BJP British Journal of Pharmacology

DOI:10.1111/bph.12856
www.brjpharmacol.org

EDITORIAL

Experimental design and analysis and their reporting: new guidance for publication in BJP

Correspondence
John C McGrath
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, Richard A Bond, Domenico Spina, Amrita Ahluwalia, Stephen P A Alexander, Mark A Giembycz, Annette Gilchrist, Daniel Hoyer, Paul A Insel, Angelo A Izzo, Andrew J Lawrence, David J MacEwan, Lawrence D F Moon, Sue Wonnacott, Arthur H Weston and John C McGrath

Editorial Office, British Journal of Pharmacology

LINKED EDITORIALS

This Editorial is part of a series. To view the other Editorials in this series, visit: http://onlinelibrary.wiley.com/doi/10.1111/bph.12956/abstract; http://onlinelibrary.wiley.com/doi/10.1111/bph.12954/abstract; http://onlinelibrary.wiley.com/doi/10.1111/bph.12955/abstract and http://onlinelibrary.wiley.com/doi/10.1111/bph.13112/abstract

Abbreviations

ALS, Amyotrophic Lateral Sclerosis; ANCOVA, analysis of covariance; ANOVA, analysis of variance; BJP, British Journal of Pharmacology; SEM, standard error of the mean; Δ, delta.

Introduction

This editorial is part of a series describing changes in the guidelines for authors making submissions to the *British Journal of Pharmacology* (BJP) (McGrath and Curtis, 2015); McGrath and Lilley, 2015; McGrath *et al*., 2015a,b). It sets out revised guidance for experimental design and analysis in preclinical pharmacological research for papers submitted to BJP. The main aim is to assist authors to provide the essential information that is required for their work to be reproducible and considered reliable.

We explain how authors are requested to follow this guidance, noting the advice applies to all work submitted after 1 August 2015. In the interests of flexibility, authors may argue for exemptions from the *design* requirements (Group sizes, power analysis, replicates and independent values; Randomization; Blinding) by including valid scientific justifications for doing so in the Methods section (explained below). The analysis requirements (normalization; statistical comparison), however, are not negotiable since they are based on straightforward statistical principles. This guide is based on the principles set out by Research Councils UK (http://www.rcuk.ac.uk/media/announcements/150415/: Research Councils UK, 2015) and U.S. National Institutes of Health 'Principles and Guidelines for Reporting Preclinical Research' (http://www.nih.gov/about/reporting-preclinical-research.htm: U.S. National Institutes of Health, 2014) (see also http://www.nigms.nih.gov/training/pages/clearinghouse-for-training-modules-to-enhance-data-reproducibility.aspx: U.S. National Institutes of Health, 2015) and is similar to the requirements of the more generic life science journal '*Nature*' (Nature Editorial, 2014), but is tailored to emphasize good practice and to eliminate weak practice, based on our experience of articles submitted to pharmacology journals.

The key points of the new guidance relating to reporting of experimental design and analysis are:

1. Experimental design should be subjected to '*a priori* power analysis' so as to ensure that the size of treatment and control groups is adequate to obtain a defined level of statistical significance, unless a valid scientific justifi-

cation is provided for reduced group size. The latter requires an *a priori* sample size calculation that should be included in Methods and should include alpha, power and effect size.

2. Owing to unreliable '*P*' values obtained with small sample sizes, group data subjected to statistical analysis should have a minimum of $n = 5$ independent samples/individuals per group, regardless of the outcome of any power analysis. Inclusion of smaller sized groups (which should not be subjected to statistical analysis) is permitted if a valid scientific justification for fewer than $n = 5$ is provided.
3. When small groups ($n < 20$) are used, they should be of equal size unless a valid scientific justification for unequal group sizes is provided. This may include variation due to loss of animals or samples; if so this should be explained, with exclusion criteria defined. Exclusions should preferably be replaced to keep the study balanced, and excluded values should certainly be replaced if the power of the study would otherwise be jeopardized.
4. In studies in which groups are compared, experimental subjects/preparations should be randomized to groups unless a valid scientific justification is provided for not doing so. The order of treatment should be randomized at the level of the experimental subject (i.e. all placebo treated animals should not be treated systematically before all drug treated animals even if animals were previously randomized into these two groups). The type of randomization should be stated explicitly (e.g. randomized block design). Given that the use of randomization designs is not ubiquitous, and cannot be applied retrospectively to studies already underway, the BJP editors are prepared to allow a moratorium on this requirement covering all papers submitted up to 1 August 2017. In the interim, all manuscripts should state explicitly whether or not and how studies were randomized (and if not, why not).
5. Assignment of subjects/preparations to groups, data recording and data analysis should be blinded to the operator and analyst unless a valid scientific justification is provided for not doing so. If it is impossible to blind the operator, for technical reasons, the data *analysis* can and should be blinded. Since blinded analyses cannot be applied retrospectively to studies already underway, we are prepared to allow a moratorium on this requirement to cover all papers submitted up to 1 August 2017. In the interim, all manuscripts should state explicitly whether or not and how studies were blinded (and if not, why not).
6. Normalization should not be undertaken unless a valid scientific justification is provided, such as normalizing to an internal standard (such as GAPDH in Western blotting) to reduce variance. It is legitimate to normalize all values (control and test) to the mean value of the experimental control group in order to set the Y axis so the control group value is 1 or 100%. If this is done, the correct units for such normalized data is % (or fold) of the control group's *mean* value, and graph Y axes and figure legends should be appropriately labelled to reflect this ('fold' or '% control' is insufficiently clear).
7. Certain types of normalization should not be undertaken under any circumstances. For example, individual experimental (test) group values should not be normalized to the matched experimental control group values and subjected to parametric analysis since the experimental control group will have no variance, and normalization should not be made to accommodate baseline differences between groups (i.e. significant differences between group mean values before intervention).
8. Transformation (e.g. by log) must be justified (i.e. by showing that it makes the SEM no longer proportional to the mean values)
9. One sample tested in triplicate is $n = 1$, not $n = 3$, and such 'pseudo-replication' should be used only to test the reliability of single values.
10. When comparing groups, a level of probability (*P*) deemed to constitute the threshold for statistical significance should be defined in Methods, and not varied later in Results (by presentation of multiple levels of significance). Thus, ordinarily $P < 0.05$ should be used throughout a paper to denote statistically significant differences between groups.
11. After analysis of variance, *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. Data transformation (e.g. log) may resolve the latter problem. If these criteria are not met, a *post hoc* test should not be run.

