Commentary

New Author Guidelines for Displaying Data and Reporting Data Analysis and Statistical Methods in Experimental Biology

✉ Martin C. Michel, T.J. Murphy, and ✉ Harvey J. Motulsky

Department of Pharmacology, Johannes Gutenberg University, Mainz, Germany (M.C.M.); Partnership for the Assessment and Accreditation of Scientific Practice, Heidelberg, Germany (M.C.M.); Department of Pharmacology and Chemical Biology, Emory University, Atlanta, Georgia (T.J.M.); and GraphPad Software, Los Angeles, California (H.J.M.)

Received November 22, 2019; accepted November 22, 2019

ABSTRACT

The American Society for Pharmacology and Experimental Therapeutics has revised the Instructions to Authors for *Drug Metabolism and Disposition*, *Journal of Pharmacology and Experimental Therapeutics*, and *Molecular Pharmacology*. These revisions relate to data analysis (including statistical analysis) and reporting but do not tell investigators how to design and perform their experiments. Their overall focus is on greater granularity in the description of what has been done and found. Key recommendations include the need to differentiate between preplanned, hypothesis-testing, and exploratory experiments or studies; explanations of whether key elements of study design, such as sample size and choice of specific statistical tests, had been specified before any data were obtained or

adapted thereafter; and explanations of whether any outliers (data points or entire experiments) were eliminated and when the rules for doing so had been defined. Variability should be described by S.D. or interquartile range, and precision should be described by confidence intervals; S.E. should not be used. *P* values should be used sparingly; in most cases, reporting differences or ratios (effect sizes) with their confidence intervals will be preferred. Depiction of data in figures should provide as much granularity as possible, e.g., by replacing bar graphs with scatter plots wherever feasible and violin or box-and-whisker plots when not. This editorial explains the revisions and the underlying scientific rationale. We believe that these revised guidelines will lead to a less biased and more transparent reporting of research findings.

Introduction

Numerous reports in recent years have pointed out that published results are often not reproducible and that published statistical analyses are often not performed or interpreted properly (e.g., Prinz et al., 2011; Begley and Ellis, 2012; Collins and Tabak, 2014; Freedman and Gibson, 2015; National Academies of Sciences Engineering and Medicine, 2019). Funding agencies, journals, and academic societies have addressed these issues with best practice statements, guidelines, and researcher checklists (<https://acmedsci.ac.uk/file-download/38189-56531416e2949.pdf>; Jarvis and Williams, 2016).

In 2014, the National Institutes of Health met with editors of many journals and established Principles and Guidelines for Reporting Preclinical Research. These guidelines were rapidly adopted by more than 70 leading biomedical research journals, including the journals of the American Society for Pharmacology and Experimental Therapeutics (ASPET). A statement of support for these guidelines was published in the Society's journals (Vore et al., 2015) along with updated Instructions to Authors (ItA). Additionally, a statistical analysis commentary was simultaneously published in multiple pharmacology

research journals in 2014, including *The Journal of Pharmacology and Experimental Therapeutics*, *British Journal of Pharmacology*, *Pharmacology Research & Perspectives*, and *Naunyn-Schiedeberg's Archives of Pharmacology*, to strengthen best practices in the use of statistics in pharmacological research (Motulsky, 2014b).

In its continuing efforts to improve the robustness and transparency of scientific reporting, ASPET has updated the portion of the ItA regarding data analysis and reporting for three of its primary research journals, *Drug Metabolism and Disposition*, *Journal of Pharmacology and Experimental Therapeutics*, and *Molecular Pharmacology*.

These ItA are aimed at investigators in experimental pharmacology but are applicable to most fields of experimental biology. The new ItA do not tell investigators how to design and execute their studies but instead focus on data analysis and reporting, including statistical analysis. Here, we summarize and explain the changes in the ItA and also include some of our personal recommendations. We wish to emphasize that guidelines are just that, guidelines. Authors, reviewers, and editors should use sound scientific judgment when applying the guidelines to a particular situation.

Explanation of the Guidelines

Guideline: Include Quantitative Indications of Effect Sizes in the Abstract. The revised ItA state that the Abstract should quantify effect size for what the authors deem to be the most important

This commentary is being simultaneously published in *Drug Metabolism and Disposition*, *The Journal of Pharmacology and Experimental Therapeutics*, and *Molecular Pharmacology*.

<https://doi.org/10.1124/dmd.119.090027>.

ABBREVIATIONS: ASPET, American Society for Pharmacology and Experimental Therapeutics; CI, confidence interval; ItA, Instructions to Authors.

Fig. 1. P-hacking refers to a series of analyses in which the goal is not to answer a specific scientific question but rather to find a hypothesis and data analysis method that results in a P value less than 0.05.

quantifiable finding(s) of their study. This can be either a numerical effect (difference, percent change, or ratio) with its 95% confidence interval (CI) or a general description, such as “inhibited by about half,” “almost completely eliminated,” or “approximately tripled.”

It is not sufficient to report only the direction of a difference (e.g., increased or decreased) or whether the difference reached statistical significance. A tripling of a response has different biologic implications than a 10% increase, even if both are statistically significant. It is acceptable (but not necessary) to also include a P value in the Abstract. P values in the absence of indicators of effect sizes should not be reported in the Abstract (or anywhere in the manuscript) because even a very small P value in isolation does not tell us whether an observed effect was large enough to be deemed biologically relevant.

Guideline: State Which Parts (if Any) of the Study Test a Hypothesis According to a Prewritten Protocol and Which Parts Are More Exploratory. The revised ItA state that authors should explicitly say which parts of the study present results collected and analyzed according to a preset plan to test a hypothesis and which were done in a more exploratory manner. The preset plan should include all key aspects of study design (e.g., hypothesis tested, number of groups, intervention, sample size per group), study execution (e.g., randomization and/or blinding), and data analysis (e.g., any normalizing or transforming of data, rules for omitting data, and choice and configuration of the statistical tests). A few of these planning elements are discussed below as part of other recommendations.

