EDITORIAL

Clarification of the basis for the selection of requirements for publication in the *British Journal of Pharmacology*

Correspondence Amrita Ahluwalia, British Journal of Pharmacology, The Schild Plot, 16 Angel Gate, City Road, London EC1V 2PT, UK. E-mail: info@bps.ac.uk

Michael J Curtis[1], John C Ashton[2], Lawrence D F Moon[1] and Amrita Ahluwalia[3]

[1]Kings College London, London, UK, [2]University of Otago, Dunedin, New Zealand, and [3]Queen Mary University of London, London, UK

In 2015 and 2018, the *British Journal of Pharmacology* (*BJP*) published guidelines on experimental design and analysis (Curtis *et al.*, 2015, 2018). The intention was to improve the credibility of papers published in *BJP* by the simplest means possible. It is all very well for a journal to elaborate a framework of best practice, with lengthy explanations for each issue considered, but if authors, reviewers and editors fail to adopt the framework because it is too complex or nuanced, then we fail as a journal. Consequently, unlike most other journals (Williams *et al.*, 2018), *BJP* has opted for *firm rules* about a *small* number of issues, rather than generalized and lengthy 'best practice advice'. We focused on inconsistent reporting of P values (e.g. $P < 0.05$, P = exact value, $P <$ different values), persistent and unjustified use of $n = 3$ (or fewer), grossly unequal group sizes and an absence of randomization and blinding (each of which typically occurs together in many papers) that are particular problems in our sector and contribute to the failed replication that is undermining the credibility of preclinical research. We received two letters that criticize some of our guidance and have written an itemized reply below.

First, we make a general point. Most of the *BJP* guidelines are 'conventions', that is, pragmatic solutions to practical challenges. This is particularly relevant to *BJP*'s requirements for group size selection. Setting $n = 5$ as the minimum allowable for comparing groups by statistical analysis (the '$n = 5$ rule') is clearly a convention. We are not claiming $n = 5$ is sufficient and necessary for all studies. In some studies, group sizes much larger than $n = 5$ are necessary to reduce the risk of false findings, whereas in other studies, where the control outcome has been established repeatedly in previous published work, group sizes of fewer than $n = 5$ may be sufficient. In the main, *BJP* publishes papers on new drugs, or using new transgenic animals, or evaluating variables that have not been evaluated previously, often a combination of all three. *Novelty* is the key. When work is novel, it is extraordinarily rare for an author to include in their Methods section a clear statement that the data are known to be drawn from a normally distributed population (the necessary prerequisite for the type of parametric analysis typically undertaken) or that they have undertaken sample size calculations *a priori* that indicate that n = X would be adequate for their design. Consequently, it seems that deciding on an appropriate group size is done by after-the-fact power analysis using the data generated by a study to justify the group size used in the study (as opposed to *a priori* power analysis) or by 'informed judgement' (guesswork). Moreover, 'group sizes as small as possible' is normally the guiding principle. The resultant problem is that studies are often favourably treated by peer review if sufficiently novel, with no questioning of group size selection. This is not a problem that can be ignored. Most statistical software programs allow tests that run on small n (even $n = 2$), but the reliability of resultant P values diminishes as group sizes become smaller (Halsey *et al.*, 2015), and low power is widespread and leads to higher rates of false findings (Button *et al.*, 2013). Because, for novel work typical of that published in *BJP*, *a priori* power calculations are normally impossible, our $n = 5$ rule is therefore a convention that precludes default selection of smaller group sizes without adequate validation and is designed to facilitate confidence in study outcome.

However, there is a recent emergence of preclinical research where safeguards have indeed been put in place before the experiments were undertaken, with pre-registration of study design limiting unreported *post hoc* manipulation of analytical methods. Emmrich *et al.* (2018) is a good example of a pre-registered study that was modified transparently after post publication peer review of the design and proposed method of analysis. As a consequence, the Editors of *BJP* will *consider* findings of $n < 5$ where the designs and analyses for a study have been approved *a priori* and *published* in a

DOI:10.1111/bph.14443

pre-registered repository (e.g. Registered Reports; https://cos.io/rr/). However, we must emphasize that without such a disclosure, the $n = 5$ rule will continue to apply.

