Eleftheriou *et al. Molecular Autism* (2025) 16:30
https://doi.org/10.1186/s13229-025-00663-3

Molecular Autism

RESEARCH Open Access

Better statistical reporting does not lead to statistical rigour: lessons from two decades of pseudoreplication in mouse-model studies of neurological disorders

Constantinos Eleftheriou[1,2], Sarah Giachetti[1,2†], Raven Hickson[1,2†], Laura Kamnioti-Dumont[1,2,3†], Robert Templaar[1,2], Alina Aaltonen[1,2,4], Eleni Tsoukala[1,2], Nawon Kim[1,2], Lysandra Fryer-Petridis[1,2], Chloe Henley[1,2], Ceren Erdem[1,2], Emma Wilson[1,7], Beatriz Maio[1,2], Jingjing Ye[1,2], Jessica C. Pierce[1,2], Kath Mazur[1,2], Lucia Landa-Navarro[1,2], Nina G. Petrović[1,2], Sarah Bendova[1,2], Hanan Woods[1,2], Manuela Rizzi[1,2], Vanesa Salazar-Sanchez[1,2], Natasha Anstey[1,2], Antonios Asiminas[8], Shinjini Basu[2,5], Sam A. Booker[1,2], Anjanette Harris[1,2], Sam Heyes[1,2], Adam Jackson[2,5], Alex Crocker-Buque[2,5], Aoife C. McMahon[6], Sally M. Till[1,2], Lasani S. Wijetunge[2,5], David JA Wyllie[1,2], Catherine M. Abbott[1,2], Timothy O'Leary[9†] and Peter C. Kind[1,2,5*†]

Abstract

Background Accurately determining the sample size ("N") of a dataset is a key consideration for experimental design. Misidentification of sample size can lead to pseudoreplication, a process of artificially inflating the number of experimental replicates which systematically underestimates variability, overestimates effect sizes and invalidates statistical tests performed on the data. While many journals have adopted stringent requirements with regard to statistical reporting over the last decade, it remains unknown whether such efforts have had a meaningful impact on statistical rigour.

Methods Here, we evaluated the prevalence of this type of statistical error among neuroscience studies involving animal models of Fragile-X Syndrome (FXS) and those using animal models of neurological disorders at large published between 2001 and 2024.

Results We found that pseudoreplication was present in the majority of publication, increasing over time despite marked improvements in statistical reporting over the last decade. This trend generalised beyond the FXS literature to

†Sarah Giachetti, Raven Hickson and Laura Kamnioti-Dumont contributed equally to this work.

†Timothy O'Leary and Peter C. Kind contributed equally to this work.

*Correspondence:
Peter C. Kind
p.kind@ed.ac.uk

Full list of author information is available at the end of the article

rodent studies of neurological disorders at large between 2012 and 2024, suggesting that pseudoreplication remains a widespread issue in the literature.

Limitations The scope of this study was limited to rodent-model studies of neurological disorders which had the potential for being pseudoreplicated, by allowing repeat observations from individual animals. We did not consider reviews or articles whose experimental design could not allow for pseudoreplication, for example studies which reported only behavioural results, or studies which did not use inferential statistics.

Conclusions These observations identify an urgent need for better standards in experimental design and increased vigilance for this type of error during peer review. While reporting standards have significantly improved over the past two decades, this alone has not been enough to curb the prevalence of pseudoreplication. We offer suggestions for how this can be remedied as well as quantifying the severity of this particular type of statistical error. Although the examined literature concerns a specific neuroscience-related area of research, the implications of pseudoreplication apply to all fields of empirical research.

Keywords Autism, Pseudoreplication, Statistics, Animal models, Fragile X

Background

The choice of what constitutes the sample size is a fundamental challenge for experimental design, forming the basis of statistical inference [1]. While ostensibly a simple decision, identifying independent replicates can be a complex task in modern neuroscience datasets. For example, a typical experiment will typically yield a large number of cells per individual animal, cells which themselves are organised in groups by cell type, with data collected frequently spanning multiple days or multiple brain areas. What constitutes independent replicates in this instance can be unclear; is it animals, cells, brain areas, experimental days or entire litters? A misidentification of sample size can lead to pseudoreplication, a form of statistical error which has long been identified as a critical flaw in statistical analysis [2, 3]. Pseudoreplication systematically inflates sample size (or "N") and overstates the power of the experiment by assuming independence in samples which are not independent; for example, cells from the same animal. This often leads to an overestimation of effect size and invalidate statistical tests performed on the data by increasing the likelihood of observing false positive results [2, 4]. An abundance of false positive results in the published literature is thought to drive a general lack of reproducibility in the field [5, 6], with pseudoreplication being a major driver [7–10].

To combat the reproducibility crisis and promote transparency and credibility in science, scientific journals have put more stringent requirements in place for the reporting of statistical details [11, 12]. Intuitively, better reporting of statistical details ought to aid editors and reviewers in catching major statistical flaws in articles prior to publication, and promote the statistical rigour of submitted and published manuscripts [13, 14]. While such measures have been commonplace for more than a decade, it is unclear what impact these measures have had on statistical rigour.

Here, we sought to determine whether improved statistical reporting has led to greater statistical rigour in articles using rodent models of neurological disorders (ND), focusing on pseudoreplication as well-known and common form of statistical error. We reviewed a total of 650 publications using rodent models of neurological disorders between 2001 and 2024, scoring for the presence of pseudoreplication and adequacy of statistical reporting. We found that pseudoreplication has remained a prevalent issue in the literature over the past two decades, present in the majority of publications, despite marked improvement in statistical reporting. We discuss how pseudoreplication affects the false discovery rate of the most common types of statistical testing, and offer suggestions for how this may be improved.

Methods

Article selection & scoring

For the FXS study cohort, a Pubmed search was performed using the gene of interest (fmr1) and the species (mouse) as keywords. Results were limited to articles published in 2001–2012 and 2013–2024. We divided articles into two groups; those published prior to the publication of Nature's reporting guidelines in 2012 (2001–2012) [13], which highlighted the issue of pseudoreplication and called for better reporting, and those published after (2013–2024).