All other types of findings should be considered exploratory. This includes multiple possibilities, including the following:

- Analysis of secondary endpoints from a study that was preplanned for its primary endpoint;
- Results from post hoc analysis of previously obtained data; and
- Any findings from experiments in which aspects of design, conduct, or analysis have been adapted after initial data were viewed (e.g., when sample size has been adapted).

A statistically significant finding has a surprisingly high chance of not being true, especially when the prior probability is low (Ioannidis, 2005; Colquhoun, 2014). The more unexpected a new discovery is, the greater the chance that it is untrue, even if the P value is small. Thus, it is much easier to be fooled by results from data explorations than by experiments in which all aspects of design, including sample size and analysis, had been preplanned to test a prewritten hypothesis. Even if the effect is real, the reported effect sizes are likely to be exaggerated.

Because exploratory work can lead to highly innovative insights, exploratory findings are welcome in ASPET journals but must be identified as such. Only transparency about preplanned hypothesis testing versus exploratory experiments allows readers to get a feeling for the prior probability of the data and the likely false-positive rate.

Unfortunately, this planning rule is commonly broken in reports of basic research (Head et al., 2015). Instead, analyses are often done as shown in Fig. 1 in a process referred to as P-hacking (Simmons et al., 2011), an unacceptable form of exploration because it is highly biased. Here, the scientist collects and analyzes some data using a statistical test. If the outcome is not $P < 0.05$ but shows a difference or trend in the hoped-for direction, one or more options are chosen from the following list until a test yields a $P < 0.05$.

- Collect some more data and reanalyze. This will be discussed below.
- Use a different statistical test. When comparing two groups, switch between unpaired t test, the Welch corrected (for unequal variances) t test, or the Mann-Whitney nonparametric test. All will have different P values, and it isn't entirely predictable which will be smaller (Weissgerber et al., 2015). Choosing the test with the smallest P value will introduce bias in the results.
- Switch from a two-sided (also called two-tailed) P value to a one-sided P value, cutting the P value in half (in most cases).
- Remove one or a few outliers and reanalyze. This is discussed in a later section. Although removal of outliers may be appropriate, it introduces bias if this removal is not based on a preplanned and unbiased protocol.
- Transform to logarithms (or reciprocals) and reanalyze.
- Redefine the outcome by normalizing (say, dividing by each animal's weight) or normalizing to a different control and then reanalyze.
- Use a multivariable analysis method that compares one variable while adjusting for differences in another.
- If several outcome variables were measured (such as blood pressure, pulse, and cardiac output), switch to a different outcome and reanalyze.
- If the experiment has two groups that can be designated “control,” switch to the other one or to a combination of the two control groups.
- If there are several independent predictor variables, try fitting a multivariable model that includes different subsets of those variables.
- Separately analyze various subgroups (say male and female animals) and only report the comparison with the smaller P value.

Investigators doing this continue to manipulate the data and analysis until they obtain a statistically significant result or until they run out of money, time, or curiosity. This behavior ignores that the principle goal of science is to find the correct answers to meaningful questions, not to nudge data until the desired answer emerges.

In some cases, investigators don't actually analyze data in multiple ways. Instead, they first look at a summary or graph of the data and then decide which analyses to do. Gelman and Loken (2014) point out that this "garden of forking paths" is a form of P-hacking because alternative analysis steps would have been chosen had the data looked different.

As might be imagined, these unplanned processes artificially select for higher effect sizes and lower P values than would be observed with a preplanned analysis. Therefore, the ItA require that the authors state their analysis methodology in detail so that any P-hacking (or practices that borderline on P-hacking) are disclosed, allowing reviewers and readers to take this into account when evaluating the results. In many cases, it makes sense to remove outliers or to log transform data. These steps (and others) just need to be part of a planned analysis procedure and not be done purely because they lower the P value.

This guideline also ensures that any HARKing (Hypothesizing After the Result is Known Kerr, 1998) is clearly labeled. HARKing occurs when many different hypotheses are tested (say, by using different genotypes or different drugs) and an intriguing relationship is discovered, but only the data supporting the intriguing relationship are reported. It appears that the hypothesis was stated before the data were collected. This is a form of multiple comparisons in which each comparison risks some level of type I error (Berry, 2007). This has been called *double dipping*, as the same data are used to generate a hypothesis and to test it (Kriegeskorte et al., 2009).

Guideline: Report Whether the Sample Size Was Determined before Collecting the Data. The methods section or figure legends should state if the sample size was determined in advance. Only if the sample size was chosen in advance can readers (and reviewers) interpret the results at face value.

It is tempting to first run a small experiment and look at the results. If the effect doesn't cross a threshold (e.g., $P < 0.05$), increase the sample size and analyze the data again. This approach leads to biased results because the experiments wouldn't have been expanded if the results of the first small experiment resulted in a small P value. If the first small experiment had a small P value and the experiment were extended, the P value might have gotten larger. However, this would not have been seen because the first small P value stopped the data collection process. Even if the null hypothesis were true, more than 5% of such experiments would yield $P < 0.05$. The effects reported from these experiments tend to be exaggerated. The results simply cannot be interpreted at face value.

Methods have been developed to adapt the sample size based on the results obtained. The increased versatility in sample size collection results in wider CIs, so larger effects are required to reach statistical significance (Kairalla et al., 2012; <https://www.fda.gov/media/78495/download>). It is fine to use these specialized "sequential" or "adaptive" statistical techniques so long as the protocol was preplanned and the details are reported.

Unlike the *British Journal of Pharmacology* requirements (Curtis et al., 2018), equal sample sizes are not required in the ASPET ItA because there are situations in which it makes sense to plan for unequal sample size (Motulsky and Michel, 2018). It makes sense to plan for unequal sample size when comparing multiple treatments to a single control. The control should have a larger n . It also makes sense to plan for unequal sample size when one of the treatments is much more expensive, time consuming, or risky than the others and therefore should have a smaller sample size than the others. In most cases, it makes sense to have the same sample size in each treatment group.

