Statistical Reporting Must Improve in Preclinical Research: A Call for a Mentality Shift

Romain-Daniel Gosselin[1]
[1]Precision medicine unit, Lausanne University Hospital (CHUV); Chemin des Roches
1a/1b, CH-1010 Lausanne, Switzerland

Abstract
It is well-established that faulty handling of statistics contributes to the reproducibility crisis in preclinical research. The widespread adoption of the myriad of existing guidelines is hard to achieve because researchers often protest against the costs of these reforms, in terms of money, workforce or ethics. However, such justifications are invalid in the case of statistical reporting since more rigorous data presentation would simply require a cultural change with no further resources. Precisely, a quantification of the existing disregard for statistical reporting in the published literature shows that bad habits (such as failure to disclose sample sizes, statistical procedures or software/code) are still ubiquitous and have largely outlived all guidelines. The ACcess to Transparent Statistics (ACTS) call to action is presented, assembling four simple pragmatic measures that are rapidly achievable by journals to enhance the quality of statistical reporting in preclinical research through a global cultural change. The ACTS call to action is a plea for concrete top-down action from publishers of scientific journals, which should spearhead the battle against irreproducibility.

Key Words: Reporting, Presentation, Publishing, Guidelines, Reproducibility, Good practices

1. Introduction

Statistics are intrinsic to quantitative biomedical research and appropriate communication of statistical protocols and results is vital for third-party data interpretation as well as to enable the execution of replication studies. Despite this, there is a widespread indulgent posture toward imperfectly reported statistics in preclinical life science, fueling the so-called reproducibility crisis that has shaken the biomedical community over the past decade. Examples of frequent flaws include partial reporting of statistical methods such as experimental design, exact statistical tests, thresholds or software used as well as culture of providing only p-values in the results with no mention of further information. Major educational efforts aiming at raising the awareness about the importance of statistical presentation in the life science community might help organically create a natural inclination to better reporting. Scores of guidelines have been published in that respect, compiled on the website of the Enhancing the QUAlity and Transparency Of health Research (EQUATOR, https://www.equator-network.org/) network, the leading one in preclinical animal research being the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines[1]. Nevertheless, the numerous editorials and guidelines regularly released with that objective have shown little impact so far[2,3]. Therefore, more coercive enforcement of rigorous reporting standards by scholarly editors might be necessary. I recently proposed the ACcess to Transparent Statistics (ACTS) call to action that recapitulates four concrete and costless changes to become mandatory in the editorial ecosystem[4]: (1) standardize the content of statistical paragraphs; (2) make the statistical subsection the opening paragraph in Methods; (3) insist on a paragraph covering statistical limitations; and (4) allocate resources on reproducibility and null results. Although a better documentation of the most frequent flaws present in the current preclinical publication landscape would help shape the most appropriate concrete corrective actions, especially for items (1) and (3), such systematic quantifications remain sparse.

The present study aims at providing a recent descriptive quantification of flawed statistical reporting in a large sample of preclinical publications. Except for the study of correlation between flaws and impact factor, the objective was not to estimate the true percentage of flaws in the global scientific literature, but rather to quantify flaws in a given simple of journals given that these flaws should be absent if journal quality control is effective. Therefore, the descriptive (non-inferential) approach was chosen. The study focuses on studies that used location tests with a specific attention given to textual reporting of some key elements of methods and results deemed vital for replication. Restrictions to location tests were justified not only by the frequent use of these tests in life science, but also by the need to limit the confounding influence various statistical approaches might have on the quality of reporting. The results indicate that flaws in statistical reporting are ubiquitous.

2. Methods

2.1 Article sampling

A mixed sampling strategy was implemented (Figure 1) to sample journals and articles as follows. First, a selection filter was applied to the Institute for Scientific Information (ISI) Journal Citation Report (https://jcr.clarivate.com) database to generate a complete list of 504 potentially pertinent life science journal titles (Table 1). Next, exclusion criteria were applied to the journal list and 245 periodicals that could not be deemed "preclinical" based on their title or content overview were identified and removed. Then, using a pseudo-random sequence of 20 numbers between 1 and 259 generated using the GraphPad QuickCalc online tool (https://www.graphpad.com/quickcalcs/randMenu), a final shortlist of 20 journals among the 259 preselected ordered by decreasing 2018 Impact Factor were selected. Four additional journals were finally excluded either because they were eventually found to be too clinical or because there was no online access granted to the local institution, leading to a final list of 16 periodicals (Table 2).