For the ND search, the following pubmed search string was used:

(mice[Title/Abstract] OR mouse [Title/Abstract]) AND (model|Title/Abstract] OR transgenic|Title/Abstract] OR knockout[Title/Abstract]) AND ("The Journal of neuroscience: the official journal of the Society for Neuroscience"[Journal] OR "Neuron" [Journal] OR "Nature neuroscience"[Journal] OR "The Journal of physiology" [Journal] OR "Journal of neurophysiology" [Journal])

Reviews and articles that were not neuroscience-related were excluded, along with articles that were purely behavioural, or which did not, by design, permit repeat-sampling from animal units. Articles with no inferential statistical analyses were also excluded. Articles were sampled in random order until the desirable total number of articles was reached (350 for the FXS literature and 300 for the ND literature).

The results and methods for each article were consulted to assess whether pseudoreplication was evident in at least one figure, and whether adequate statistical details were reported to unequivocally determine this to be the case. An article was considered to contain adequate statistical details if the degrees of freedom for each test performed were explicitly stated, the unit of replication was clearly defined and the statistical tests applied were clearly stated. If adequate statistical details were not reported and there was no obvious evidence of pseudoreplication that could be inferred from the figure legend or main text, articles were given the benefit of the doubt and assumed not to have committed pseudoreplication.

All articles were assessed by two independent scorers for each category, who were blind to each other's assessment. Where discrepancies occurred, these were discussed between all authors to reach a consensus. Where consensus could not be reached, articles were excluded from further analysis and another article was randomly drawn from the literature. This applied to a total of 4 articles.

Data analysis

All data was analysed using custom scripts written in Python v3.11. Jupyter notebooks showing all analysis steps can be found at https://doi.org/10.5281/zenodo.13882178.

Confidence interval estimates (95% CI) of article scores for each year (e.g. Figure 1A) were calculated by bootstrap resampling article scores for pseudoreplication and

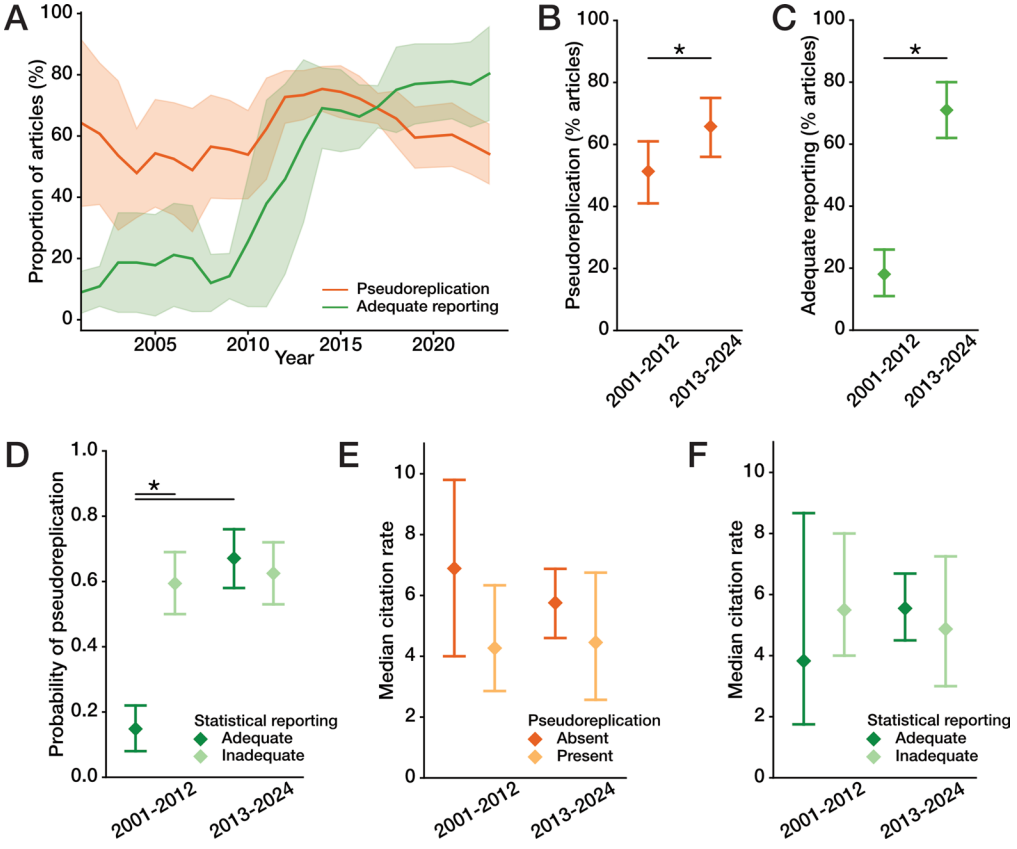

Fig. 1 The prevalence of pseudoreplication in the Fragile-X mouse model literature has remained fairly constant despite marked improvements in statistical reporting. (**A**) Proportion of articles suspected of pseudoreplication in at least one figure (orange line), and proportion reporting adequate statistical details (green line) in articles sampled between 2001 and 2024. Each time-point shows the bootstrap resampled median and 95% percentile interval of the bootstrapped distribution. (**B**) Average percentage of articles suspected of pseudoreplication between 2001–2012 and 2013–2024. Whisker plots show median ± 95% CI of bootstrapped distribution per group. (**C**) Average percentage of articles reporting adequate statistical details between 2001–2012 and 2013–2024. (**D**) Probability of an article being suspected of pseudoreplication when reporting adequate (dark green) or inadequate (light green) statistical details between 2001–2012 and 2013–2024. (**E**) Median citation rate for articles where pseudoreplication was present (dark orange) or absent (light orange) between 2001–2012 and 2013–2024. (**F**) Median citation rate for articles reporting adequate (dark green) or inadequate (light green) statistical details between 2001–2012 and 2013–2024

statistical reporting at each year point. In each bootstrap replicate, articles were resampled 100 times with replacement, and the number of articles suspected of pseudoreplication constituted the bootstrap sample. The process was repeated 10,000 times, and the final distribution was visually inspected to determine whether the bootstrap had converged (i.e. differences were normally distributed as per the central limit theorem [15]). Bootstraps that had not converged were repeated with 100,000 replicates. Confidence intervals were drawn using the percentile method [16], whereby the 2.5th and 97.5th percentiles of the bootstrapped distribution constituted the 95% confidence interval. The median of the bootstrapped distribution was used as the average number of articles which were suspected of pseudoreplication. This process was repeated to estimate confidence intervals for the number of articles which reported adequate statistical details.