Situations exist in which a difference in sample sizes between groups was not planned but emerges during an experiment. For instance, an investigator wishes to measure ion currents in freshly isolated cardiomyocytes from two groups of animals. The number of successfully isolated cardiomyocytes suitable for electrophysiological assessment from a given heart may differ for many reasons. It is also possible that some samples from a planned sample size undergo attrition, such as if more animals in the diseased group die than in the control group. This difference in attrition across groups may in itself be a relevant finding.

The following are notes on sample size and power.

- Sample size calculations a priori are helpful, sometimes essential, to those planning a major experiment or to those who evaluate that plan. Is the proposed sample size so small that the results are likely to be ambiguous? If so, the proposed experiment may not be worth the effort, time, or risk. If the experiment uses animals, the ethical implications of sacrificing animals to a study that is unlikely to provide clear results should be considered. Is the sample size so large that it is wasteful? Evaluating the sample size calculations is an essential part of a full review of a major planned experiment. However, once the data are collected, it doesn't really matter how the sample size was decided. The method or justification used to choose sample size won't affect interpretation of the results. Some readers may appreciate seeing the power analyses, but it is not required.
- Some programs compute "post hoc power" or "observed power" from the actual effect (difference) and S.D. observed in the experiment. We discourage reporting of post hoc or observed power because these values can be misleading and do not provide any information that is useful in interpreting the results (Hoenig and Heisey, 2001; <http://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html>; Lenth, 2001; Levine and Ensom, 2001).
- The *British Journal of Pharmacology* requires a minimum of $n = 5$ per treatment group (Curtis et al., 2018). The ItA do not recommend such a minimum. We agree that in most circumstances, a sample size < 5 is insufficient for a robust conclusion. However, we can imagine circumstances in which the differences are large compared with the variability, so smaller sample sizes can provide useful conclusions (<http://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html>). This especially applies when the overall conclusion will be obtained by combining data from a set of different types of experiments, not just one comparison.

Guideline: Provide Details about Experimental and Analytical Methods. The methods section of a scientific article has the following two purposes:

- To allow readers to understand what the authors have done so that the results and conclusions can be evaluated.
- To provide enough information to allow others to repeat the study.

Based on our experience as editors and reviewers, these goals are often not achieved. The revised ItA emphasize the need to provide sufficient detail in the methods section, including methods used to analyze data. In other words, the description of data analysis must be of sufficient granularity that anyone starting with the same raw data and applying the same analysis will get exactly the same results. This includes addressing the following points.

- Which steps were taken to avoid experimental bias, e.g., prespecification, randomization, and/or blinding? If applicable, this can be a statement that such measures were not taken.

Fig. 2. Hypothetical data illustrating how data points may appear as outliers on a linear scale but not after log transformation. The five data tests are all randomly drawn from a lognormal distribution. The left panel uses a linear scale. Some of the points look like outliers. The right panel shows the same data on a logarithmic axis. The distribution is symmetrical, as expected for lognormal data. There are no outliers.

- Have any data points or entire independent experimental replicates (“outliers”) been removed from the analysis, and how was it decided whether a data point or experiment was an outlier? For explanations, see the next section.
- The full name of statistical tests should be stated, for example, “two-tailed, paired *t* test” or “repeated-measures one-way ANOVA with Dunnett’s multiple comparisons tests.” Just stating that a “*t* test” or “ANOVA” has been used is insufficient. When comparing groups, it should be stated whether *P* values are one or two sided (same as one or two tailed). One-sided *P* values should be used rarely and only with justification.
- The name and full version number of software used to perform nontrivial analyses of data should be stated.

We realize that including more details in the methods section can lead to long sections that are difficult to read. To balance the needs of transparency and readability, general statements can be placed in the methods section of the main manuscript, whereas details can be included in an online supplement.

Guideline: Present Details about Whether and How “Bad Experiments” or “Bad Values” Were Removed from Graphs and Analyses. The removal of outliers can be legitimate or even necessary but can also lead to type I errors (false positive) and exaggerated results (Bakker and Wicherts, 2014; Huang et al., 2018).

Before identifying outliers, authors should consider the possibility that the data come from a lognormal distribution, which may make a value look as an outlier on a linear but not on a logarithmic scale. With a lognormal distribution, a few really high values are expected. Deleting those as outliers would lead to misleading results, whereas testing log-transformed values is appropriate (Fig. 2).

If outlier removal is done by “gut feel” rather than preset rules, it can be highly biased. Our brains come up with many apparently good reasons why a value we do not like in the first place should be considered an outlier! Therefore, we recommend that outlier removal should be based on prespecified criteria. If no such rules had been set, a person blinded to group allocation may be less biased.

The choice of the appropriate method for handling apparent outliers depends on the specific circumstances and is up to the investigators. The ItA ask authors to state in the methods or results section what quality control criteria were used to remove “bad experiments” or outliers, whether these criteria were set in advance, and how many bad points or experiments were removed. It may also make sense to report in an online supplement the details on every value or experiment removed as outliers, and to report in that supplement how the results would differ if outliers were not removed.

Guideline: Report Confidence Intervals to Show How Large a Difference (or Ratio) Was and How Precisely It Was Determined.

When showing that a treatment had an effect, it is not enough to summarize the response to control and treatment and to report whether a *P* value was smaller than a predetermined threshold. Instead (or additionally), report the difference or ratio between the means (effect size) and its CI. In some cases, it makes sense to report the treatment effect as a percent change or as a ratio, but a CI should still be provided.