In addition to the comments in both letters regarding the issue discussed above, we respond to further points raised by the authors of the two letters below.

From the letter by Neuhäuser and Ruxton (2018), the first comment refers to our recommendation that authors design a study to have equal group sizes. The rationale for this was not made clear in either the 2015 or 2018 papers. Like the $n = 5$ rule, it is a convention, and the main reasons for it are as follows.

- Pre-registration of experimental design and intended methods of analysis is not yet common in our sector. We agree that optimally unbalanced groups can lead to improved sensitivity and power when the *a priori* decision is made to analyse them without ANOVA and with (for instance) Dunnett's tests back to a single comparator rather than all pairwise comparisons (Bate and Karp, 2014). However, in our experience, reviewers and editors often cannot tell whether experiments with unbalanced groups result from planned excellent design or unconsidered design and inadequate transparency, with attrition unreported and exclusions undeclared.
- Some investigators do not undertake blinded and randomized studies, and animals are added to, or removed from, the study after preliminary analysis. Typically, no explanation is given for such variation, and this is not picked up during peer review.
- When limited numbers of rare samples are available, an equal group size design is the safest way to minimize the risk that if there are lost samples, this will render the study unfit for analysis (e.g. $n = 6, 6, 6$ becoming $n = 6, 5, 5$ is preferable to $n = 12, 3, 3$ becoming $n = 12, 2, 2$).

By requiring authors to *declare* they have *designed* their study to have *equal* group sizes, we are requiring authors to think about their design. Nevertheless, we note the comment and have determined that the author guidance should be modified to state 'Exceptions to these guidelines will be considered (*e.g.*, normalized data analysed parametrically without a preceding ANOVA arising from unbalanced experiments with low n in treatment groups) for a result where a full description of the intended experimental designs and analyses have been published in a date-stamped, peer-reviewed preclinical registry together with *a priori* sample size calculations for each group involving adequate power (*e.g.*, Registered Reports; https://cos.io/rr/)'.

Neuhäuser and Ruxton (2018) go on to say 'a further reason for unequal group sizes is unequal variances. To increase power, a larger proportion of the total sample size should be allocated to a group with a larger variance'. Our comments on pre-planned and published protocols (above) apply here, and without this, we would expect studies to be designed to have equal group sizes. Otherwise, unequal variances means that the data are not fit for parametric statistical analysis (if transformation fails to homogenize variances). Also, it is difficult to see how one can undertake a randomized and blinded study *and* manipulate group sizes to 'accommodate' high variance in one group *unless* it were known *a priori* that one group will have a disproportionate variance, otherwise the accommodation would be a form of '*P* hacking' (Head *et al.*, 2015).

The next comment is 'a general principle to "add 50% to the calculated minimum group sizes" (Curtis *et al.*, 2018) is unusual and not reasonable'. Adding 50% was proposed as helpful advice rather than part of our list of requirements, so it is not one of our 'conventions'. Systematic reviews show that, usually, most studies are unreasonably underpowered which inflates the incidence of false findings (e.g. Button *et al.*, 2013). The 'general principle' alluded to above is a simple way to add to the '$n = 5$ rule' to encourage individuals to further increase group size.

Neuhäuser and Ruxton (2018) next state that 'We agree that significance in classical ANOVA can be caused by inhomogeneity in variances. But the recommendation not to carry out post-hoc tests in case of a significant variance inhomogeneity is not satisfying. A better strategy would be performing an ANOVA designed for possible variance inhomogeneity. Several methods have been proposed for this'. Variance inhomogeneity (which by necessity includes *large* variance in *some* groups) may cause false *negative* findings to be reported. We are saying that when conditions do not permit conventional parametric analysis, then an alternative must be found (we mentioned nonparametric tests and use of transforms). Moreover, we have said nothing to stop authors doing what is suggested in the statement quoted above.

Neuhäuser and Ruxton (2018) finally state 'Clearly, asymptotic or approximate tests are not acceptable for very small samples sizes, but the minimum 5 is completely arbitrary'. This is also noted by Motulsky and Michel (2018) in the second letter. We have addressed this important point at some length in our second paragraph, above.