For between-group comparisons (e.g. Figure 1B), effect size and significance were estimated by bootstrapping median differences between groups [17, 18]. In each bootstrap replicate, articles from each group were resampled 100 times with replacement, and counts of scores were subtracted between groups. This procedure was repeated 10,000 times to yield a distribution of differences. The distribution was visually inspected to determine whether the bootstrap had converged, and bootstraps that had not converged were repeated with 100,000 replicates. Confidence intervals were drawn using the percentile method [16] and were used to determine whether differences between groups were significant (i.e. whether or not the 95% CI overlapped zero) [18]. The median of the distribution (M_{diff}) was reported as the effect size. P-values were derived from the 95% CI as described by [19].

To estimate the probability of pseudoreplication given adequate or inadequate statistical reporting (e.g. Figure 1D), articles were randomly resampled 100 times with replacement for each level of statistical reporting (i.e. adequate and inadequate). For each level of reporting, the number of articles suspected of pseudoreplication per 100 article samples constituted the probability of pseudoreplication for that level of reporting. The process was repeated 10,000 times to yield a distribution of probabilities, where the 95% percentile interval was used as the 95% CI and the median as the average probability per level of reporting.

Results

How prevalent is pseudoreplication in rodent model studies of neurological disorders? To address this question, we first examined articles published between 2001 and 2024 which utilised mouse models of FXS. FXS is of particular interest to our own research programme and, as one of the most extensively studied monogenic neurological disorders, we anticipated that findings would be representative of the wider literature. In total, we identified 345 articles (150 between 2001 and 2012 and 195 between 2013 and 2024) which reported a quantitative measurement prone to pseudoreplication, such as a biochemical, physiological or anatomical assay. For each article, we assessed whether pseudoreplication was suspected in at least one figure, and whether adequate statistical details were reported in the text to unequivocally determine its presence. We found that pseudoreplication was a widespread phenomenon in the FXS literature, with the majority of publications suspected of committing this type of error in at least one figure (Fig. 1). Interestingly, the proportion of articles suspected of pseudoreplication showed a slight increase over time despite the dramatic increase in adequate statistical reporting observed after 2012 (Fig. 1A). In the period between 2001 and 2012 and 2013–2024, the proportion of articles suspected of pseudoreplication increased by roughly 14% (Fig. 1B; M_{diff} = 14.0 [1.0 28.0] 95% CI, $p = 0.0418$), while the proportion of articles reporting adequate statistical details increased by 53% (Fig. 1C, M_{diff} = 53.0 [41.0 65.0] 95% CI, $p = 5.83e^{-17}$). Overall, these results suggest that pseudoreplication remains a widespread phenomenon despite efforts by the community to improve the standard of statistical reporting in the FXS literature.

Improved statistical reporting is often seen as a panacea for reproducible science [13, 20–24]. The majority of mainstream journals now enforce statistical reporting guidelines (such as [11]) as a means of promoting the statistical rigour and reproducibility of their published articles. As we found that pseudoreplication remained prevalent in spite of marked improvements in statistical reporting over the past two decades, we questioned whether articles reporting adequate statistical details were indeed less prone to pseudoreplication. We calculated the probability of suspecting pseudoreplication in articles that reported adequate or inadequate statistical details published between 2001 and 2012 and 2013–2024 (Fig. 1D). We found that between 2001 and 2012, when adequate statistical reporting was not commonplace, articles which did report adequate statistical details were considerably less likely to pseudoreplicate (P(pseudoreplication | $reporting_{adequate}$) - P(pseudoreplication | $reporting_{inadequate}$); M_{diff} = 0.45 [0.32 0.56] 95% CI, $p = 8.94e^{-13}$). However, between 2013 and 2024, when adequate statistical reporting was more prevalent, articles reporting adequate statistical details were just as likely to pseudoreplicate as those which did not (M_{diff} = -0.05 [-0.18 0.09] 95% CI, $p = 0.741$). The probability of pseudoreplication amongst articles which reported adequate statistical details (i.e. P(pseudoreplication | $reporting_{adequate}$)) was also increased between 2001 and 2012 and 2013–2024

(M_{diff} = 0.53, [0.41 0.64] 95% CI, $p = 2.79e^{-18}$). In contrast, the probability of pseudoreplication amongst articles which did not report adequate statistical details (i.e. P(pseudoreplication | $reporting_{inadequate}$)) remained roughly the same (M_{diff} = 0.03, [-0.11 0.17] 95% CI, $p = 0.688$). Together, the data suggests that improved statistical reporting does not lead to improved statistical practice with respect to pseudoreplication.

We next investigated whether the impact of articles in the FXS field was affected by pseudoreplication or the adequacy of their statistical reporting. Since the share of publications does not necessarily equate to influence or impact, we attempted to quantify impact as the yearly citation rate of an article as reported by the iCite database [25]. We found no evidence that citation rate was affected by pseudoreplication (Figs. 1E and 2001–2012 M_{diff} = -1.44, [-3.75 0.17] 95% CI, $p = 0.847$; 2013–2024 M_{diff} = -1.04, [-2.39 0.28] 95% CI, $p = 0.885$) nor adequacy of statistical reporting (Figs. 1F and 2001–2012 M_{diff} = -1.58, [-3.28 0.11] 95% CI, $p = 0.823$; 2013–2024 M_{diff} = 0.68, [-0.60 2.10] 95% CI, $p = 0.329$). Therefore, pseudoreplication and inadequate statistical reporting are not issues confined to articles that receive little to no attention, but rather are widespread across the entire spectrum of articles in the FXS literature.