The CI provides a range of possible values for an estimate of some effect size. This has the effect of quantifying the precision of that estimate. Based on the outer limits of the CI, readers can determine whether even these can still be considered biologically relevant. For instance, a novel drug in the treatment of obesity may lower body weight by a mean of 10%. In this field, a reduction of at least 5% is considered biologically relevant by many. But consider two such studies with a different sample size. The smaller study has a 95% CI ranging from a 0.5% reduction to a 19.5% reduction. A 0.5% reduction in weight would not be considered biologically relevant. Because the CI ranges from a trivial effect to a large effect, the results are ambiguous. The 95% CI from the larger study ranges from 8% to 12%. All values in that range are biologically relevant effect sizes, and with such a tight confidence interval, the conclusions from the data are clearer. Both studies have $P < 0.05$ (you can tell because neither 95% confidence interval includes zero) but have different interpretations. Reporting the CI is more informative than just the *P* value. When using a Bayesian approach, report credible intervals rather than CI.

Fig. 3. Variability of P values. If the null hypothesis is true, then the distribution of P values is uniform. Half the P values will be less than 0.50, 5% will be less than 0.05, etc. But what if the null hypothesis is false? The figure shows data randomly sampled from two Gaussian populations with the S.D. equal to 5.0 and populations means that differ by 5.0. Top: three simulated experiments. Bottom: the distribution of P values from 2500 such simulated experiments. Not counting the 2.5% highest and lowest P values, the middle 95% of the P values range from 0.00016 to 0.73, a range covering almost 3.5 orders of magnitude!

Guideline: Report P Values Sparingly. One of the most common problems we observe as statistical reviewers is overuse of P values. Reasons to put less emphasis on P values were reviewed by Greenland et al. (2016) and Motulsky (2018).

- P values are often misunderstood. Part of the confusion is that the question a P value answers seems backward. The P value answers the following question: If the null hypothesis is true (as well as other assumptions), what is the chance that an experiment of the same size would result in an effect (difference, ratio, or percent change) as large as (or larger than) that observed in the completed experiment? Many scientists think (incorrectly) that the P value is the probability that a given result occurred by chance.
- Statisticians don't entirely agree about the best use of P values. The American Statistical Association published a statement about P values introduced by Wasserstein and Lazar (2016) and accompanied it with 21 commentaries in an online supplement. This was not sufficient to resolve the confusion or controversy, so a special issue of *The American Statistician* was published in 2019 with 43 articles and commentaries about P values introduced by Wasserstein et al. (2019).

- P values are based on tentatively assuming a null hypothesis that is typically the opposite of the biologic hypothesis. For instance, when the question would be whether the angiotensin-converting enzyme inhibitor captopril lowers blood pressure in spontaneously hypertensive rats, the null hypothesis would be that it does not. In most pharmacological research, the null hypothesis is often false because most treatments cause at least *some* change to most outcomes, although that effect may be biologically trivial. The relevant question, therefore, is *how big* the difference is. The P value does not address this question. P values say nothing about how important or how large an effect is. With a large sample size, the P value will be tiny even if the effect size is small and biologically irrelevant.
- Even with careful replication of a highly repeatable phenomenon, the P value will vary considerably from experiment to experiment (Fig. 3). P values are "fickle" (Halsey et al., 2015). It is not surprising that random sampling of data leads to different P values in different experiments. However, we and many scientists were surprised to see how much P values vary from experiment to experiment (over a range of more than three

Fig. 4. Comparison of bar graph (mean and S.D.), box and whiskers, scatter plot, and violin plot for a large data set ($n = 1335$). Based on data showing number of micturitions in a group of patients seeking treatment (Amiri et al., 2018). Note that the scale of the y-axis is different for the bar graph than for the other graphs.

orders of magnitude). Note that Fig. 3 is the best case, when all variation from experiment to experiment is due solely to random sampling. If there are additional reasons for experiments to vary (perhaps because of changes in reagents or subtle changes in experimental methods), the P values will vary even more between repeated experiments.

Guideline: Dichotomous Decisions Based on a Single P Value Are Rarely Helpful. Scientists commonly use a P value to make a dichotomous decision (the data are, or are not, consistent with the null hypothesis). There are many problems with doing this.

- Interpretation of experimental pharmacology data often requires combining evidence from different kinds of experiments, so it demands greater nuance than just looking at whether a P value is smaller or larger than a threshold.
- When dichotomizing, many scientists always set the threshold at 0.05 for no reason except tradition. Ideally, the threshold should be set based on the consequences of false-positive and false-negative decisions. Benjamin et al. (2018) suggest that the default threshold should be 0.005 rather than 0.05.
- Sometimes, the goal is to show that two treatments are equivalent. Just showing a large P value is not enough for this purpose. Instead, it is necessary to use special statistical methods designed to test for equivalence or noninferiority. These are routinely applied by clinical pharmacologists in bioequivalence studies.
- Many scientists misinterpret results when the P value is greater than 0.05 (or any prespecified threshold). It is not correct to conclude that the results prove there is “no effect” of the treatment. The corresponding confidence interval quantifies how large the effect is likely to be. With small sample sizes or large variability, the P value could be greater than 0.05, even though the difference could be large enough to be biologically relevant (Amrhein et al., 2019).
- When dichotomizing, many scientists misinterpret results when the P value is less than 0.05 or any prespecified threshold. One misinterpretation is believing that the chance that a conclusion is a false positive is less than 5%. That probability is the false-positive rate. Its value depends on the power of the experiment and the prior probability that the experimental hypothesis was correct, and it is usually much larger than 5%. Because it is hard to estimate the prior probability, it is hard to estimate the false-positive rate. Colquhoun (2019) proposed flipping the problem by computing the prior probability required to achieve a desired false-positive rate. Another misinterpretation is that a P value less than 0.05 means the effect was large enough to be biologically relevant. In fact, a trivial effect can lead to a P value less than 0.05.

In an influential *Nature* paper, Amrhein et al. (2019) proposed that scientists “retire” the entire concept of making conclusions or decisions based on a single P value, but this issue is not addressed in the ASPET ItA.