Practicalities of collecting and reviewing necessary information

We require authors to address the issues outlined above (along with details of statistical tests used) in a subsection of their Methods called *'Compliance with design and statistical analysis requirements'*: a template will be provided in which to enter this information. In addition, when submitting the manuscript, authors will be required to box-tick a 'declaration' based on this, which is shown in Table 1, called *'Declaration on Experimental Design, Statistical Analysis and Requirements for Studies using Animals'*. The declaration will additionally contain a further section covering aspects of research involving animals such as ethics and welfare, shown in the bottom part of Table 1. This is discussed in more detail in a separate editorial (McGrath and Lilley, 2015). The latter takes account of the ARRIVE guidelines (Kilkenny *et al*., 2010) and the recommendations of the Basel Declaration Society (2011), who are seeking to internationalize good practice (see also McGrath *et al*., 2015a).

The declaration will be used by authors submitting a manuscript to help them include everything that is required, and by referees to monitor compliance. Instruction is given on where in the actual paper the relevant information should appear. To facilitate this, submission will require uploading specified parts of the paper into a template. This process facilitates transparency since it requires authors to include as well as declare (and referees to check) more issues and items than previously.

In the remainder of this article we explain the reasoning behind this new advice, and include some guidance on how to address the issues.

Table 1

Declaration on Experimental Design, Statistical Analysis and Requirements for Studies using Animals

Group sizes

1. The exact group size (*n*) for each experimental group/condition is provided, not a range, and 'n' refers to independent values, not replicates.
2. Data subjected to statistical analysis have *n* of at least 5/group, and a scientific justification is provided for inclusion of any data not subjected to statistical analysis due to smaller group sizes.
3. Group sizes are equal by design, and any variation, due to experimental losses or violation of predetermined exclusion criteria, are explained.

Randomization

4. An explanation of how samples or animals were randomized to treatment is provided, and a valid scientific justification is provided if this was not the case.

Blinding

5. An explanation of how the operator and data analyst were blinded is provided, and if blinding was not undertaken, a valid scientific justification is provided.

Normalization

6. Where data normalization is employed, a valid scientific justification is provided and the units of the derived variables are correct.
7. Data obtained for parametric statistical analysis were not normalized so all control group values became 1, and any data normalization (e.g., to baseline or maximum values) has been justified by valid scientific explanation
8. A justification is provided for any data transformation (such as log transformation).

Statistical comparison

9. Group mean values and statistical analysis use independent values (any replicates have been used only to ensure reliability of an individual single value).
10. When comparing groups, a level of probability (*P*) deemed to constitute the threshold for statistical significance is defined in Methods, and not varied later in Results (by presentation of multiple levels of significance).
11. After ANOVA, a post hoc test was run only if F achieved the necessary level of statistical significance (i.e. $P < 0.05$) and there was no significant variance inhomogeneity.

If no animals were used in this study, go to item 18.

Validity of animal species or model selection

12. A scientific justification for the animal species and model selected for study is provided.

Ethical statement

13. An ethical statement indicating the body that approved the research is provided.

Animals

14. Source, species, strain, sex, range of age and weight of animals and any additional data that are relevant to the study are provided.

Experimental procedures

15. Details are provided of anaesthesia and analgesia, surgical procedures including asepsis, and method of killing for experimental procedures.

Housing and husbandry

16. Details are provided of:
    a. Housing (type of facility e.g. specific pathogen free [SPF]; type of cage or housing; bedding material; number of cage companions; tank shape and material etc. for fish).
    b. Husbandry conditions (e.g. breeding programme, light/dark cycle, temperature, quality of water etc., for fish, type of food, access to food and water, environmental enrichment).
    c. Welfare-related assessments, measurements and interventions (e.g. humane end points) that were carried out prior to, during, or after the experiment.

Interpretation

17. A statement is made concerning whether the study has any implications for replacement, refinement or reduction (the 3Rs).

Translation

18. A statement is made concerning the possible clinical relevance of the study.

Discussion of the issues involved in the new guidance for reporting experimental design and analysis

We (BJP senior editors with advice and assistance from other colleagues) have modified the guidance on design and analysis for manuscripts submitted to BJP to provide clearer advice on what we expect from authors, and what we expect to be evaluated by peer review. The goal is to increase the likelihood that published findings are found to be reproducible and reliable.

In peer review, standards apply to all aspects of a manuscript, such as inclusion of fair and scholarly reflection of the literature, cogent structuring of hypotheses, and appropriate description of methods used. Adherence to such standards is typically evaluated by referees and editors, applying their experience, expertise and knowledge, using subject-specific criteria. Separate from this, however, is a clear need for an *agreed* minimum level of general standards in experimental design and analysis that reflects essential requirements of all scientific investigation, which should be elaborated *as journal policy*. The current guidelines given by the BJP to authors and referees appear to be insufficiently explicit, judging by the variation in standards of design and analysis (and their reporting) in submitted manuscripts, and the approaches taken by referees and editors when evaluating papers as part of peer review. Consequently we have refined our guidelines and the way they are implemented.

To make manuscripts easier to review, authors are invited to include details of adherence in a specific place in the Methods section. Manuscripts submitted without conforming to the new design and analysis guidelines or that fail to provide the appropriate mandatory information regarding design and data analyses will be returned to the authors and may be rejected without full peer review.

Guidance for authors on appropriate statistical tests

Differences between groups – the t*-test and when to modify it*

Before addressing this question we must preface this section with a warning. In pharmacology, the most common question asked is whether a drug has induced a response, and statistics are used to establish whether or not the response was due to chance alone or not (i.e. 'statistically significant'). The first point to make is that one should not collect data then try out various statistical tests until one obtains the 'desired' result, known as 'P hacking' (Head *et al.*, 2015).