Finally, we sought to determine whether our findings generalised beyond the FXS literature to mouse-model studies of neurological disorders at large. We examined articles published between 2001 and 2024 in highly respected neuroscience journals (Neuron, Nature Neuroscience, Journal of Neurophysiology, Journal of Physiology and Journal of Neuroscience) which used mice as models of any neurological disorder (ND). We sampled a total of 300 articles (100 between 2001 and 2012 and 200 between 2013 and 2024) which reported a quantitative measurement prone to pseudoreplication, and assessed whether pseudoreplication was present in at least one figure and whether adequate statistical details were reported in text. We found that pseudoreplication was present in the majority of ND articles, with its prevalence increased between 2001 and 2012 and 2013–2024 (Fig. 2A; M_{diff} = 17.0 [4.0 30.0] 95% CI, $p = 0.0104$) despite a dramatic increase in adequate statistical reporting during that time (Fig. 2B; M_{diff} = 56.0 [44.0 67.0] 95% CI, $p = 3.72e^{-20}$). Similar to the FXS literature, between 2001 and 2012 ND articles which did report adequate statistical details were considerably less likely to pseudoreplicate than articles which did not (Fig. 2C; M_{diff} = 0.39, [0.26 0.52] 95% CI, $p = 8.37e^{-09}$). This trend did not persist in articles published in 2013–2024, with the probability of pseudoreplication amongst articles reporting adequate statistical details showing an increase (M_{diff} = 0.52 [0.40 0.63] 95% CI, $p = 1.12e^{-17}$). Intriguingly, ND articles which did not report adequate statistical details in 2013–2024 were less likely to be suspected of pseudoreplication than articles which did (M_{diff} = -0.22 [-0.35 -0.10] 95% CI, $p = 0.048$). We attribute this finding to the relative degree of leniency we applied when assessing pseudoreplication in articles which did not report adequate statistical details, as these were generally given the benefit of the doubt unless the balance of probability suggested that a figure was pseudoreplicated (for example, by showing more data points in a plot than the number of animals reported in the methods). Lastly, we found that the impact of ND articles, calculated as their annual citation rate [25], was not affected by the presence of pseudoreplication (Figs. 2D and 2001–2012 M_{diff} = 3.63 [-1.11 6.65] 95% CI, $p = 0.104$; 2013–2024 M_{diff} = -0.33 [-2.75 2.28] 95% CI, $p = 0.855$) nor by the adequacy of statistical reporting (Figs. 2E and 2001–2012 M_{diff} = 0.92 [-2.11 3.15] 95% CI, $p = 0.505$; 2013–2024 M_{diff} = 3.51 [-0.93 6.64] 95% CI, $p = 0.158$). Overall, our results for ND articles recapitulate those for FXS, suggesting that the issues we identified generalise beyond the FXS literature.

Discussion

Our results demonstrate that pseudoreplication is a widespread phenomenon in the FXS and wider ND literature, featuring in the majority of articles published in the past two decades despite improvements in statistical reporting ushered by the advent of statistical reporting guidelines enforced by journals. Far from a fringe phenomenon, pseudoreplication is abundant in articles published in highly respected neuroscience journals and spans the entire spectrum of impact and influence in mouse-model studies of neurological disorder. Better statistical reporting alone has done little to curb its prevalence. Rather, the increase in the percentage of articles suspected of pseudoreplication we observed is likely due to the fact that it is easier to detect when more statistical details are provided.

But how much of a problem is pseudoreplication? After all, if an error in statistical analysis has little effect on the conclusions of a study, or does not affect the overall emergence of spurious findings, then perhaps it may not harm reproducibility within the field and its prevalence may be viewed as rather inconsequential [26, 27].

Let's consider the following example, which is fairly representative of the studies we reviewed here. An experiment is designed to determine the effect of deleting a gene on a physiological phenotype measured at cellular resolution, with multiple cells recorded per animal (Fig. 3A). In this instance, individual animals would constitute the appropriate sample size ("N") rather than individual cells, as the properties of cells within a given animal will inevitably show some correlation with each other irrespective of genetic manipulation [28, 29]. Treating within-group samples (cells in this case) as