To summarize, $P < 0.05$ does not mean the effect is true and large, and $P > 0.05$ does not prove the absence of an effect or that two treatments are equivalent. A single statistical analysis cannot inform us about the truth but will always leave some degree of uncertainty. One way of dealing with this is to embrace this uncertainty, express the precision of the parameter estimates, base conclusions on multiple lines of evidence, and realize that these conclusions might be wrong. Definitive conclusions may only evolve over time with multiple lines of evidence coming from multiple sources. Data can be summarized without pretending that a conclusion has been proven.

Guideline: Beware of the Word “Significant.” The word “significant” has two meanings.

- In statistics, it means that a P value is less than a preset threshold (statistical α).
- In plain English, it means “suggestive,” “important,” or “worthy of attention.” In a pharmacological context, this means an observed effect is large enough to have physiologic impact.

Both of these meanings are commonly used in scientific papers, and this accentuates the potential for confusion. As scientific communication should be precise, the ItA suggests not using the term “significant” (Higgs, 2013; Motulsky, 2014a).

If the plain English meaning is intended, it should be replaced with one of many alternative words, such as “important,” “relevant,” “big,” “substantial,” or “extreme.” If the statistical meaning is intended, a better wording is “ $P < 0.05$ ” or “ $P < 0.005$ ” (or any predetermined threshold), which is both shorter and less ambiguous than “significant.” Authors who wish to use the word “significant” with its statistical meaning should always use the phrase “statistically significant.”

Guideline: Avoid Bar Graphs. Because bar graphs only show two values (mean and S.D.), they can be misleading. Very different data distributions (normal, skewed, bimodal or with outliers) can result in the same bar graph (Weissgerber et al., 2015). Figure 4 shows alternatives to bar graphs (see Figs. 7–9 for examples for smaller data sets). The preferred option for small samples is the scatter plot, but this kind of graph can get cluttered with larger sample sizes (Fig. 4). In these cases, violin plots do a great job of showing the spread and distribution of the values. Box-and-whisker plots show more detail than a bar graph but can’t show bimodal distributions. Bar graphs should only be used to show data

expressed as proportions or counts or when showing scatter, violin, or box plots would make the graph too busy. Examples for the latter include continuous data when comparing many groups or showing X versus Y line graphs (e.g., depicting a concentration-response curve). In those cases, showing mean $\pm$ S.D. or median with interquartile range is an acceptable option.

A situation in which bar graphs are not helpful at all is the depiction of results from a paired experiment, e.g., before-after comparisons. In this case, before-after plots are preferred, in which the data points from a single experimental unit are connected by a line so that the pairing effects become clear (see Fig. 8). Alternately, color can be used to highlight individual replicate groups. Authors may consider additionally plotting the set of differences between pairs, perhaps with its mean and CI.

Guideline: Don't Show S.E. Error Bars. The IItA discourage use of S.E. error bars to display variability. The reason is simple; the S.E. quantifies precision, not variability. The S.E. of a mean is computed from the S.D. (S.D. that quantifies variation) and the sample size. S.E. error bars are smaller than S.D., and with large samples, the S.E. is always tiny. Thus, showing S.E. error bars can be misleading, making small differences look more meaningful than they are, particularly with larger sample sizes (Weissgerber et al., 2015) (Fig. 5).

Variability should be quantified by using the S.D. (which can only be easily interpreted by assuming a Gaussian distribution of the underlying population) or the interquartile range.

When figures or tables do not report raw data, but instead report calculated values (differences, ratios, EC_{50} s), it is important to also report how precisely those values have been determined. This should be done with CI rather than S.E. for two reasons. First, CIs can be asymmetric to better show uncertainty of the calculated value. In contrast, reporting a single S.E. cannot show asymmetric uncertainty. Second, although a range extending from the computed value minus one S.E. to that value plus one S.E. is sort of a CI, its exact confidence level depends on sample size. It is better to report a CI with a defined confidence level (usually 95% CI).

Guideline: ASPET Journals Accept Manuscripts Based on the Question They Address and the Quality of the Methods and Not Based on Results. Although some journals preferentially publish studies with a “positive” result, the ASPET journals are committed to publishing papers that answer important questions irrespective of whether a “positive” result has been obtained, so long as the methods are suitable, the sample size was large enough, and all controls gave expected results.

Why not just publish “positive results,” as they are more interesting than negative results?

- Studies with robust design, e.g., those including randomization and blinding, have a much greater chance of finding a “neutral” or “negative” outcome (Sena et al., 2007; MacLeod et al., 2008). ASPET doesn't want to discourage the use of the best methods because they are more likely to lead to “negative” results.
- Even if there is no underlying effect (of the drug or genotype), it is possible that some experiments will end up with statistically significant results by chance. If only these studies, but not the “negative” ones, get published, there is selection for false positives. Scientists often reach conclusions based on multiple studies, either informally or by meta-analysis. If the negative results are not published, scientists will only see the positive results and be misled. This selection of positive results in journals is called “publication bias” (Dalton et al., 2016).
- If there is a real effect, the published studies will exaggerate the effect sizes. The problem is that results will vary from study to study, even with the same underlying situation. When random factors make the effect large, the P value is likely to be <0.05 , so the paper gets published. When random factors make the effect small, the P value will probably be >0.05 , and that paper won't get published. Thus, journals that only publish studies with small P values select for studies in which random factors (as well as actual factors) enhance the effect seen. So, publication bias leads

Fig. 5. Comparison of error bars. Based on Frazier et al. (2006) showing maximum relaxation of rat urinary bladder by norepinephrine in young and old rats; the left panel shows the underlying raw data for comparison as scatter plot.