The most common scenario for this is a comparison between control and test groups. Statistical analysis is valid only if the data are stochastic (randomly determined – derived from a randomized study); whether this is the case depends upon whether an appropriate experimental design and subsequent execution were performed. These fundamentals are often ignored or not understood.

Analysis of numerical data is determined by whether values are 'all or none' (e.g. group incidence data) or part of an arithmetic continuum. If the latter, analysis is determined by whether the data are Gaussian (normally distributed) or not. Analysis of Gaussian distributed data is typically undertaken using a *t*-test or similar, generating a *P*-value that is used to identify an 'effect' (Altman *et al.*, 1983).

It has been cogently argued that the *t*-test can fail to take into account a 'false discovery rate' (Colquhoun, 2014). Even the meaning of *P* is misunderstood by some. Many interpret $P < 0.05$ as meaning the apparent difference between two groups has less than a 5% probability of occurring by chance. This may be reduced to 'less than a 5% chance of being wrong' about what we claim to be 'significant', that is, scientifically significant. However this is not strictly accurate. As Colquhoun (2014) has stated: 'If there were actually no effect (if the true difference between means were zero) then the probability of observing a value for the difference equal to, or greater than, that actually observed would be $P = 0.05$'. In other words, there is a 5% chance of seeing a difference at least as big as we have done, by chance alone. Even this does not accommodate the false discovery rate. 'In order to avoid making a fool of yourself you need to know how often you are right when you declare a result to be statistically significant, and how often you are wrong. In this context, being wrong means that you declare a result to be real when the true value of the difference is actually zero, that is, when the treatment and placebo are really identical. We can call this our 'false discovery rate' (Colquhoun, 2014) (defined on a scale of 0 to 1, with 0.8 meaning an 80% probability that when we report a "statistically significant" result, there is actually no real effect (the result is a false positive) and only a 20% probability that there actually is a true effect.

Unfortunately the false discovery rate is normally not known *a priori*. The solution to this problem is not agreed. Suggestions made include making the threshold *P* much harder to achieve; <0.001 has been suggested to be sufficient to reduce false discovery to close to zero if false discovery rate is 0.8 (Colquhoun, 2014), but this will introduce a high risk of false negative findings. *We hope to address this issue again in a few years' time when the problem has entered into to the consciousness of the pharmacology community.* Until then we propose that group sizes should never be less than 5; in a binomial distribution increasing group size from $n = 3–5$ improves the best attainable *P*-value from $P = 0.125$ to $P = 0.031$, and a similar improvement (though less easy to calculate accurately) occurs for the Gaussian ('normally') distributed data that we would subject to a *t*-test. It must be noted, that in more mathematically rigorous sciences, a *P*-value of 0.05 is not even considered (as a way of determining an effect): *P*-values used in determining the existence of the Higgs boson, for example, were many orders of magnitude lower (CMS Collaboration, 2012).

Pharmacologists are mostly interested in whether the means of groups differ, and in a majority of cases we use *P* as the tool. Although the true meaning of *P* is as defined above, we will hereon refer to *P* in less accurate but more succinct and accessible terms (e.g. probability of there being a difference).

Prospective (i.e. planned hypothesis-testing) investigations are the most common preclinical studies. Upon completion, if data are stochastic according to the design, variables presumed to be Gaussian distributed are subjected to ANOVA or co-variance (ANCOVA). This will determine whether the groups' means are similar. If they are not (i.e. F is found to be statistically significant) then a *post hoc* test may be undertaken (see below). In precise terms, the *P*-value from ANOVA/ANCOVA is a probability estimate of whether the groups are from one or more than one population. If there are only two groups, this analysis of variance is called a *t*-test and provides a *P*-values for a direct comparison between the two groups (since $F = t^2$).

If there *are* more than two groups, a further test (carried out '*post hoc*') is needed to determine *where* there are differences between groups (the F test having already established *whether* there is difference), that is, identifying the group(s) different from the others. A *t*-test will not suffice because it has to be repeated (two more times if there are three groups and five more times if there are four groups). Every time a *t*-test is done, with a probability of 0.05 as the cut-off for 'significance', the outcome of the test has a one in twenty chance of being false. Thus, if the same critical 't' statistic is used three times (as it would be in a study with three groups, each compared with each other), the likelihood of 'significance' being a chance event at least once increases to almost three in twenty (the exact value being 14.2%). To accommodate this difficulty, and restore the original one in twenty of the result being chance, the value of the critical t statistic is increased to accommodate the repeated measures. The modification varies according to whether groups are all compared with each other, or just with controls, giving rise to *post hoc* tests named after the originators (such as Bonferroni's, Dunnett's or Tukey's tests).

When *post hoc* tests are used, the level of *P* deemed to constitute 'statistical significance' should be decided before the experimental work is carried out, and should be described in the Methods. In pharmacology studies (and manuscripts reporting results of such studies), this is typically $P < 0.05$. Additionally, this means that it is inappropriate in the same paper to provide a variety of different *P* values (<0.01, <0.001, etc.) for different group versus group comparisons in Results and use such differences to imply that some differences are 'more significant' than others. If $P < 0.05$ is accepted as denoting that two groups are significantly different from one another, and 'statistical significance' is used as a means of identifying an 'effect', then using (e.g.) $P < 0.01$ to say the same thing about two groups elsewhere in the same paper is unhelpful.

It is useful here to remind ourselves that the *P*-value does not say anything about the size of the effect – only the likelihood that the effect is nominally 'real' (i.e. that the null hypothesis, that no difference exists between groups, has been rejected). As a guide, authors should keep in mind the question they are attempting to address.

Finally, in studies with complex experimental designs involving factors other than one experimental factor 2-way or 3-way ANOVA, it is important to conduct *post hoc* tests between types of data (e.g. between doses or between genotypes) only where the ANOVA indicates there is a source of variance.