Fifteen articles were collected per journal using a convenient sampling methodology. Online contents of each journal were explored, starting from the last issue released in 2019 and browsing backwards to previous issues if needed to reach 15 publications. This time window was decided to prevent any interference of the abundant literature on Coronavirus disease 2019 (Covid-19) published since January 2020, which might show confounding statistical standards. The inclusion and exclusion criteria are presented in Table 3. Studies using human data were acceptable when they used ex-vivo / in-vitro approaches on extracted tissues, cells or samples. From this intermediate list of 240 articles, 17 were finally excluded during the analysis due to previously unnoticed violation of inclusion criteria or congruity with exclusion criteria, giving a final sample of 223 articles included in the study.

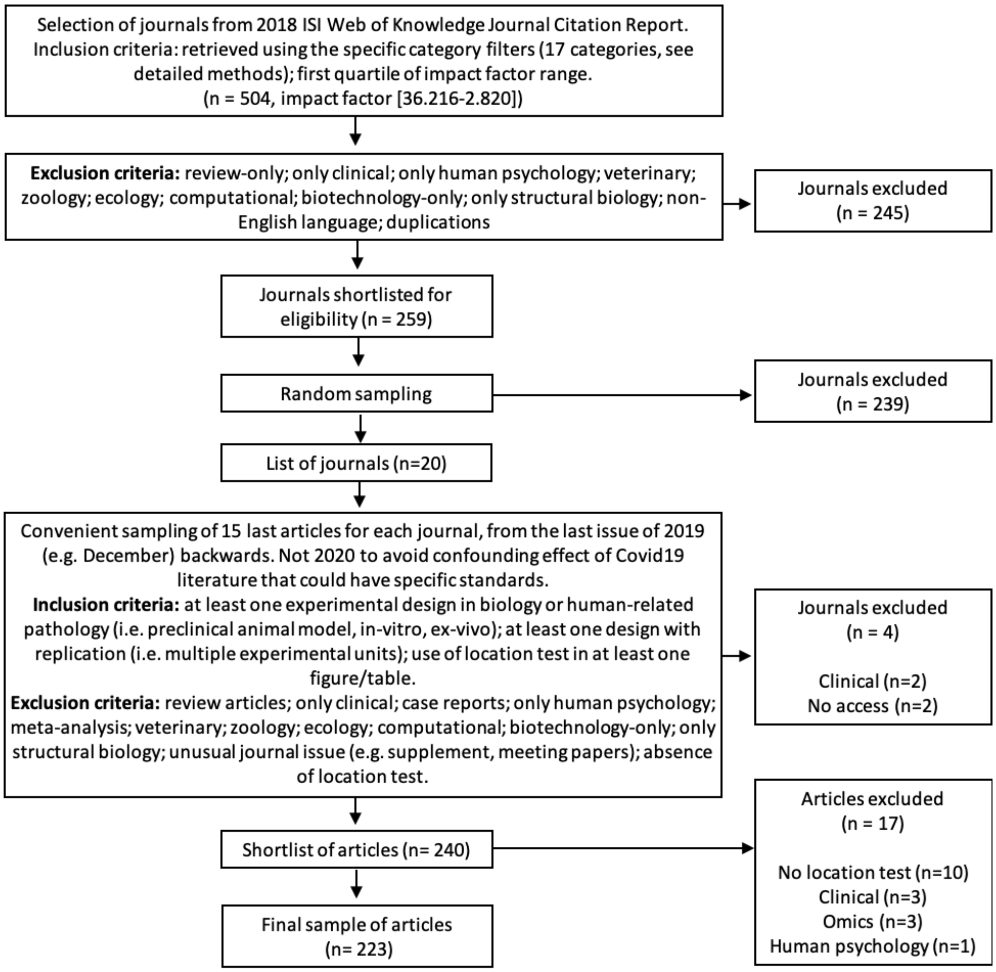
Selection of journals from 2018 ISI Web of Knowledge Journal Citation Report. Inclusion criteria: retrieved using the specific category filters (17 categories, see detailed methods); first quartile of impact factor range. (n = 504, impact factor [36.216-2.820])
Exclusion criteria: review-only; only clinical; only human psychology; veterinary; zoology; ecology; computational; biotechnology-only; only structural biology; non-English language; duplications
Journals excluded (n = 245)
Journals shortlisted for eligibility (n = 259)
Random sampling
Journals excluded (n = 239)
List of journals (n=20)
Convenient sampling of 15 last articles for each journal, from the last issue of 2019 (e.g. December) backwards. Not 2020 to avoid confounding effect of Covid19 literature that could have specific standards.
Inclusion criteria: at least one experimental design in biology or human-related pathology (i.e. preclinical animal model, in-vitro, ex-vivo); at least one design with replication (i.e. multiple experimental units); use of location test in at least one figure/table.
Exclusion criteria: review articles; only clinical; case reports; only human psychology; meta-analysis; veterinary; zoology; ecology; computational; biotechnology-only; only structural biology; unusual journal issue (e.g. supplement, meeting papers); absence of location test.
Journals excluded (n = 4)
Clinical (n=2)
No access (n=2)
Shortlist of articles (n= 240)
Articles excluded (n = 17)
No location test (n=10)
Clinical (n=3)
Omics (n=3)
Human psychology (n=1)
Final sample of articles (n= 223)