Fig. 6. How P value selection from underpowered studies and publication bias conspire to overestimate effect size. The simulations draw random data from a Gaussian (normal) distribution. For controls, the theoretical mean is 4.0. For treated, the theoretical mean is 5.0. So, the true difference between population means is 1.0. The S.D. of both populations was set to 1.0 for the simulations in (A) and was set to 0.5 for those in (B). For each simulation, five replicates were randomly drawn for each population, an unpaired t test was run, and both the difference between means and the two-sided P value were tabulated. Each panel shows the results of 1000 simulated experiments. The left half of each panel shows the difference between means for all the simulated experiments. Half the simulated experiments have a difference greater than 1.0 (the simulated population difference), and half have a difference smaller than 1.0. There is more variation in (A) because the S.D. was higher. There are no surprises so far. The right half of each panel shows the differences between means only for the simulated experiments in which $P < 0.05$. In (A), this was 32% of the simulations. In other words, the power was 32%. In (B), there was less experimental scatter (lower S.D.), so the power was higher, and 79% of the simulated experiments had $P < 0.05$. Focus on (A). If the sample means were 4.0 and 5.0 and both sample S.D.s were 1.0 (in other words, if the sample means and S.D.s match the population exactly), the two-sided P value would be 0.1525. P will be less than 0.05 only when random sampling happens to put larger values in the treated group and smaller values in the control group (or random sampling leads to much smaller S.D.s). Therefore, when $P < 0.05$, almost all of the effect sizes (the symbols in the figure) are larger than the true (simulated) effect size (the dotted line at $Y = 1.0$). On average, the observed differences in (A) were 66% larger than the true population value. (B) shows that effect magnification also occurs, but to a lesser extent (11%), in an experimental design with higher power. If only experiments in which $P < 0.05$ (or any threshold) are tabulated or published, the observed effect is likely to be exaggerated, and this exaggeration is likely to be substantial when the power of the experimental design is low.

to overestimation of effect sizes. This is demonstrated by simulations in Fig. 6. Gelman and Carlin (2014) call this a Type M (Magnitude) Error.

- If journals only only accept positive results, they have created an incentive to find positive results, even if P -hacking is needed. If journals accept results that answer relevant questions with rigorous methodology, they have created an incentive to ask important questions and answer them with precision and rigor.

It is quite reasonable for journals to reject manuscripts because the methodology is not adequate or because the controls do not give the expected results. These are not negative data but rather bad data that cannot lead to any valid conclusions.

Examples of How to Present Data and Results

The examples below, all using real data, show how we recommend graphing data and writing the methods, results, and figure legends.

Fig. 7. Unpaired t test example.

Fig. 8. Paired *t* test example.

Example: Unpaired *t* Test

Figure 7 shows a reanalysis of published data (Frazier et al., 2006) comparing maximum relaxation of urinary bladder strips in young and old rats by norepinephrine. The left half of the figure shows the data (and would be sufficient), but the results are more complete with the addition of the right half showing an estimation plot (Ho et al., 2019) showing the difference between means and its 95% CI.

Suggested Wording of Statistical Methods. Relaxation was expressed as percent, with 0% being the force immediately prior to the start of adding norepinephrine and 100% a force of 0 mN. Assuming sampling from a Gaussian distribution, maximum relaxation by norepinephrine in the two groups was compared by an unpaired, two-tailed *t* test.

Results. The mean maximum relaxation provoked by norepinephrine was 24 percentage points smaller (absolute difference) in old rats than in young rats (95% CI: 9% to 38%; $P = 0.0030$).

Figure Legend. In the left panel, each symbol shows data from an individual rat. The lines show group means. The analysis steps had been decided before we looked at the data. The right panel shows the difference between the mean and its 95% CI. The sample sizes were unequal at the beginning of the experiment (because of availability), and the sample sizes did not change during the experiment.

Example: Paired *t* Test

Figure 8 shows a reanalysis of published data (Okeke et al., 2019) on cAMP accumulation in CHO cells stably transfected with human β_3 -adrenoceptors pretreated for 24 hours with 10 μ M isoproterenol

(treated) or vehicle (control) and then rechallenged with freshly added isoproterenol.

Suggested Wording of Statistical Methods. The log $E_{\max}$ of freshly added isoproterenol (as calculated from a full concentration-response curve) was determined in each experiment from cells pretreated with isoproterenol or vehicle. The two sets of $E_{\max}$ values were compared with a two-tailed ratio-paired *t* test (equivalent to a paired *t* test on log of $E_{\max}$).

Suggested Wording of Results

Pretreating with isoproterenol substantially reduced maximum isoproterenol-stimulated cAMP accumulation. The geometric mean of the ratio of $E_{\max}$ values (pretreated with isoproterenol divided by control) was 0.26 (95% confidence interval: 0.16 to 0.43; $P = 0.0002$ in two-tailed ratio-paired *t* test).

Suggested Wording of Figure Legend. In the left panel, each symbol shows the $E_{\max}$ of isoproterenol-stimulated cAMP accumulation from an individual experiment. Data points from the same experiment, pretreated with isoproterenol versus pretreated with vehicle, are connected by a line. Sample size had been set prior to the experiment based on previous experience with this assay. In the right panel, each symbol shows the ratio from an individual experiment. The error bar shows the geometric mean and its 95% confidence interval.

Example: Nonlinear Regression

Figure 9 shows a reanalysis of published data (Michel, 2014) comparing relaxation of rat urinary bladder by isoproterenol upon a 6-hour pretreatment with vehicle or 10 μ M isoproterenol.

Suggested Wording of Statistical Methods. Relaxation within each experiment was expressed as percent, with 0% being the force immediately prior to start of adding isoproterenol and 100% a force of 0 mN. Concentration-response curves were fit to the data of each experiment (where X is the logarithm of concentration) based on the equation

$$Y = \text{bottom} + (\text{top} - \text{bottom}) / (1 + 10^{(\log EC_{50} - X)})$$

to determine top (maximum effect, $E_{\max}$) and $\log EC_{50}$, with bottom constrained to equal zero. Note that this equation does not include a slope factor, which is effectively equal to one. Unweighted nonlinear regression was performed by Prism (v. 8.1; GraphPad, San Diego, CA). Values of $E_{\max}$ and $-\log EC_{50}$ fit to each animal's tissue with pretreatment with vehicle or 10 μ M isoproterenol were compared by unpaired, two-tailed *t* test.