In summary, BJP requires that appropriate statistical tests be used, and *P*-values described conservatively in accordance with the level of *P* defined as significant in Methods (typically $P < 0.05$, and consistently at the same level throughout a manuscript). 'Appropriate' statistics is however contingent on appropriate design, that is, the fulfilment of requirements concerning group sizes and randomization.

Guidance for authors on experimental design

Group sizes, power analysis, replicates and independent values

First, it is important to emphasize that individual values in a group that contribute to calculation of the group's mean value are derived from independent samples (biological replicates), not repeated measurements of single samples (technical replicates). If values are obtained by repeating runs of a single sample, the variance of these repeats can be small. However, technical replicates estimate only the precision of the value of that one sample, such that the variance of the estimate is simply a measure of the reliability of the measurement system (which includes the investigator). Repeating measurements of a single sample does not provide information about the *accuracy* of the value as representative of the population (the experimental group). That information requires the accumulation of a sufficient number of independent samples to fairly represent the experimental group. A simple example to illustrate this more clearly is, if one wanted to know the average height of males aged 56 living in Kent in the UK, then one would measure the height of an adequate sample of men aged 56 from Kent (let's say 20 men), not measure the height of one man aged 56 from Kent 20 times. This mistake is known as pseudo-replication (Lazic, 2010) and increases the chance of false positive results. The literature contains numerous examples of very small variance in a data set that in some cases represent average values of very few samples repeated several times (Lazic, 2010). Thus it is worth reiterating and emphasizing that, for example, when 6 samples are run 3 times each, the group size is 6, not 18, the SEM is the SD divided by $\sqrt{6}$, not $\sqrt{18}$, and the statistics should be done on 6 samples per group, not 18.

The value of the critical t-statistic (or its modifications – Bonferroni, Dunnett, Tukey, etc.) is determined by group size, and variability of values. Statistical power analysis to determine the minimum necessary group size is undertaken (Lehr, 1992) when the investigator knows the likely variability (defined for Gaussian distributed data as the variance) of the numerical values and the anticipated magnitude of the difference between groups. However, it is not easy to estimate exact group sizes that are necessary, and it is impossible if the likely variance or effect is not known. Group sizes below a certain level have low power and will generate frequent false positive results (Button *et al.*, 2013). Some statistical packages may not permit analysis with group sizes below $n = 5$, whereas others may, thereby making it possible to obtain $P < 0.05$ with group sizes of less than 5. These probabilities are not reliable.

Defining an exact minimum necessary group size has become important owing to a 'default' in parts of the pharmacology community (especially in studies that report Western blot experiments) to opt for a very low group size of $n = 3$. In the interim, before we can make a better informed proposal that acknowledges a convincing argument based on a large body of accumulated data for each test, BJP will require that group size for parametric testing be *at least* $n = 5$, regardless of any statistical power analysis calculation. Some investigators (including some authors of this article) argue that $n = 5$, as didactic guidance, is too small, and we will certainly revisit this when we update the guide.

Where it is possible, authors should provide clear statements about statistical power calculations in the Methods section of a submitted manuscript. Such power analyses are usefully viewed in terms of how much extra power is introduced by increasing group sizes beyond $n = 5$ and should not be used, for example, to determine the minimum feasible group size that would *just allow*, statistically, detection of an effect, since the necessary effect size would need to be extreme, it being approximately inversely proportional to group size.

For binomially distributed variables (yes/no outcomes such as group incidence as a % of group size) similar considerations apply (Mainland *et al*., 1956).

The guidelines on group sizes end with: 'unless a valid scientific justification for fewer (than $n = 5$) is provided'. Clearly we have argued that when conducting statistical analysis there can be no valid scientific justification for using fewer than $n = 5$/group, owing to issues of data reliability. However, if statistical analysis is not undertaken then the rule need not apply. Of course this puts great demands on the author to provide a scientific justification for including data that, owing to BJP's group size requirements, cannot be subjected to statistical analysis. Clearly such data need to be labelled as 'exploratory' or preliminary', and should constitute only a small proportion of the data in the paper. Any scientific justification for inclusion of such data should make reference to the impracticality of increasing group size and/or the value of including such preliminary findings in the paper, set against the bulk of the other data in the study that *do* have group sizes adequate to permit statistical analysis.

In summary, for data subjected to statistical analysis, BJP requires group sizes of at least $n = 5$ independent values whether or not power analysis has been undertaken (with use of larger group sizes if power analysis indicates this is needed). Replicates (e.g. 'in triplicate') provide one numerical value, and must not be used to generate multiple values (e.g. three from a triplicate) for statistical analysis and comparison between groups.

Randomization

All of the above is predicated on the measured variable being stochastic (drawn as random samples from a population). Thus, if we have two groups, one treated and one control, and we measure a variable to examine if the treatment affects it, the intervention (control vs. treatment) must be randomly assigned to the sample of the population (the test subjects).

The consequence of not doing this should be self-evident. For example, imagine if, on a Monday, a series of 6 pieces of guinea pig ileum are used to generate a concentration-response curve to noradrenaline, and on the next day another 6 pieces of ileum are used to replicate the experiment in the presence of a drug (for example, a putative antagonist). It is not possible to ascribe any difference to the drug, no matter how big the difference, since one cannot distinguish an effect of the drug from an effect of 'day of the week' (which could result from innumerable possible sources of variation, such as differences in preparation of solutions). Likewise, even if the entire experiment is conducted on Monday but the drug effect is determined by 'before versus after' comparison ('paired' analysis of two consecutive noradrenaline concentration-response curves, one before and the other after the drug's administration) the design does not allow discrimination between an effect of the drug and the possibility of time-dependent 'rundown' or 'warm up' (depending on how the agonist responses change). It may be possible to control for such sources of error by washing out the drug and showing the noradrenaline responses (slope, EC_{50} and maximum) are fully restored after washout, but if they are not fully restored (which is quite likely owing to the difficulty of 'washing out' an antagonist, and the unknown influence of time-dependent variation in responsiveness) then the data will be un-interpretable. In addition, regardless of the design, the drug's vehicle could have an effect, and one must 'control' for such an effect as well.

