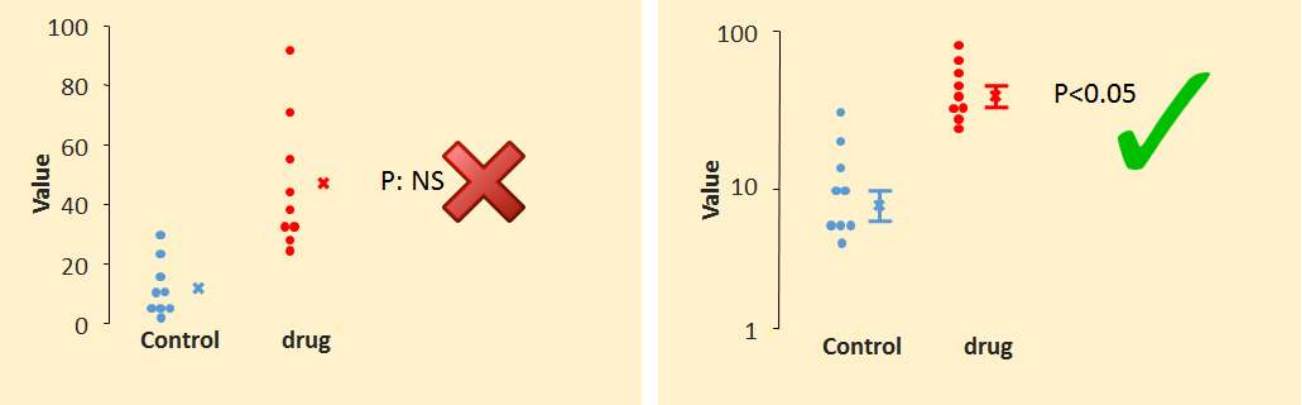


Figure 3. Individual data points (circles), mean values (x) and s.e.mean values are shown. On the left, owing to the large variance in the drug group, a t test identifies the two groups as not significantly different. However, closer examination shows this to be a false inference. The individual data values are not equally distributed around the arithmetic mean, and an arithmetic s.e.mean is not an appropriate way to summarize the distribution (and has therefore been omitted from the figure). If the s.e.mean were to be drawn its size would be proportional to the mean (i.e., the larger the mean, the larger the s.e.mean). It is possible to analyse the data in the left-hand figure using a non-parametric statistical test (such as a U test), but non-parametric tests are less powerful than parametric tests and their use can result in false negative findings. In the right-hand panel, the same data are log transformed and the Y axis uses the log scale. Here the s.e.mean is no longer proportional to the mean, and the values are Gaussian distributed. It is appropriate to show the s.e.mean (error bars in the figure). A t test correctly identifies a statistically significant difference between groups. This transformation unsettles some investigators as it appears to be a manipulation of data. However, in nature, many variables are log normally distributed. Sound (decibels) and acidity (pH) are units on a log scale, used because the variable is log Gaussian. In pharmacology the pA_2, and even the relationship between a response and a drug concentration are log Gaussian. This is why we express agonist and antagonist "affinity" values as pK_A and pK_B, respectively not K_A and K_B. It should be no surprise that many other variables in biology are log-Gaussian (for example the number of ectopic beats occurring in experimental myocardial infarction). The key issue here is that authors and peer reviewers should look at figures to ensure that data like those in the left-hand part of the figure are not included in a paper – if they are the data should be re-analysed.

Group sizes – what is a group size and what is a group?

We have indicated that for BJP n=5 is the minimum acceptable group size for statistical analysis but on this issue two other factors warrant consideration. The first is 'what is a group size'? We allude above to a group being comprised of independent samples. But what is an independent sample? For BJP, 'independent' signifies that the sample represents a bias-free representation of the population. Thus, 5 cells from one mouse is not n=5. On the other hand it may be argued that for *in vivo* experimentation datasets derived from 5 mice housed in the same cage should be averaged to provide n=1. BJP does ***not*** require that data from n=5 mice housed in the same cage should be treated as n=1. We will review this position as the field progresses. The second factor is 'what is a

group?' A group is a collection of independent samples rather than technical replicates. A technical replicate is one sample, run/tested/probed again (and again). A single sample of cells, divided into three and treated the same way, on the same day, by the same person is a technical triplicate. It is n=1. Thus, we expect statistical analysis to be undertaken on independent samples, demonstrably independent, and of a minimum group size.

Regarding minimum group sizes and power analysis, we have stated that a good rule of thumb is to add 50% to the n determined by power analysis, in order to avoid under powering. This is arbitrary guidance. It is based on the fact that power analysis defines the minimum group size to obtain significance that is reliable 50% of the time, and thus, that it makes sense to add 50% to the group. We may revisit this recommendation in the future.

Experimental outliers

We have debated how to handle outliers within datasets. A recommendation is summarised in the guidance list shown above. We begin by defining what we mean.

Outliers may be defined as data values that digress from the central tendency (often also described as 'central location'). There are a range of issues that need to be considered before an investigator can decide how to deal with them. First, how does one identify an outlier? With small group sizes ($n<12$) this is difficult or even impossible. It is feasible to define an outlier in a control population, but only if the control group has been generated numerous times previously so that a large number of control values are available for inspection.