Fig. 9. Nonlinear regression example.

Fig. 10. Multiple comparisons example.

Suggested Wording of Results. Freshly added isoproterenol was similarly potent in bladder strips pretreated with vehicle or 10 μ M isoproterenol but was considerably less effective in the latter. The mean difference in potency ($-\log EC_{50}$) between control and pretreated samples was -0.22 (95% CI: -0.52 to $+0.09$; $P = 0.1519$). The mean absolute difference for efficacy ($E_{\max}$) was 26 percentage points (95% CI: 14% to 37%; $P = 0.0005$).

Suggested Wording of Figure Legend. The left panel shows mean of relaxation from all experiments and a curve fit to the pooled data (for illustration purposes only). The center and right panels show $-\log EC_{50}$ and $E_{\max}$, respectively, for individual experiments. The lines show means of each group. All analysis steps and the sample size of $n = 6$ per group had been decided before we looked at the data. Note that the y-axis is reversed, so bottom (baseline) is at the top of the graph.

Example: Multiple Comparisons

Figure 10 is based on a reanalysis of published data (Michel, 2014) comparing relaxation of rat urinary bladder by isoproterenol upon a 6-hour pretreatment with vehicle or 10 μ M isoproterenol, fenoterol, CL 316,243, or mirabegron.

Suggested Wording of Statistical Methods. Relaxation within each experiment was expressed as percent, with 0% being the force immediately prior to the start of adding isoproterenol and 100% a lack of contractile force (measured tension 0 mN). Concentration-response curves were fit to the data of each experiment (where X is logarithm of concentration) based on the equation

$$Y = \text{bottom} + (\text{top} - \text{bottom}) / (1 + 10^{(\log EC_{50} - X)})$$

to determine top (maximum effect; $E_{\max}$) and $\log EC_{50}$ with bottom constrained to equal zero. Note that this equation does not include a slope factor, which is effectively equal to one. Unweighted nonlinear regression was performed by Prism (v. 8.1; GraphPad). $E_{\max}$ and $-\log EC_{50}$ values from tissue with pretreatment with a β -adrenergic agonist were compared with those from preparations pretreated with vehicle by one-way analysis of variance, followed by Dunnett's multiple comparison test and reporting of multiplicity-adjusted P values and confidence intervals. As sample sizes had been adapted during the experiment, the P values and CIs should be considered descriptive and not as hypothesis-testing.

Suggested Wording of Results. Freshly added isoproterenol was similarly potent in bladder strips pretreated with vehicle or any of the

β -adrenergic agonists. The $E_{\max}$ of freshly added isoproterenol was reduced by pretreatment with isoproterenol (absolute mean difference 22 percentage points; 95% CI: 6% to 39%; $P = 0.0056$) or fenoterol (24% points; 8% to 41%; $P = 0.0025$) but less so by pretreatment with CL 316,243 (11% points; -6% to 28% ; $P = 0.3073$) or mirabegron (14% points; -1% to 29% ; $P = 0.076$).

Suggested Wording of Figure Legend. The left panel shows $E_{\max}$ for individual rats, and the lines show means of each group. The right panel shows the mean difference between treatments with multiplicity-adjusted 95% confidence intervals. All analysis steps had been decided before we looked at the data, but the sample size had been adapted during the course of the experiments.

Summary

The new ItA of ASPET journals do not tell investigators how to design and execute their studies but instead focus on data analysis and reporting, including statistical analysis. Some of the key recommendations are as follows.

- Provide details of how data were analyzed in enough detail so the work can be reproduced. Include details about normalization, transforming, subtracting baselines, etc. as well as statistical analyses.
- Identify whether the study (or which parts of it) was testing a hypothesis in experiments with a prespecified design, which includes sample size and data analysis strategy, or was exploratory.
- Explain whether sample size or number of experiments was determined before any results were obtained or had been adapted thereafter.
- Explain whether statistical analysis, i.e., which specific tests to use and which groups being compared statistically, was determined before any results were obtained or had been adapted thereafter.
- Explain whether any outliers (single data points or entire experiments) were removed from the analysis. If so, state the criteria used and whether the criteria had been defined before any results were obtained.
- Describe variability around the mean or median of a group by reporting S.D. or interquartile range; describe precision, e.g., when reporting effect sizes as CI. S.E. should not be used.
- Use P values sparingly. In most cases, reporting effect sizes (difference, ratio, etc.) with their CIs will be sufficient.
- Make binary decisions based on a P value rarely and define that decision.

- Beware of the word “significant.” It can mean that a P value is less than a preset threshold or that an observed effect is large enough to be biologically relevant. Either avoid the word entirely (our preference) or make sure its meaning is always clear to the reader.
- Create graphs with as much granularity as is reasonable (e.g., scatter plots).

Acknowledgments

Work on data quality in the laboratory of M.C.M. is funded by the European Quality In Preclinical Data (EQIPD) consortium as part of the Innovative Medicines Initiative 2 Joint Undertaking [Grant777364], and this Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industry Associations.

M.C.M. is an employee of the Partnership for Assessment and Accreditation of Scientific Practice (Heidelberg, Germany), an organization offering services related to data quality. T.J.M. declares no conflict of interest in matters related to this content. H.J.M. is founder, chief product officer, and a minority shareholder of GraphPad Software LLC, the creator of the GraphPad Prism statistics and graphing software.

Authorship Contributions

Wrote or contributed to the writing of the manuscript: All authors.