It is possible to argue that real drug effects are 'bloody obvious', and that if effects are not, then they are probably not very interesting or important (Kitchen, 1987), with self-evident implications about design and analysis. However, small effects may be very important. For example a 5% reduction in risk of sudden cardiac death would lead to 100–200 lives saved per 1 000 000 of the population per annum in Europe and North America, according to the present rate of sudden cardiac death reported by John *et al.* (2012). Moreover, no scientific experimental outcome is so obvious that it precludes the necessity of experimental verification, and the reproducibility of results is fundamental to experimental verification. Thus, seeking to remove avoidable unwanted sources of error is inescapably mandatory, and doing so facilitates the reproducibility of results.

In the example above, a simple solution is for a third party to code the drug and the drug's vehicle stock bottles, blind the operator and analysts, and randomize tissues to two groups. The tissues are first exposed to noradrenaline to ensure that the slope, the maximum and the EC_{50} each attain threshold levels for inclusion in the study (by reference to predetermined values; the exclusion criteria). The tissues are then exposed to drug or vehicle, and then to noradrenaline, again, in the presence of drug or vehicle. The second set of curve parameters are used to compare between groups by unpaired analysis. The first set may also be compared to show that the two groups are indistinguishable at baseline (a necessary prerequisite of quality control). By this design, a *t*-test will reliably determine whether the curve parameters come from the same population or not, at a predetermined *P* level.

Most problems of interpretation arise from difficulties in detecting and assimilating the value of weak drug effects. While it is true that large drug effects (or dramatic and unanticipated outcomes such as death) are hard to miss, no matter how poorly the experiment has been designed, it is best practice and safer to remove bias from an experimental design by randomizing the study. This has been recognized for decades (e.g. Mainland, 1957). Moreover it is *necessary* to randomize when the statistical analysis (e.g. ANOVA, *t*-tests, the variations thereon, and chi^2) requires a random sample of the population for validity (which assumes stochastic data).

Randomization can be achieved easily. If in doubt, tables have been published illustrating different types of randomization based on numbers of groups, maximum number of individual experiments achievable per day, and cost/availability of test article (Curtis *et al*., 2013).

Blinding

Blinding to eliminate a placebo effect and/or observer bias is obligatory in typical clinical trials and studies that assess drug effects in the treatment of patients. However, unfortunately, in preclinical research, experiments are often not blinded. One reason is that placebo effects are not generally thought to occur in preclinical (animal) research. Nevertheless, observer bias can occur and this can be subtle or even subconscious. It is possible to affect an outcome in innumerable imperceptible ways. Blinding eliminates all avoidable bias, irrespective of the source.

It is sometimes claimed that blinding is too difficult to incorporate in preclinical research. It may not be appropriate for the 'Principal Investigator'/'Supervisor' to do the blinding, as he/she may be involved during data analysis in making decisions about inclusion/exclusion/outlier handling. However, blinding can usually be achieved easily with an independent third party. If it is impossible to blind the operator to assignment to groups and data recording owing, for example, to a test group having an obvious phenotype (e.g. obese), the data *analysis* can and should still nevertheless be blinded by blinding the analyst. This may be achieved by having records (physiological data traces, samples for biochemical analysis etc.) re-coded by a second person, or by giving the records to a second person to analyse. None of us work in isolation. It can be argued that part of the responsibility of an organization in which research is conducted should be to foster co-operation among scientists such that those from different laboratories assist in blinding one another's studies. Moreover, although the benefit may not be obvious in the short term, it will become increasingly more evident as a laboratory accumulates data, knowing that all of it was free of observer bias. Blinding has the additional value that beyond eliminating unconscious bias, the possibility of fraud is greatly diminished.

Blinding a study will, intuitively, reduce bias. When intra- and inter-individual variability are low, and sample size relatively small (under 100 in the Gensburger example cited below), blinding may have little effect on study outcome (Gensburger *et al.*, 2012). However, it has been shown that blinding can affect outcome, even with binary outcomes as determined from a meta-analysis of a large population sample (over 4000 in the example cited; Hróbjartsson *et al.*, 2012). There is, additionally, the famous case of Benveniste who published a paper in Nature (which was never retracted) that identified an apparent activity of antibodies at a 'concentration' below that necessary to provide a single molecule in the solutions used in the experiment (Davenas *et al.*, 1988). The veracity of the outcome was later tested by an investigatory team from Nature who failed to replicate the outcome *in a blinded study* (see Ball, 2004).

Guidance for authors on data normalization and transformation

By far the most common form of data processing prior to analysis is normalization (expression of values as a percentage or ratio of another set of values). Normalization is undertaken to correct for an identified source of covariance, and should be practised when a source of covariance in data sets is identified. For example, If an investigator wishes to see if a drug lowers blood pressure and compares the drug to a vehicle control, and baseline varies between animals, a *t*-test comparing mmHg blood pressure after drug or vehicle may yield $P > 0.05$ owing to baseline variability obscuring the drug effect. This can be overcome by expressing individual blood pressure values after drug or vehicle as % of baseline, using each individual as its control. The two groups can thus be compared by (paired) *t*-test in terms of % baseline or the change (Δ) from baseline in mmHg. This type of approach is standard practice in clinical research.

The objective of normalization, therefore, is always to minimize the influence of within-group variability of individual baseline values. It is not to be used to 'normalize' unexplained or unanticipated differences between groups at baseline that may have arisen due to lack of randomization, or accommodate for a lack of calibration. If significant baseline differences in mean values exist between groups, then the experiment becomes uninterpretable, even with normalization. Note that all the caveats about controlling for time-dependent changes in values, etc., discussed above, apply.