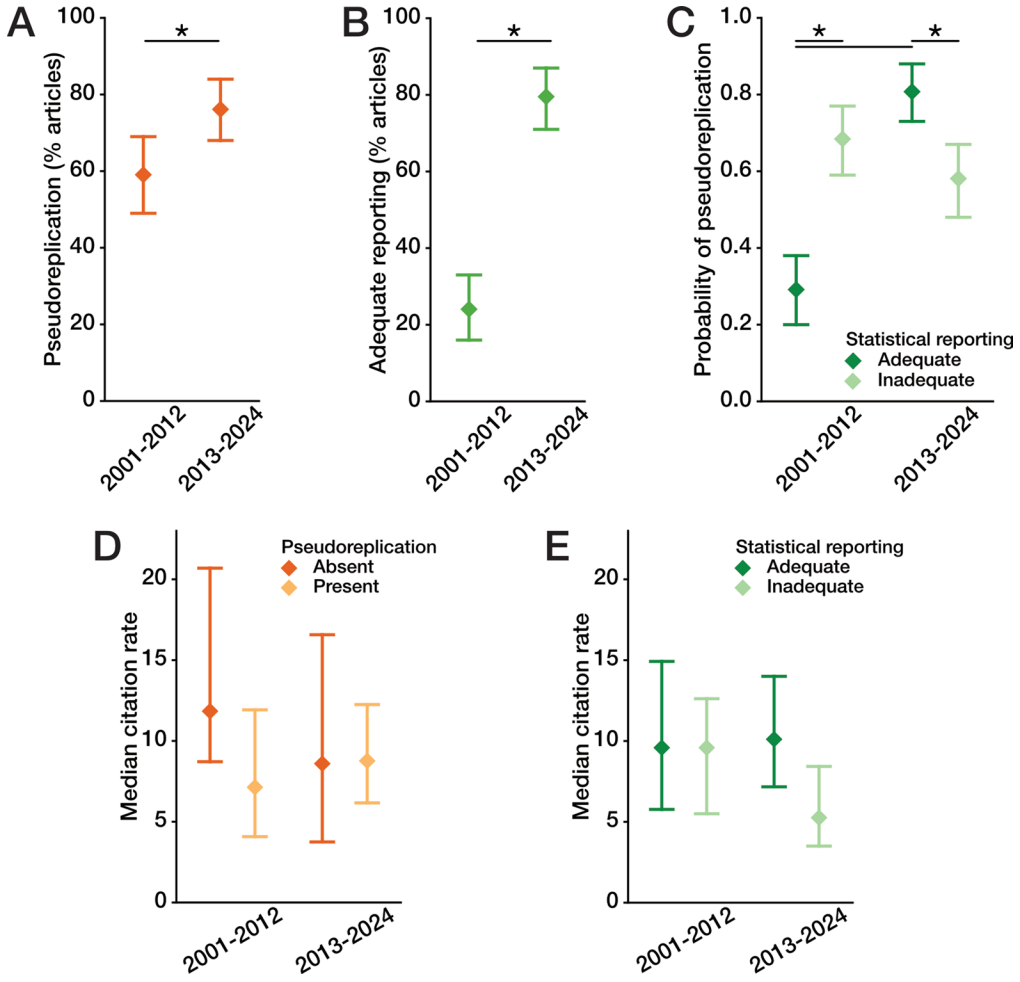

Fig. 2 The prevalence of pseudoreplication across all publications using animal models of neurological disorder has remained high despite improvements in statistical reporting. (**A**) Average percentage of articles suspected of pseudoreplication between 2001–2012 and 2013–2024. Whisker plots show median ± 95% CI of bootstrapped distribution per group. (**B**) Average percentage of articles reporting adequate statistical details between 2001–2012 and 2013–2024. (**C**) Probability of an article being suspected of pseudoreplication when reporting adequate (dark green) or inadequate (light green) statistical details between 2001–2012 and 2013–2024. (**D**) Median citation rate for articles where pseudoreplication was present (dark orange) or absent (light orange) between 2001–2012 and 2013–2024. (**E**) Median citation rate for articles reporting adequate (dark green) or inadequate (light green) statistical details between 2001–2012 and 2013–2024

independent would therefore underestimate variability and constitute pseudoreplication, as sample size would be artificially inflated. The disparity in between-animal and within-animal variance can be quantified using the intraclass correlation coefficient [30, 31], or ρ_{IC} (Fig. 3A). The higher the ρ_{IC}, the higher the discrepancy in between-animal and within-animal variability, meaning multiple samples from a small animals will be less representative of the wider population (Fig. 3B). As the faithful representation of observed variance is a critical consideration for the bulk of statistical tests, including ANOVA or the Student's t-test, misrepresentation of variance can be a critical error for statistical inference [3, 29, 32].

Classical statistical inference, used in the overwhelming majority of published articles, assumes a threshold p-value for determining whether an observation is "statistically significant", with values below the threshold considered grounds for rejecting the null hypothesis. This threshold (α) represents the accepted false positive rate and is typically set to 5% or less for a given test. Pseudoreplication results in artificially low p-values, an effect which becomes more prominent the more the replicate number is inflated raising the false discovery rate beyond the stipulated threshold value [3]. For example, if we were to apply the Student's t-test on a typical dataset comprising of 2 groups, with 3 animals per group and 6 cells from each animal, assuming that the within-animal and between-animal variance is equal (i.e. $\rho_{IC} = 0.5$) the true false discovery would be 33% rather than the stipulated 5% (Fig. 3C). Ostensibly increasing the stringency of our test by lowering the p-value threshold would only modestly diminish the effect of pseudoreplication on false