References

- Amiri M, Murgas S, and Michel MC (2018) Do overactive bladder symptoms exhibit a Gaussian distribution? Implications for reporting of clinical trial data. *Neurourol Urodyn* **37** (Suppl 5): S397–S398.
- Amrhein V, Greenland S, and McShane B (2019) Scientists rise up against statistical significance. *Nature* **567**:305–307.
- Bakker M and Wicherts JM (2014) Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychol Methods* **19**:409–427.
- Begley CG and Ellis LM (2012) Drug development: raise standards for preclinical cancer research. *Nature* **483**:531–533.
- Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, et al. (2018) Redefine statistical significance. *Nat Hum Behav* **2**:6–10.
- Berry DA (2007) The difficult and ubiquitous problems of multiplicities. *Pharm Stat* **6**:155–160.
- Collins FS and Tabak LA (2014) Policy: NIH plans to enhance reproducibility. *Nature* **505**: 612–613.
- Colquhoun D (2014) An investigation of the false discovery rate and the misinterpretation of p -values. *R Soc Open Sci* **1**:140216.
- Colquhoun D (2019) The false positive risk: a proposal concerning what to do about p -values. *Am Stat* **73** (Suppl 1):192–201.
- Curtis MJ, Alexander S, Cirino G, Docherty JR, George CH, Giembycz MA, Hoyer D, Insel PA, Izzo AA, Ji Y, et al. (2018) Experimental design and analysis and their reporting II: updated and simplified guidance for authors and peer reviewers. *Br J Pharmacol* **175**: 987–993.
- Dalton JE, Bolen SD, and Mascha EJ (2016) Publication bias: the elephant in the review. *Anesth Analg* **123**:812–813.
- Frazier EP, Schneider T, and Michel MC (2006) Effects of gender, age and hypertension on β -adrenergic receptor function in rat urinary bladder. *Naunyn Schmiedeberg Arch Pharmacol* **373**:300–309.
- Freedman LP and Gibson MC (2015) The impact of preclinical irreproducibility on drug development. *Clin Pharmacol Ther* **97**:16–18.
- Gelman A and Carlin J (2014) Beyond power calculations: assessing type S (Sign) and type M (Magnitude) errors. *Perspect Psychol Sci* **9**:641–651.
- Gelman A and Loken E (2014) The statistical crisis in science. *Am Sci* **102**:460–465.
- Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, and Altman DG (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* **31**:337–350.
- Halsey LG, Curran-Everett D, Vowler SL, and Drummond GB (2015) The fickle P value generates irreproducible results. *Nat Methods* **12**:179–185.
- Head ML, Holman L, Lanfear R, Kahn AT, and Jennions MD (2015) The extent and consequences of p -hacking in science. *PLoS Biol* **13**:e1002106.
- Higgs MD (2013) Macroscopic: do we really need the S -word? *Am Sci* **101**:6–9.
- Ho J, Tumkaya T, Aryal S, Choi H, and Claridge-Chang H (2019) Moving beyond P values: Everyday data analysis with estimation plots. *Nature Methods* **16**:565–566. doi: 10.1038/s41592-019-0470-3.
- Hoening JM and Heisey DM (2001) The abuse of power. *Am Stat* **55**:19–24.
- Huang M-W, Lin W-C, and Tsai C-F (2018) Outlier removal in model-based missing value imputation for medical datasets. *J Healthc Eng* **2018**:1817479.
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* **2**:e124.
- Jarvis MF and Williams M (2016) Irreproducibility in preclinical biomedical research: perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol Sci* **37**:290–302.
- Kairalla JA, Coffey CS, Thomann MA, and Muller KE (2012) Adaptive trial designs: a review of barriers and opportunities. *Trials* **13**:145.
- Kerr NL (1998) HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* **2**: 196–217.
- Kriegeskorte N, Simmons WK, Bellgowan PSF, and Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* **12**:535–540.
- Lenth RV (2001) Some practical guidelines for effective sample size determination. *Am Stat* **55**: 187–193.
- Levine M and Ensom MHH (2001) Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy* **21**:405–409.
- Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, and Donnan GA (2008) Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* **39**:2824–2829.
- Michel MC (2014) Do β -adrenoceptor agonists induce homologous or heterologous desensitization in rat urinary bladder? *Naunyn Schmiedeberg Arch Pharmacol* **387**:215–224.
- Motulsky H (2014a) Opinion: never use the word ‘significant’ in a scientific paper. *Adv Regen Biol* **1**:25155.
- Motulsky H (2018) *Intuitive Biostatistics*, 4th ed, Oxford University Press, Oxford, UK.
- Motulsky HJ (2014b) Common misconceptions about data analysis and statistics. *J Pharmacology and Experimental Therapeutics* **351**:200–205. doi: 10.1124/jpet.114.219170.
- Motulsky HJ and Michel MC (2018) Commentary on the BJP’s new statistical reporting guidelines. *Br J Pharmacol* **175**:3636–3637.
- National Academies of Sciences Engineering and Medicine (2019) *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC.
- Okeke K, Michel-Reher MB, Gravas S, and Michel MC (2019) Desensitization of cAMP accumulation via human β_3 -adrenoceptors expressed in human embryonic kidney cells by full, partial, and biased agonists. *Front Pharmacol* **10**:596.
- Prinz F, Schlange T, and Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* **10**:712–713.
- Sena E, van der Worp HB, Howells D, and Macleod M (2007) How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci* **30**:433–439.
- Simmons JP, Nelson LD, and Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* **22**: 1359–1366.
- Vore M, Abernethy D, Hall R, Jarvis M, Meier K, Morgan E, Neve K, Sibley DR, Traynelis S, Witkin J, et al. (2015) ASPET journals support the National Institutes of Health principles and guidelines for reporting preclinical research. *J Pharmacol Exp Ther* **354**:88–89.
- Wasserstein RL and Lazar NA (2016) The ASA statement on p -values: context, process, and purpose. *Am Stat* **70**:129–133.
- Wasserstein RL, Schirm AL, and Lazar NA (2019) Moving to a world beyond “ $p < 0.05$ ”. *Am Stat* **73**:1–19.
- Weissgerber TL, Milic NM, Winham SJ, and Garovic VD (2015) Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol* **13**:e1002128.

Address correspondence to: Dr. Harvey J. Motulsky, GraphPad Software, 1100 Glendon Ave., 17th floor, Los Angeles, CA 90024. E-mail: hmotulsky@graphpad.com