Normalization is common in the quantitative analysis of immunoblots, or analyses of expression of mRNA. Individual control values can vary enormously in a study owing to the numerous sources of variation (such as duration of exposure in blot studies). Consequently, many investigators express the intervention measurement as % of control. If there is an independent control included in each run (an intensity control) then it may be possible to express all values, including the *experimental* control group's value, as a % of the intensity of the control (using individual intensity controls from each blot). However, BJP is increasingly receiving articles for peer review that have no independent control for blot intensity. Moreover, authors often present the control group values as 1, or 100%, with no variance (±0). The authors then compare the control group with test groups (with all values expressed as % or ratio of control), by parametric analysis (ANOVA and modified *t*-test). This is not valid because there is a variance inhomogeneity; one group (control) has no variance. Statistical software that is run properly will tell the investigator to desist from *post hoc* analysis because of the lack of variance in one group and the resultant 'variance inhomogeneity'. However, an investigator may be able to force the software package to do a *post hoc* test. Unfortunately, the answers obtained are not valid.

BJP (and doubtless other journals) has also received manuscripts with data sets of this sort in which the control group has a mean of 1 or 100%, yet the value has apparent variance (i.e. an SEM is shown). It appears that authors have calculated a mean of raw values for the control group (intensity of blot, for example), then expressed each individual control value as a % or ratio of this mean. The authors then compare these values with test groups whose individual values are calculated from the ratio or % of the control value in the individual experiment (e.g. gel run). There are two problems with this.

First the explanation of the calculation is often missing from the Methods. Second, although the authors provide (in their figure or data table) an SEM for the control group, the individual control values used in the statistical test calculation have no variance (and an SEM of zero) since each individual control value entered into the statistical test calculation is 1 or 100%. Consequently the presentation looks convincing (the controls appear to have a variance similar to the test groups) but the underlying statistical method is flawed, and the control SEM shown is misleading.

We note that with certain data sets, some authors calculate the control mean, and then express all the individual control values and all the individual test values as a % of the control mean, and conduct statistical analysis on these normalized values. This is mathematically acceptable. However, it is not a technique that reduces variability. This is because the ratios of each mean to each SEM, and each mean to each mean (control to test) are identical to the ratios of the equivalent 'raw' numerical values. This is a technique that does nothing more than re-label the Y axis of a figure so that the control group's mean value is designated 1 (or 100%). This may seem harmless, but it gives the false impression that control values (blot intensity, amount of protein, etc.) are identical from one experimental data set to another, and from one study to another. Moreover, authors rarely explain how they derived their values in Methods, and do not label the Y axis correctly (it should be '% of control *mean*', not '% control' or 'fold change').

Transformation is the conversion of data to a form that fits a known statistical distribution, for example, Gaussian, Poisson, etc., and is a method of managing data sets in which the SEM is proportional to the mean (which introduces a variance inhomogeneity that precludes use of parametric statistical testing). Individual numerical values may be transformed by an appropriate mathematical formula to generate data that are Gaussian distributed. An example is the log transformation, which has been commonly used to allow certain types of cardiac arrhythmia data to be subjected to ANOVA (Curtis *et al*., 2013). Such an approach may be helpful to authors seeking to ensure compliance with BJP requirements for ensuring that appropriate statistical tests are used. It is, of course, important to show the transform is valid before using it (e.g. it removes the proportionality between mean and SEM).

BJP requires that any normalization or transformation should be justified and explained in the Methods, avoided where possible, never used to mask baseline differences (that ought not to be present) between groups, and have correct labelling of Y axes of figures with normalized values.

Exploratory versus confirmatory studies

In some instances, it may not be possible to obtain data of sufficient quantity to permit statistical analysis yet the information may be of value to confirm other findings that are part of the same study. This might be considered 'preliminary' since although the sample is too small to be used statistically, it can, nevertheless, be a guide within a broader narrative if it constitutes preliminary data that stimulates a full study.

BJP is open to publishing preliminary data provided the experiment adheres to the fundamental requirements of scientific investigation. Thus, the provenance of preliminary data should be made clear (i.e. the author must provide details of randomization and blinding, and provide an explanation for the experiment's underpowered nature). Moreover if an author includes underpowered data, it is imperative they include *all* the underpowered data sets that were part of the original study or the work will be biased by exclusion of preliminary data that do not 'fit' the argument.

Finally, it is important to reiterate that all studies should be *designed* before they are analysed. The choice of method of analysis is dictated by the experimental design. Studies should be designed to be *appropriate* to permit statistical analysis, meaning that underpowered data will ordinarily not be generated.

There may, also, be cases where an assay is used for guidance towards determining, for example, the most 'potent' of many compounds and '*n*' for each compound or concentration may be small. The key here is to be transparent about how the design decision and the basis for data interpretation have been made, rather than to engage in spurious statistical analysis. Appropriate statistics can be undertaken later in appropriately designed studies with the lead compounds.

This section illustrates the subtle difference between guidelines for design versus guidelines for analysis, and illustrates that BJP will consider a wide range of approaches, according to context, provided convincing justification and full transparency are provided.

Even if statistics suggest a difference, is this different enough to matter?

Statistical significance and biological significance are different and should not be confused and, although they may coincide, they cannot be interchanged. Numerous examples of drugs that produce, for example, a 20% reduction in a pathological variable in a disease model at a level of high statistically significance exist, but this is not very useful if at least a *50%* reduction is required to 'treat' the condition. We require that authors clarify the biological significance of their data, and urge referees to ask for the information to be elaborated if it is unclear in a manuscript. It is, at times, not easy to define the threshold for biological significance and a target effect size may be difficult to determine ahead of studies: a 10% change may be very relevant in some gene expression studies, whereas a 90% reduction in virus titre may yet be irrelevant in some viral infection studies.

In addition, there are important issues relating to the level of probability selected to denote statistical significance. As noted earlier, Colquhoun (2014) has argued that $P < 0.05$ (the minimum level we accept) is not an appropriate cut-off to denote significance; in his view authors may run the risk of generating false positives, thus 'making a fool of themselves' if they select $P < 0.05$. He advocates opting for $P < 0.001$. This may be wise but we will not require it for BJP articles (presently).

False positives can arise more often when a study has not been randomized, and/or is grossly underpowered (Button *et al*., 2013), but false positives may nevertheless arise from chance alone. False negatives may arise because of poor design but, on the other hand, they are also dependent on variance (innate variance, and the variance introduced by poor experimental technique). Thus there is clearly a risk of type 2 error (failure to detect a real effect) if one attempts to correct for false positives by increasing the height of the 'P bar', and this will be without obtaining any extra benefit of avoiding false negatives (type 1 error).