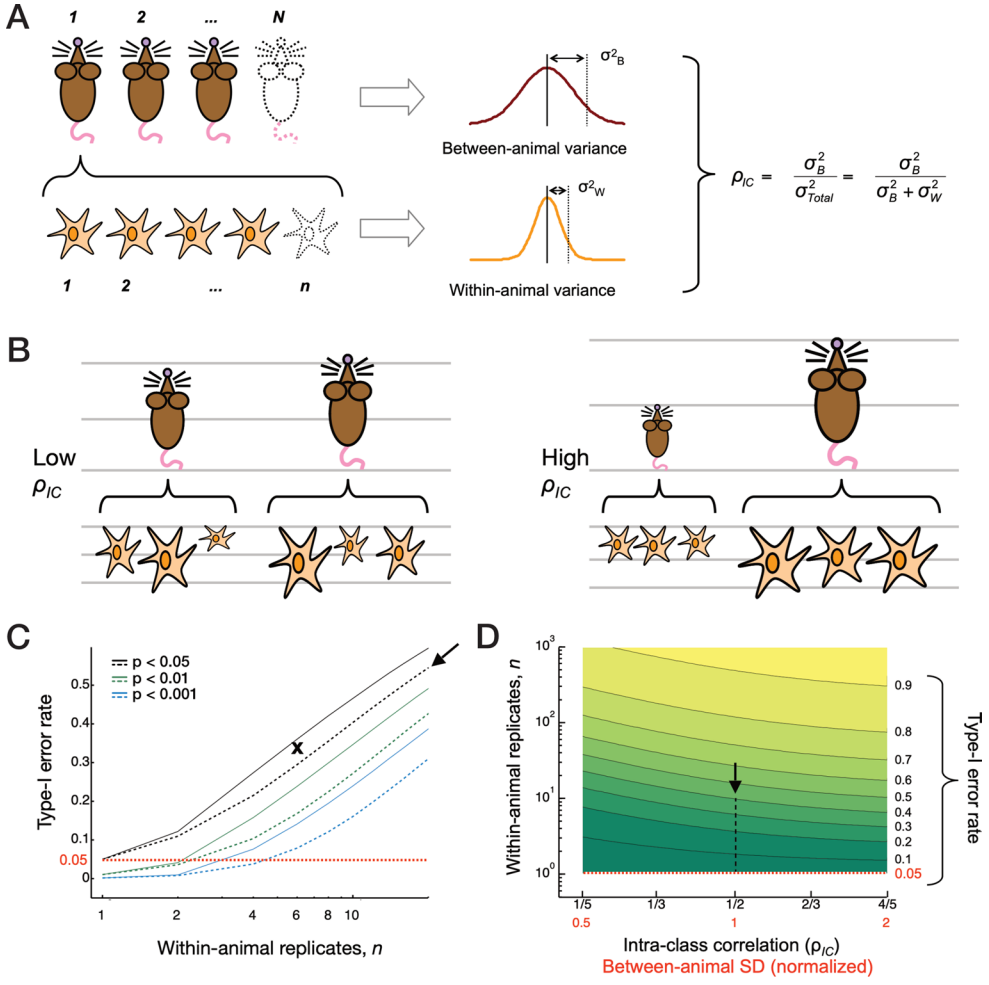

Fig. 3 A typical example of pseudoreplication and its consequences for Student's t-test. (**A**) Many experimental designs use within-animal samples to draw conclusions about the effect of a gene or environmental condition. The total variability in an effect can be split into a within-animal component (between cells, in this case) and a between-animal component. The intra-class correlation coefficient, ρ_{IC}, is a measure of how these sources of variation are related. (**B**) Schematic of variability relationships between and within animals for low (left population) and high (right population) intra-class correlation. Note that animals in the population on the left have high variance between cells (within animal), whereas animals in the population on the right have low cell variance in any given animal. (**C**) Pseudoreplicating by considering within-animal replicates as experimental replicates inflates the true Type-I error rate (false positive rate). **X** indicates the example case given in the text. The curves show how the true Type-I error rate varies with the number of within animal replicates for commonly stipulated significance levels (5%, 1%, 0.1%) and for the possible range of between-animal replicates (solid curves = 2 animals, dotted curves = infinite animals). For all curves, ρ_{IC} is set to 0.5. (**D**) The combined effect of within-animal replicates and intra-class correlation (ρ_{IC}) on the Type-I error rate for a significance threshold of 5% in the presence of pseudoreplication. Between animal standard deviation is shown normalised to within-animal standard deviation for comparison with corresponding values of ρ_{IC}

discovery rate (blue and green traces in Fig. 3C), particularly for higher numbers of pseudoreplicates. Increasing the number of true replicates (i.e. number of animals) would also do little to curb the effect of pseudoreplication; assuming infinite true replicates, false discovery rates would still rise to unacceptable levels as the number of pseudoreplicates increases (dotted traces in Fig. 3C). This discrepancy between stipulated false discovery rate and true false discovery rate dramatically increases as the number of pseudoreplicates per true replicate rises beyond 10. This can be particularly problematic for studies which consider individual synaptic terminals, dendritic spines or cell bodies in counting fields as pseudoreplicates can readily inflate sample size by several orders of magnitude, rendering inferential statistics meaningless [10].

The degree to which pseudoreplication inflates the false discovery rate depends on the degree of correlation within and between replicates, captured by ρ_{IC} [29, 30]. The higher the ρ_{IC}, the worse the consequences of pseudoreplication, as false discovery rates rise beyond the stipulated threshold (Fig. 3D).

While all research approaches can suffer from pseudoreplication, approaches where replicates from a single animal are most possible showed a higher prevalence of pseudoreplication (and where pseudoreplication is

likely to have the greatest effect on the statistical outcome). Because measurements form multiple cells is most common in anatomical studies, these most commonly resulted in pseudoreplication; however, cellular physiological measurements were also commonly pseudoreplicated. In vivo 2-photon imaging and high-density probe electrophysiological recordings were very commonly pseudoreplicated (although, to date, the latter are rarely encountered in NDD research). Indeed, these types of approaches are particularly vulnerable to this type of statistical error because they make measurements from 10s to hundreds of cells. As we note below, measuring from multiple cells from individual animals can increase the power of the experiment if the within and between animal variation is accounted for in a single statical approach. Biochemical and transcriptomic data tended to fare better by virtue of relying on more complex statistical tools (such as mixed-effects models).

How can pseudoreplication be avoided? For classical null hypothesis significance tests (NHSTs), such as t-tests or ANOVAs, the best course of action would most often be pooling within-animal replicates to obtain an average measurement for each animal; for example, averaging all cells recorded from an animal and treating animal means as independent replicates [33]. This relatively blunt approach, however, is not without its shortcomings. Statistical inference on animal averages fails to capture within-animal variability and can obscure effects where they may be present [34], leading to inflated false negative rates [35]. A more desirable - albeit more complex - approach is to model between and within group factors using linear mixed effects models [30], which can model variability between and within animals. While such models are themselves not immune to pseudoreplication [36], they allow for the interrogation of complex relationships in data which would otherwise violate the assumptions of classical NHSTs [37]. Bayesian inference offers another alternative approach to NHSTs, which can be used to make valid predictions about biological entities even if they are pseudoreplicates [38]. Irrespective of the choice of statistic, we encourage authors to interrogate complex hierarchical datasets using interactive computational tools such as "LabAID" [39] as a standard step in their statistical analyses, as such exploration is often sufficient to avoid many common statistical errors including pseudoreplication [40]. Overall, the most reliable way to avoid pseudoreplication is to design experiments that yield adequate sample sizes for quantities of interest, as no statistical approach can ameliorate an underpowered study [5, 9].

Pseudoreplication was identified as an issue nearly four decades ago [2], with many articles since highlighting the dangers it engenders for neuroscience as a field [3, 7, 9, 41, 42]; why does it remain so pervasive to this day? We speculate that part of the issue is an inappropriate emphasis on p-values by authors, reviewers and editors. The widely held belief that a low p-value indicates a "real" result creates a pressure to analyse data in a way that generates low p-values and suppress data that do not fall under an arbitrary significance threshold [43]. Furthermore, this emphasis on p-values often forces authors to adopt NHST approaches where their use is not warranted, for example in observational or preliminary studies which would otherwise be better served by reporting summary statistics. The current culture surrounding scientific publishing offers little scope for such studies to be published, as editors and reviewers will often reject manuscripts that are not deemed to be "hypothesis-driven". This may lead to authors developing post-hoc hypotheses (a process known as HARKing) in order to carry out NHSTs, a practice which is widely considered inappropriate [44]. This is not an issue unique to the use of p-values for determining significance; the use of small confidence intervals to determine significance would be equally problematic. Rather, the emphasis that our community places on producing statistically significant results therefore leads to a use of statistics that is reducing the reproducibility of our science and - by extension - the integrity of our field [10, 45–49].