Figure 1: Flow chart of the summarized sampling protocol.

Table 1: Selection criteria applied to the ISI JCR database to generate a first list of 504 life science journals and exclusion criteria applied to this list to obtain the final list of journals.

Year	2018
Impact Factor quartile	Q1
Categories (in alphabetical order)	Behavioral sciences; Biomedical research methods; Biochemistry & Molecular biology; Biology; Cell &; Tissue engineering; Cell biology; Developmental biology; Genetics & Heredity; Immunology; Microbiology; Multidisciplinary sciences; Neurosciences; Oncology; Pharmacology & Pharmacy; Physiology; Psychology; Biological; Psychology; Experimental; Toxicology; Virology
Exclusion criteria	Reviews only; Clinical only; Human psychology only; Veterinary; Zoology; Ecology; Computational; Biotechnology only; Structural biology only; Non-English language; Duplications

Table 2: Final list of journals included in the study

Journal name	2018 IF	Number of articles
Science	41.06	12
Cancer Cell	23.92	12
Nature Neuroscience	21.13	13
Science Translational Medicine	17.20	15
Journal of Pineal Research	15.22	15
Biological Psychiatry	11.50	15
Cancer Immunology Research	8.62	14
Cell Death and Differentiation	8.09	14
Neuropsychopharmacology	7.16	14
Science Signaling	6.56	14
Molecular Therapy - Oncolytics	5.71	14
Cellular Oncology	5.02	15
European Journal of Pharmaceutics and Biopharmaceutics	4.71	15
Neuropharmacology	4.37	15
Journal of Virology	4.32	14
Toxins	3.89	12

Table 3: Selection criteria used to sample articles from journal contents.

Period	12.2019 and previous
Inclusion criteria	At least one experimental design in biology or human-related pathology (i.e. preclinical animal model, in-vitro, ex-vivo); at least one design with replication (i.e. multiple experimental units); use of location test in at least one figure/table.
Exclusion criteria	Review articles; Only clinical; Case reports; Only human psychology; Meta-analysis; Veterinary; Zoology; Ecology; Computational; Only big data (genomics, transcriptomics…); Biotechnology-only; Only structural biology; Unusual journal issue (e.g. supplement, meeting papers); Absence of location test.

2.2 Quantification of reporting flaws

Each sampled article abiding by the inclusion and exclusion criteria was explored and three types of statistical attributes, one of primary interest and two of secondary interest, were quantified (Table 4).

Primary binary items relate to the transparency of study protocols and are coded as 0 (presence of all information in the text) or 1 (absence of information in the text for at least one figure or table). These primary binary data were aggregated as proportions of articles that present a flaw (non-disclosure) for the given item. *Secondary quantitative* items monitor the article structure, are given as total counts of given items. *Secondary* *qualitative* items represent the article contents and are summarized as an inventory of information of interest.

Supplemental methods and information were considered fully fledged contents for methodological information (e.g. disclosure of statistical tests, package or sample sizes) but supplementary figures and tables presenting results were not eligible for quantification of their statistical flaws, even if reporting location tests.

Results aggregated for the whole dataset, each article being an experimental unit, voluntarily ignoring the confounding effect of journals. However, for quantifying the correlation between journal impact factor and scores of different items, the results were broken down with "journals" as experimental units and "articles" as sampling units (technical replicates).