Thus our recommendation is to design the study with adequate group sizes, opting for safety rather than just the bare minimum that a power analysis may indicate. There are instances where power analysis is difficult to perform before the experiments because, depending on the type of investigation, the relationship between effect size and biological significance are difficult to predict, especially if large baseline variations exist. With a homogeneous cell population, $n = 5$ may provide robust data whereas, in clinical research, 60 patients may not be sufficient, depending on disease complexity. We ask authors to justify their pre-determined group size selection (and, as noted above, *never undertake statistical analysis between groups if group size is* $n < 5$).

The future

Some investigators have suggested that using '*P*' to denote 'statistical significance' as a way to denote detection of an 'effect' is inappropriate, and offer other solutions such as provision of effect size estimates and their precision from confidence intervals (Colquhoun, 2014; Halsey *et al*., 2015). There is certainly value in such suggestions and we intend to revisit them once we have evaluated the impact of our present, more modest, proposals.

It is certainly the case, as explained by Halsey *et al*. (2015), that unreliability of findings, and 'lack of repeatability' are the most concerning problems that presently face publication of biomedical research. Our first attempt to address this is to ask authors to ensure they do not violate the fundamental requirements of experimentation by the use of inappropriate design, and by avoiding egregious errors of analysis.

We are also aware of the growing pressure to 'reduce' the numbers of experiments we conduct (e.g. see Parker and Browne, 2014). A radical change to how most of us approach statistical analysis will not resolve the damage done by the use of very small group sizes, etc., so we must take change in stages. Indeed, some colleagues consulted or co-authoring this article have argued that no study with $n < 10$ should be considered for publication, in order to ensure greater reliability of findings. On the other hand, other colleagues argue that no design and analysis requirements should be set 'in stone' for authors, and peer review should be allowed to take its course on a case by case basis. Consequently this article may appear arbitrary and dictatorial in places to some, but to others it may seem too lax.

In order to emphasize the need for a set of precise minimum necessary requirements, let us consider the threat posed by the alternative: offering 'best practice advice' and leaving decisions about the acceptability of a paper for publication entirely to the discretion and judgment of individual referees (peer review).

The threats posed by failure to change practices

A recent evaluation of a random sample of preclinical research suggests that fewer than 30% of studies in the biomedical preclinical literature report randomization and fewer than 5% report blinded analysis (Ioannidis *et al*., 2014), a level that has been described, correctly in our view, as 'pitifully infrequent' (Macleod *et al*., 2014). Although the consequences of this cannot be measured directly, one would anticipate that a proportion of published findings and conclusions are not reproducible because they are incorrect.

A variety of different types of analysis appears to confirm this problem. For example, in the amyotrophic lateral sclerosis (ALS) field, more than 50 published papers, published prior to 2008, identified survival benefit from a range of drugs in the standard mouse model (SOD1^{G93A}), yet almost all failed to translate and provide human benefit (Scott *et al*., 2008). In a disturbing analysis, Scott *et al*. (2008) concluded that 'the high noise floor of the model and the failure of the selected studies to replicate support the conclusion that the bulk of published studies using the SOD1^{G93A} mouse model may unfortunately be measurements of biological variability due to inappropriate study design'.

To that we would add 'and analysis'. A recent large study of preclinical research publications in neurological science found that the preponderance of statistically significant effects was not consistent with expected outcome based on study sample sizes, leading the authors to conclude that 'selective outcome and analysis reporting' was the most plausible source of the unequivocal biases identified (Tsilidis *et al*., 2013).

Inadequate design and analysis therefore have potentially devastating consequences, sabotaging the path to translation of results from preclinical studies to the clinical environment and undermining confidence in the preclinical drug discovery process, particularly at the early stages of target validation. It has been suggested that 'at least 50% of published studies from academic laboratories cannot be repeated in an industrial setting' (Mullard, 2011). We cannot sit idly by and tolerate this level of misinformation. Our modest recommendations are thus intended to shift the balance in favour of reliability.

It is beyond the scope of this article to offer a statistics primer, or a comprehensive guide on the theory of experimentation. We recommend that all authors read and note the advice of the Animal Research Reporting In Vivo Experiments (ARRIVE) guidelines (Kilkenny *et al*., 2010), the series of articles in BJP collected under the title 'Best practice in statistical reporting' (including the recent article by Motulsky, 2015), Colquhoun (2014), Marino (2014), Kenakin *et al*. (2014) and Halsey *et al*. (2015). The ARRIVE guidelines cover many areas and are open to interpretation, and reportedly are not being followed or assiduously enforced in published studies (Baker *et al*., 2014). Indeed, we agree with the recent assessment that rather than attempting to enforce the reporting of details that have previously been published, a 'pragmatic approach might be to implement the most important aspects of the guidelines', such as requiring the reporting of the extent of blinding and randomization, ensuring adequate group sizes (Baker *et al*., 2014), and that these aspects are checked by peer review. There is no point generating an exhaustive list of manuscript requirements if it is impossible for compliance to be properly checked. We thus contend that it is better to define a minimum level of mandatory requirements, and ensure they are clear, and verifiable by peer review.

Conclusion

Pharmacologists seek to determine whether an intervention (drug, genetic modification, etc.) has an effect. To achieve this there is a need to eliminate factors other than the intervention that may affect the readout. This means eliminating untoward sources of variance by blinding a study, properly randomizing the treatment, and utilizing adequate power in the study design to allow detection of a scientifically meaningful difference, and appropriately processing the numerical readout. Many of our referees and editors already ask authors to ensure their work adheres to these standards. After 1 August 2015 this will be a *requirement* for publication in BJP. There will be a moratorium on blinding and randomization until 1 August 2017. In the interim, manuscript authors must state explicitly whether or not randomization and blinding were performed, if so how, and if not, why not. A declaration for authors will be required at the point of submission of an article (see Table 1) and essential information will be gathered in a template of the Methods section.