If it is possible to define an outlier (i.e. a value that lies outside a defined range) then the next question to address is why a value is an outlier. The reason may be that the value is false, contaminated or in some other way incorrect. On the other hand, it may be *correct,* and the result of a natural wide spread of data, or even a bimodal distribution of data. The latter would arise, for example, if one were to analyse readouts in a population that expresses a polymorphism, such as deficiency in acetylcholinesterase in a small section of the human population. Excluding such outlier values/subjects is justified only in an experiment defined to be relevant only to the larger population, for instance, in the acetylcholinesterase example, those with typical enzyme activity. It is important to know what the explanation is for an outlier, since it is hazardous and potentially inappropriate to exclude correct values just because they alter the data distribution.

Genuinely false values *should* be excluded from a sample, but this must be done using *predefined* criteria. Using a formula based on the standard deviation is one. A number of others are available in routinely used statistical packages but their use can be problematic (Leys et al., 2013). Alternatively, one may use an arbitrary limit of acceptability (e.g. to exclude animals whose surgery has lowered blood pressure to below a value appropriate for testing drug effects especially, in this example, effects on blood pressure). This approach is entirely acceptable to BJP, however, the key is to define the exclusion criteria beforehand and apply them on blinded data to avoid bias. Exclusion criteria should be fully described in the Methods section.

With a novel type of study (i.e. where there are no historical controls and database of records with which to consult) it is essential to ensure unbiased quality control, which means undertaking preliminary studies in order to allow generation of arbitrary exclusion criteria, keeping in mind that

the justification cannot be 'scientific', merely pragmatic (to ensure that data can be analysed statistically without the need for onerously large group sizes). It is wise to revisit any criteria as new data emerge. We encourage authors to present such data within manuscripts.

When the reasons for exclusion of data are impossible to justify, the best solution may be to include all data including 'outliers' and ensure group sizes are large (increased by 50% from the value determined by Power analysis is advisable). When inclusion of outliers is decided to be the best option, this may generate non-Gaussian datasets. These can be accommodated by use of transformations or non-parametric statistical tests, as discussed above. In summary, outliers should be *included* in a data set unless a predefined and defensible set of exclusion criteria can be generated and applied.

Conclusions

Here we have updated and simplified journal requirements for experimental design and analysis. The objective is pragmatic – to facilitate manuscript preparation and peer review that is more consistent and transparent, and that generates research articles whose data are more likely to be reproducible.

The caveat is that there is no panacea since implementation of any process is entirely dependent on stakeholders engaging with it. If guidance is too onerous, too detailed or ambiguous, or presented as optional ('best practice') it will likely fail. If authors ignore the guidance and peer review fails to recognise this, we will make no progress. Keeping guidance and the process of its implementation simple has a better chance of success than elaboration of complex and detailed guidance on every nuance. We will revisit the guidance in 2020, but will also conduct 6 monthly audits, in order to monitor its effects, and will introduce new guidance as appropriate.

One issue that we have avoided concerns the meaning of statistical significance. It has been suggested that the threshold P that constitutes significance be lowered from the 'pharmacology standard' of $P<0.05$ to as low as $P<0.001$. Others have suggested that 'significance' be a forbidden word, and that exact P values be presented with the author (or even the reader) offered the freedom to make his/her own judgement as to whether to 'believe' if an 'effect' is 'real' or not. BJP continues with the policy that authros must define what level of P they define to constitute statistical significance within the Methods section of the paper. Authors may select a more stringent P threshold that the current norm of $P<0.05$, but this must not be varied from one part of a manuscript to another. We will remain watchful and will assess the evidence and revisit this issue as and when the discourse dictates and at the very least in 2020.

Statistical analysis does not guarantee that a finding is necessarily biologically true, and we will allow an author the right to argue that a false positive or a false negative finding may have been generated. Therefore, some room is needed for expert opinion and the judgement of experience (when argued persuasively). In addition, the calculated P value is almost always bigger than it seems ('less significant') owing to the false discovery rate which is one reason why some investigators argue that most (rather than just some) research findings are false (Colquhoun 2014; Begley 2013; Begley & Ioannidis 2015).

Thus, to summarise, this update describes a modified approach to issues that have arisen from our experiences following the publication in 2015 of our design and analysis guidelines. The areas of focus have been selected from our internal audits as issues requiring reconsideration and aspects of experimentation that we did not consider in the first iteration. We will continue to use this approach of internal audit and reappraisal to assess and to ensure that BJP remains at the forefront of the agenda encouraging good experimental practice. Our intention is to continue to support the Pharmacology community in identifying strategies that support and enable transparency and reproducibility. As we stated previously (Curtis et al 2015) some of our guidance is arbitrary, and some will change. We will make progress, but it will need clear requirements and vigilance with progress won in stages. This is our stage II.

List of references

Anonymous (2017) Transparency upgrade for Nature journals. Nature, 543: 288

Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of *p*-values. R. Soc. open sci. 1: 140216

Begley CG (2013) Reproducibility: six red flags for suspect work. Nature 497 , 433-434

Begley CG & JPA Ioannidis (2015). Reproducibility in science. Circulation research 116, 116-126

Curtis MJ, Bond RA, Spina D, Ahluwalia A, Alexander SPA, Giembycz MA, Gilchrist A, Hoyer D, Insel P, Izzo AA, Lawrence AJ, MacEwan DJ, Moon LDF, Wonnacott S, Weston AH, McGrath JC (2015) Experimental design and analysis and their reporting: new guidance for publication in BJP. Br J Pharmacol 172:2671-2674, 2015

Leys C, Ley C, Klein O, Bernard P (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. J Exp Social Psychol 49:764-766

McGrath JC & Lilley E (2015) Implementing guidelines on reporting research using animals (ARRIVE etc.): new requirements for publication in BJP. BPH12955