Motulsky and Michel assert that following ANOVA, it is acceptable to conduct 'follow-up' tests even if F is not significant. We very much oppose the notion of encouraging investigators to routinely conduct ANOVA then routinely ignore the F value. ANOVA is undertaken to examine whether a factor (e.g. treatment) is a significant source of variance. If it is, then a *post hoc* test to identify *which* treatment (which level of the factor) is the source of variance is justified. If a study is not blinded or randomized, and indeed in addition is made up of groups with small n, there is every chance that variance inhomogeneity may undermine scope for F to reach significance and that real effects may be missed if *post hoc* tests are not undertaken. Essentially, false negatives may arise owing to failure to use a suitable design and not because ANOVA is intrinsically flawed. In many respects, this exemplifies why we created the *BJP* requirements – the experimental design and *a priori* choice of analysis is paramount, but this is out of the scope for peer review to thoroughly validate. Meanwhile, the choice of statistical test and its execution must follow the design. If the study has been designed and executed appropriately, the scope for false negative findings predicated by ANOVA will be minimized. However, we do agree to consider during the review process one exception to this rule, that is, in the case where planned comparisons (i.e. not simply pairwise *post hoc* comparisons) have been pre-registered and peer-reviewed *a priori* as explained above, these may be undertaken in the absence of a preceding ANOVA.

Motulsky and Michel (2018) additionally criticize the journal requirement that normalized data be analysed with nonparametric statistics. Our intention in the *BJP* guidance with this requirement was to stop the routine use of two-sample *t*-tests (or equivalent tests for multiple group comparisons) when the control group has no variance. There are certainly occasions where the use of one-sample *t*-test is valid so long as there is evidence that the assumptions of this parametric test are not violated. However, as we have argued elsewhere, it is not possible to determine with any confidence whether, for example, five randomly selected samples come from a population with a normal distribution or not, and so a nonparametric test is preferable, avoiding the need for baseless assumptions.

In conclusion, we thank the authors of the two letters for their interest in our guidance. We acknowledge the comments raised and agree that there are specific variations to some *BJP* design and analysis requirements that are acceptable and their inclusion should not preclude consideration of a manuscript by *BJP*. This will result in small changes to the design and analysis guidance. We will incorporate this into the journal's author guidelines and capture it in the next update editorial concerning design and analysis, which is likely to be published in 2021.

Conflict of interest

The authors declare no conflicts of interest.

References

Bate S, Karp NA (2014). A common control group – optimising the experiment design to maximise sensitivity. PLoS One 9: e114872.

Button KS, Ioannides JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ *et al*. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14: 365–376.

Curtis MJ, Alexander SPA, Cirino G, Docherty JR, George CH, Giembycz MA *et al*. (2018). Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers. Brit J Pharmacol 175: 987–993.

Curtis MJ, Bond RA, Spina D, Ahluwalia A, Alexander SPA, Giembycz MA *et al*. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. Br J Pharmacol 172: 2671–2674.

Emmrich JV, Neher JJ, Boehm-Sturm P, Enders M, Dirnagl U, Harms C (2018). Stage 1 registered report: effect of deficient phagocytosis on neuronal survival and neurological outcome after temporary middle cerebral artery occlusion (tMCAo). F1000Res 6: 1827.

Halsey LG, Curran-Everett D, Vowler SL, Drummond GB (2015). The fickle P value generates irreproducible results. Nat Methods 12: 179–185.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015). The extent and consequences of P-hacking in science. PLoS Biol 13: e1002106.

Motulsky HJ, Michel MC (2018). Commentary on the *BJP*'s new statistical reporting guidelines. Br J Pharmacol 175: 3636–3637.

Neuhäuser M, Ruxton GD (2018). Some comments on the update to *BJP* guidance on experimental design and analysis. Br J Pharmacol 175: 3638–3639.

Williams M, Mullane K, Curtis MJ (2018). Addressing reproducibility: peer review, impact factors, checklists, guidelines, and reproducibility initiatives. In: Williams M, Mullane K, Curtis MJ (eds). Research in the Biomedical Sciences. Elsevier: New York, USA, pp. 197–306.