2.3 Data collection, analysis and presentation

Data were collected, organized and processed for basic calculation using Microsoft Excel for Mac (version 16). Descriptive statistics (medians, means) and Spearman's rank order correlations were calculated using GraphPad Prism for Mac (version 8, GraphPad Software LLC). GraphPad Prism was used for creating graphs.

Table 4: Description of items used to quantify the reporting flaws in articles.

Type of items	**Description**
Primary binary	Presence of a dedicated statistical paragraph Unambiguous disclosure all statistical tests performed Disclosure of all statistical software used Unambiguous disclosure of all exact sample sizes Absence of contradictory information about methods
Secondary quantitative	Total number of figures and tables Number of figures with location tests
Secondary qualitative	List of all statistical location tests/procedures used List of all statistical software/packages used

3. Results

3.1 Location tests are very frequent and insufficiently described

The analysis of secondary quantitative outcomes is presented in Figure 2. The median number of figure or tables per article in our sample is 6.66 (range [4.58 – 9.08], n=16 journals) and among these, a large proportion display the results of at least one location test (median 79.72%, range [43.22% – 95.41%], n=16 journals). Figure 3 shows the quantification of primary binary outcomes (quality of reporting in figures and tables). The proportions of flawed articles among articles that report location tests differ depending on the items considered. Insufficient disclosure of tests (median 44.76%, range [23.08% – 78.57%], n=16 journals), packages (median 30.95%, range [13.33% – 57.14%], n=16 journals) and exact sample sizes (median 44.17%, range [6.67% – 80.00%], n=16 journals) being particularly frequent. About one fifth (median 18.33%, range [0% – 33.33%], n=16 journals) of articles present contradictory information. Notably, half (8/16) of the sampled journals contain at least one article that fails to display a dedicated statistical paragraph in the method section (median 3.33%, range [0% – 33.33%], n=16 journals).

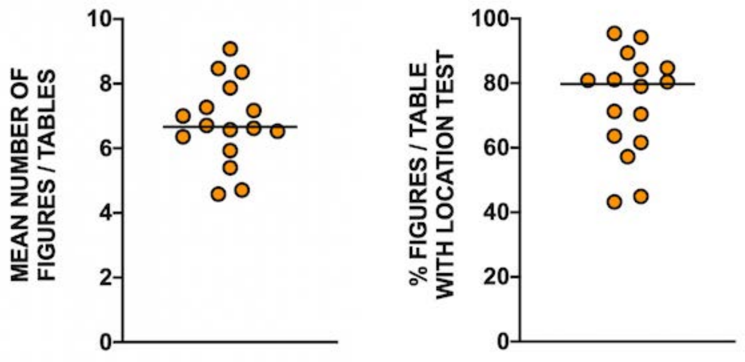

Figure 2: Dot plots showing the mean number of figures/tables per publication (left) and the percentage these with at least a location test (right). Each dot represents one journal, n=16.

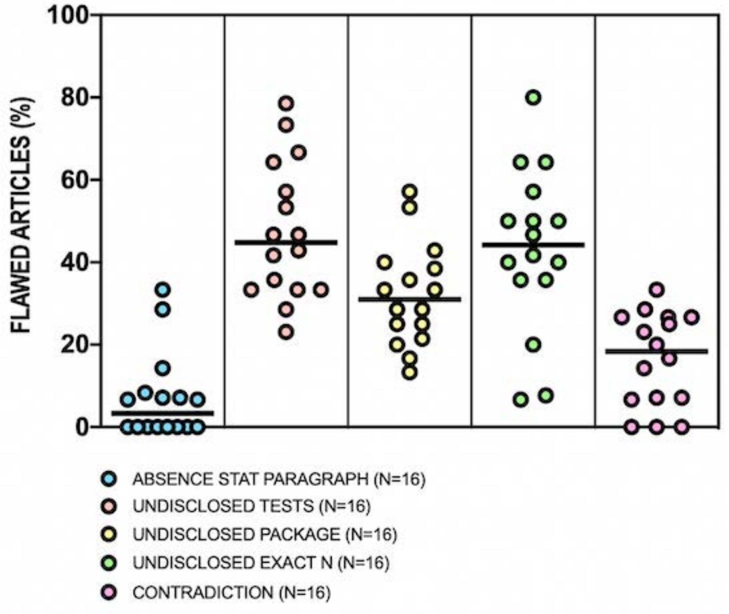

Figure 3: Quantification of primary binary outcomes. Each dot represents one journal, n=16.