It would be rather rash to suggest that large portions of the literature in the last twenty years ought to be discarded solely due to the presence of pseudoreplication. Studies that have been reproduced multiple times by different groups, for example, are likely to be genuine; on the other hand, caution is warranted for "one-off" observations based on pseudoreplicated data. Conclusions in studies reporting large effect sizes are also likely to survive correcting for pseudoreplication, while small effect sizes on pseudoreplicated data warrant caution. Studies attempting to address phenotypes mechanistically should certainly not be discounted on the basis of pseudoreplication alone. It is important to note that during the course of our survey, we observed many instances of statistical errors, such as inappropriate use of statistical tests, and basic errors of interpretation which were entirely separate from pseudoreplication. Critical judgement of a paper's merits is therefore required irrespective of the presence or absence of pseudoreplication.

Limitations

The scope of this study was limited to a specific neuroscience-related area of research; while the implications of pseudoreplication apply to all fields of empirical research, we did not attempt to quantify the prevalence of pseudoreplication in articles which did not use animal models of neurological disorders. However, our results are consistent with previous work which reported similarly high

prevalence of pseudoreplication in neuroscience studies at large [3].

Manual scoring of articles for pseudoreplication could be a potential source of bias, as the subjective assessment of an individual reviewer on the presence of pseudoreplication could be coloured by their prior belief on its prevalence. This could particularly be the case for articles which did not report adequate statistical details to definitively determine whether pseudoreplication was present. In this study, we took a number of steps to eliminate bias. Firstly, all articles were scored independently by two reviewers who were blind to each other's result, which highlighted instances where the presence of pseudoreplication was disputed. We also employed a policy of giving articles where the presence of pseudoreplication was disputed on account of inadequate statistical reporting the "benefit of the doubt", which effectively lead to scoring under the assumption that an article was not pseudoreplicated unless there was evidence to suggest otherwise.

Finally, we assessed whether individual reviewers were biased in their scoring for pseudoreplication using the proportion of articles they deemed to be pseudoreplicated; if the percentage of articles an individual reviewer deemed to be pseudoreplicated was above two standard deviations of the overall mean across all reviewers, a reviewer could be considered biased. We did not detect bias in our reviewing using this method.

It is important to note that our estimates on the prevalence of pseudoreplication apply only to articles which had the potential for being pseudoreplicated, i.e. articles which used inferential statistics in experiments whose design permitted repeat sampling from animal units. As articles which did not fit these criteria, such as those reporting only behavioural results or non-experimental articles such as reviews, were not considered, our findings reflect that pseudoreplication is prevalent in the majority of articles which could be pseudoreplicated, rather than the majority of articles across the entire literature.

Conclusions

Our results here have shown that paying lip service to statistical reporting guidelines is not an adequate means of improving statistical rigour. Authors, journal editors and reviewers should be more vigilant about the validity and rigour of their statistical analyses, in addition to ensuring a high standard of statistical reporting. Effectively tackling the lack of reproducibility and statistical rigour will require an open discussion on the inherent variability in biology, and a genuine embrace of negative results. As neuroscience is attempting to tackle some of the most complex questions in contemporary biology, a frank dialogue and an acute awareness of our statistical tools is our only means to discoveries that will stand the test of time.

Abbreviations

CI	Confidence interval
FXS	Fragile-X syndrome
HARKing	Hypothesizing after the results are known
M_{diff}	Bootstrapped median difference
N	Number of replicates or sample size
ND	Neurological disorder
NHST	Null hypothesis significance test
ρ_{IC}	Intra-class correlation coefficient

Acknowledgements

The authors would like to thank the members of the SIDB community and the Patrick Wild Centre for their feedback during the development of this manuscript.

Author contributions

Conceptualisation: D.J.W., T.O'L., and P.C.K. Data collection: C.E., S.G., R.H., L.K-D., R.T., S.Ba., S.Bo., A.H., S.H., A.J., A.C-B., A.C.M., S.M.T., L.S.W., and T.O'L. Article Scoring: C.E., S.G., E.H., L.K-D., R.T., A.A., E.T., N.K., L.F-P., C.H., C.Er., E.W., B.M., J.Y., J.P., K.M., L.L-N., N.P., S.B., H.W., M.R., V.S., N.A., A.As., S.Ba., S.Bo., A.H., S.H., A.J., A.C-B., A.C.M., S.M.T., L.S.W., D.J.W., C.M.A., T.O'L., and P.C.K. Data analysis: C.E. and T.O'L. Writing - draft: C.E., T.O'L and P.C.K. Writing - review and editing: C.E., S.G., R.H., and P.C.K. Supervision: D.J.W., C.M.A., T.O'L., and P.C.K.

Funding

This study was supported by the Simons Initiative for the Developing Brain and the Patrick Wild Centre at the University of Edinburgh.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Peter Kind is an Associate Editor for Molecular Autism. The authors declare no competing interests.

Author details

[1]Simons Initiative for the Developing Brain, University of Edinburgh, Edinburgh, UK
[2]Centre for Discovery Brain Sciences, Deanery of Biomedical Sciences, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK
[3]Scottish Brain Sciences, Edinburgh, UK
[4]Department of Neuroscience, Karolinska Institutet, Stockholm, Sweden
[5]Patrick Wild Centre, University of Edinburgh, Edinburgh, UK
[6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[7]Centre for Clinical Brain Sciences, Deanery of Clinical Sciences, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK
[8]Centre for Translational Neuromedicine, University of Copenhagen, København, Denmark
[9]Department of Engineering, University of Cambridge, Cambridge, UK

Received: 5 December 2024 / Accepted: 2 May 2025
Published online: 26 May 2025

References

1. Gardenier J, Resnik D. The misuse of statistics: concepts, tools, and a research agenda. Account Res. 2002;9(2):65–74.

2. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr. 1984;54(2):187–211.
3. Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? BMC Neurosci. 2010;11(1):5.
4. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings – a practical guide. Biol Rev. 2017;92(4):1941–68.
5. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
6. Billheimer D. Predictive inference and scientific reproducibility. Am Stat. 2019;73(sup1):291–5.
7. Lazic S. Pseudoreplication invalidates the results of many neuroscientific studies. 2009 [cited 2024 Jun 17]; Available from: https://ora.ox.ac.uk/objects/uuid:5f12da48-3001-4972-8c83-fbb28213fce2
8. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179–85.
9. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76.
10. Li BZ, Sumera A, Booker SA, McCullagh EA. Current best practices for analysis of dendritic spine morphology and number in neurodevelopmental disorder research. ACS Chem Neurosci. 2023;14(9):1561.
11. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, Baker M, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. BMC Vet Res. 2020;16(1):242.
12. Making methods clearer. Nat Neurosci. 2013;16(1):1–1.
13. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187–91.
14. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie Du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1(1):0021.
15. Laplace PS. Mémoire Sur les approximations des formules Qui Sont fonctions de Très grands nombres et Sur leur applications aux probabilités. Memoires de l'Academie des Sciences de Paris. 1810.
16. Stine R. An introduction to bootstrap methods: examples and ideas. Sociol Methods Res. 1989;18(2–3):243–91.
17. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Statist [Internet]. 1979 Jan 1 [cited 2022 Aug 12];7(1). Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-1/Bootstrap-Methods-Another-Look-at-the-https://doi.org/10.1214/aos/1176344552.full
18. Mooney CZ, Mooney CF, Mooney CL, Duval RD, Duvall R. Bootstrapping: A nonparametric approach to statistical inference. sage; 1993.
19. Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ. 2011;343(aug08 1):d2304–2304.
20. Gosselin RD. Insufficient transparency of statistical reporting in preclinical research: a scoping review. Sci Rep. 2021;11(1):3335.
21. Baker D, Lidster K, Sottomayor A, Amor S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for Pre-Clinical animal studies. PLoS Biol. 2014;12(1):e1001756.
22. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016;8(341):ps34112–34112.
23. Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg LS, et al. Badges to acknowledge open practices: A simple, Low-Cost, effective method for increasing transparency. PLoS Biol. 2016;14(5):e1002456.
24. Weissgerber TL, Garovic VD, Winham SJ, Milic NM, Prager EM. Transparent reporting for reproducible science. J Neurosci Res. 2016;94(10):859–64.
25. ICite, Hutchins BI, Santangelo G. iCite Database Snapshots (NIH Open Citation Collection) [Internet]. The NIH Figshare Archive; 2022 [cited 2024 Aug 16]. Available from: https://nih.figshare.com/collections/iCite_Database_Snapshots_NIH_Open_Citation_Collection_/4586573
26. Oksanen L. Logic of experiments in ecology: is pseudoreplication a pseudoissue? Oikos. 2001;94(1):27–38.
27. Schank JC, Koehnle TJ. Pseudoreplication is a pseudoproblem. J Comp Psychol. 2009;123(4):421–33.
28. Lazic SE, Clarke-Williams CJ, Munafò MR. What exactly is 'N' in cell culture and animal experiments? PLoS Biol. 2018;16(4):e2005282.
29. Yu Z, Guindani M, Grieco SF, Chen L, Holmes TC, Xu X. Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron. 2022;110(1):21–35.
30. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. 2nd ed. Los Angeles: Sage; 2012. p. 354.
31. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation - A discussion and demonstration of basic features. PLoS ONE. 2019;14(7):e0219854.
32. Boneau CA. The effects of violations of assumptions underlying the t test. Psychol Bull. 1960;57(1):49–64.
33. Pernet C. Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. 2017 Oct 12 [cited 2024 Oct 2]; Available from: https://f1000research.com/articles/4-621
34. n der Goot MH, Kooij M, Stolte S, Baars A, Arndt SS, van Lith HA. Incorporating inter-individual variability in experimental design improves the quality of results of animal experiments. PLoS ONE. 2021;16(8):e0255521.
35. Scariano SM, Davenport JM. The effects of violations of independence assumptions in the One-Way ANOVA. Am Stat. 1987;41(2):123–9.
36. Arnqvist G. Mixed models offer no freedom from degrees of freedom. Trends Ecol Evol. 2020;35(4):329–35.
37. Aarts E, Verhage M, Veenvliet JV, Van Der Dolan CV. A solution to dependency: using multilevel analysis to accommodate nested data. Nat Neurosci. 2014;17(4):491–6.
38. Lazic SE, Mellor JR, Ashby MC, Munafo MR. A bayesian predictive approach for dealing with pseudoreplication. Sci Rep. 2020;10(1):2366.
39. Dando OR, Kozic Z, Booker SA, Hardingham GE, Kind PC. Brain Commun. 2024;6(2):fcae074. Laboratory Automated Interrogation of Data: an interactive web application for visualization of multilevel data from biological experiments.
40. Zuur AF, Ieno EN, Elphick CS. A protocol for data exploration to avoid common statistical problems: data exploration. Methods Ecol Evol. 2010;1(1):3–14.
41. Lazic SE. Genuine replication and pseudoreplication. Nat Rev Methods Primers. 2022;2(1):1–2.
42. Running the numbers. Nat Neurosci. 2005;8(2):123–123.
43. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1(3):140216.
44. Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev. 1998;2(3):196–217.
45. Cohen J. The earth is round (p <.05). American Psychologist. 1994 Dec;49(12):997–1003.
46. Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. Am Stat. 1995;49(1):108–12.
47. Goodman SA, Dirty Dozen. Twelve P-Value misconceptions. Semin Hematol. 2008;45(3):135–40.
48. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25(4):225–30.
49. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond P < 0.05. Am Stat. 2019;73(sup1):1–19.

Publisher's note