Acknowledgements

We thank Michael JA Walker and Elspeth M McLachlan for their very helpful edits, and David Colquhoun for his challenging advice and comments. We are also grateful to many other colleagues for comments and suggestions, particularly Mick Bakhle, Michael Lew and Harvey Motulsky.

Author contributions

The article originated from discussions at the regular meetings of the Senior Editors of BJP during 2013 and 2014. M. J. C. coordinated the writing of the manuscript with contributions and edits from all of the other authors.

Conflict of interest

None.

References

Altman DG, Gore SM, Gardner MJ, Pocock SJ (1983). Statistical guidelines for contributors to medical journals. Brit Med J 286: 1489–1493.

Baker D, Lidster K, Sottomayor A, Amor S (2014). Two years later: journals are not yet enforcing the ARRIVE Guidelines on reporting standards for pre-clinical animal studies. PLoS Biol 12: e1001756.

Ball P (2004). The memory of water. The life and work of Jacques Benveniste taught us valuable lessons about how to deal with fringe science. Available at: http://www.nature.com/news/2004/041004/full/news041004-19.html (accessed 15/6/2014).

Basel Declaration Society (2011). The Basel Declaration. Available at: http://www.basel-declaration.org (accessed 10/2/2015).

Best practice in statistical reporting. British Journal of Pharmacology Virtual Issue: Best Practice in Statistical Reporting http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm

Button KS, Ioannidis 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.

CMS Collaboration (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys Lett B 716: 30–61.

Colquhoun D (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. Royal Society Open Science. 20141:140216. DOI: 10.1098/rsos.140216.

Curtis MJ, Hancox JC, Farkas A, Wainwright CL, Stables CL, Saint DA *et al*. (2013). The Lambeth Conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias. Pharmaco Ther 139: 213–248

Davenas E, Beauvais F, Amara J, Oberbaum M, Robinzon B, Miadonnai A *et al*. (1988). Human basophil degranulation triggered by very dilute antiserum against IgE. Nature 333: 816–818.

Gensburger D, Roux JP, Arlot M, Sornay-Rendu E, Ravaud P, Chapurlat R (2012). Influence of blinding sequence of radiographs on the reproducibility and sensitivity to change of joint space width measurement in knee osteoarthritis. Arthritis Care Res 62: 1699–1705.

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

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

Hróbjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I *et al*. (2012). Surgeon observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012: e1119.

Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D *et al*. (2014). Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383: 166–175.

John RM, Tedrow UB, Koplan BA, Albert CM, Epstein LM, Sweeney MO *et al*. (2012). Ventricular arrhythmias and sudden cardiac death. Lancet 380: 1520–1529.

Kenakin T, Bylund DB, Toews ML, Mullane K, Winquist RJ, Williams M (2014). Replicated, replicable and relevant-target engagement and pharmacological experimentation in the 21st century. Biochem Pharmacol 87: 64–77.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8: e1000412.

Kitchen I (1987). Statistics and pharmacology: the bloody obvious test. Trends Pharmacol Sci 8: 252–253.

Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci 11: 5.

Lehr R (1992). Sixteen S-squared over D-squared: a relation for crude sample size estimates. Statistic in Medicine 11: 1099–1102. http://www.gpower.hhu.de (accessed 04/03/15).

McGrath JC, Curtis MJ (2015). BJP is changing its requirements for scientific papers to increase transparency. Br J Pharmacol 172: 2671–2674.

McGrath JC, Lilley E (2015). Implementing guidelines on reporting research using animals (ARRIVE etc.): new requirements for publication in BJP. Br J Pharmacol 172: 3189–3193.

McGrath JC, McLachlan EM, Zeller R (2015a). Transparency in Research involving Animals: The Basel Declaration and new principles for reporting research in BJP manuscripts. Br J Pharmacol 172: 2427–2432.

McGrath JC, Pawson AJ, Sharman JL, Alexander SPH (2015b). BJP is linking its articles to the IUPHAR/BPS Guide to PHARMACOLOGY. Br J Pharmacol 172: 2929–2932.

Macleod MR, Michie S, Roberts I, Ulrich Dirnagl U, Chalmers I, Ioannidis JPA *et al*. (2014). Design animal studies better. Nature 510: 35.

Mainland D, Herrera L, Sutcliffe MI (1956). Statistical tables for use with binomial samples – contingeny tests, confidence limits and sample size estimates. Department of Medical Statistics, New York University College of Medicine.

Mainland D (1957). Safety in numbers. Circulation 16: 784–790.

Marino MJ (2014). The use and misuse of statistical methodologies in pharmacology research. Biochem Pharmacol 87: 78–92.

Motulsky H (2015). Common misconceptions about data analysis and statistics and how to avoid them. Br J Pharmacol 172: 2126–2132.

Mullard A (2011). Reliability of 'new drug target' claims called into question. Nat Rev Drug Discov 10: 643–644.

Nature Editorial (2014). Journals unite for reproducibility Nature 515: 7.

Parker RM, Browne WJ (2014). The Place of Experimental Design and Statistics in the 3Rs. ILAR J 55: 477–485.

Research Councils UK (2015). Updated RCUK guidance for funding applications involving animal research. Available at: http://www.rcuk.ac.uk/media/announcements/150415/ (accessed 24/04/2015).

Scott S, Ktranz JE, Cole J, Lincecum JM, Thompson K, Kelly N *et al*. (2008). Design, power, and interpretation of studies in the standard murine model of ALS. Amyotroph Lateral Scler 9: 4–15.

Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW *et al*. (2013). Evaluation of excess significance bias in animal studies of neurological disease. PLoS Biol 11: e1001609.

U.S. National Institutes of Health (2014). Principles and Guidelines for Reporting Preclinical Research: Available at: http://www.nih.gov/about/reporting-preclinical-research.htm (accessed 10/2/2015).

U.S. National Institutes of Health (2015). http://www.nigms.nih.gov/training/pages/clearinghouse-for-training-modules-to-enhance-data-reproducibility.aspx: (accessed 4/4/2015).