3.2 Impact factor is not convincingly correlated with the quality of reporting

Next, the possible relationship between journal impact factor and the frequency flaws in articles was explored (Figure 4). No statistically convincing correlation could be detected for any investigated type of flaw (95% confidence intervals for Spearman r all included 0).

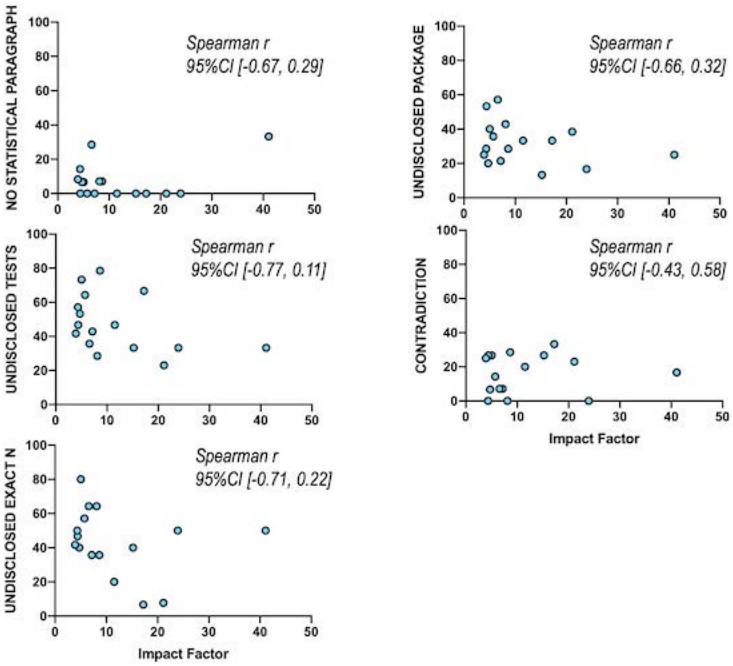

Figure 4: Study of correlation between journal impact factor and percentages on flaws. Each dot represents one journal, n=16.

3.3 Parametric tests are over-represented, especially Student's t test and ANOVA

The analysis of frequency distribution of location tests used in articles (secondary qualitative outcome) is presented in Figure 5. The most frequently used tests were one way analysis of variance (ANOVA, used in 53.15% of articles, n=223 articles), two way ANOVA (28.83%), repeated measure one way ANOVA (9.46%), unpaired Student's t test (38.74%) and Student's t test of undefined laterality (26.83% of articles). Non-parametric tests were less frequently used than parametric methods, with Mann-Whitney test (19.37% of articles) being the most frequently present.

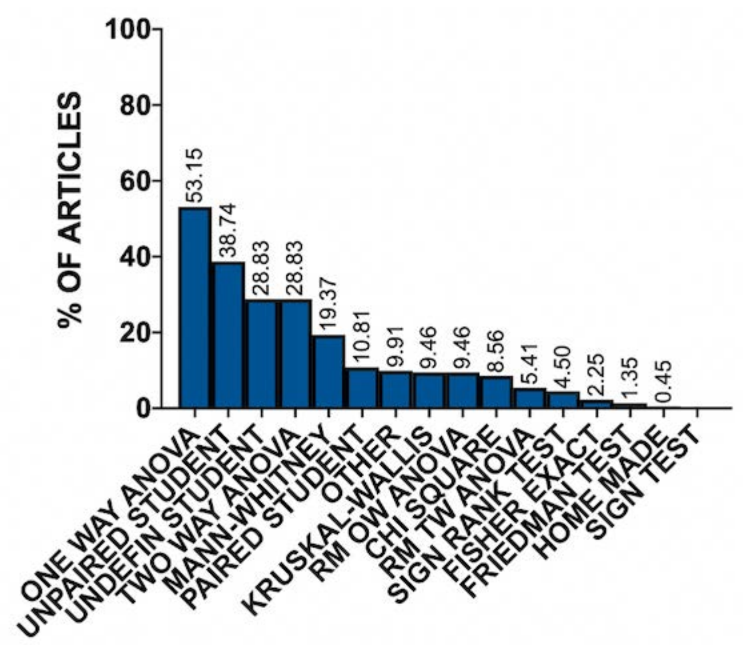

Figure 5: Histogram showing the frequency distribution of location tests in samples articles (n=223 articles).

3.4 Commercial software largely outnumber open source packages

Figure 6 depicts the frequency distribution of statistical packages used in articles (secondary qualitative outcome). The most frequent software used are Prism (mentioned in 59.01% of publications, n=223) and SPSS (16.22%). The only non-proprietary package mentioned in the sampled articles is R (used in 4.50% of articles).

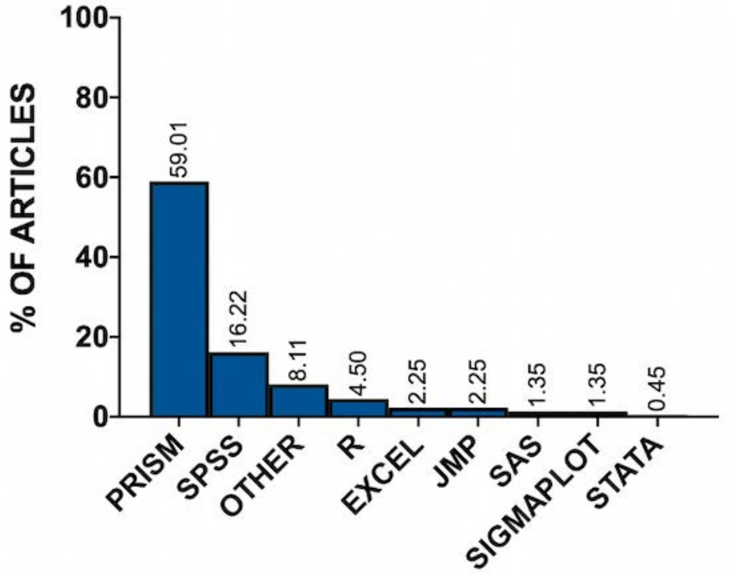

Figure 6: Histogram showing the frequency distribution of statistical packages in samples articles (n=223 articles).

4. Discussion

The broad validity of publications in the preclinical ecosystem are called into question due to a tolerance of faulty statistics. Among the various components of biostatistics that contribute to the insufficiencies in published materials, the poor quality of data/method reporting is considered an important contributor of the observed difficulties of replication attempts[5,6]. I used a rigorous quantification of misuse of statistical reporting to identify the most frequent statistical flaws in 223 articles published in 2019 in 16 journals. I identified the most frequent statistical flaws and hypothesize that they are more likely linked to underlying ignorance and disregard for the importance of good presentation than to deliberate manipulation.

I show here that location tests are highly prevalent but reported with insufficient standards in the preclinical literature, which suggests that the presentation of location tests should become an important target for new measures. The over-representation of ANOVA and Student's t tests not only confirms the entrenched culture of using parametric tests in life sciences, but also points at the importance of educating researchers to the specificities of parametric testing. All of the quantified flaws were present in articles, although at different frequencies. The most frequent flaws were insufficient test disclosure, sample size description and package disclosure, all reaching alarming proportion (median percentage of flawed articles reaching 45%). The low percentage of articles that do not display a dedicated statistical paragraph (median 3.33%, range [0% – 33.33%]) masks the disturbing fact that half of the sampled journals contain at least one publication that falls into that category. Interestingly, the journal impact factor does not seem to be associated with better or worse percentages of flaws articles, regardless the investigated flaw. This prompts for broad educational measures that would target the entire spectrum of scholarly periodicals. Finally, the omnipresence of proprietary software alongside the observed rarity of open-source packages strongly advocates against making mandatory the disclosure of codes used by researchers in the field of preclinical sciences.

I advocate in favor of active changes in mentality in the life science community, by the enforcement (and not mere recommendations) of strict standards in statistical reporting. Many of the pinpointed mistakes in my study, could be efficiently corrected at no significant extra cost. I recently released the ACcess to Transparent Statistics (ACTS) call to action that appeal for the enforcement of four concrete and inexpensive changes in publications: (1) standardize the content of statistical paragraphs; (2) make the statistical subsection the opening paragraph in Methods; (3) insist on a paragraph covering statistical limitations; and (4) allocate resources on reproducibility and null results. In addition, it is crucial that the editorial system takes seriously the importance of the statistical training of peer-reviewers and to encourage the recruitment of statistical reviewers in the peer-review process when needed. Pushing the reflection to its limit, one could even argue that none of the 223 articles sampled in the present study should have contained even one article with the flaws quantified herein, which points at the responsibility of the quality control established by publishers.

The present study has notable limitations. First, the quantification was restricted to a shortlist of items among others that could have been included. For example, the unambiguous reporting of errors, the pervasive and inappropriate usage of standard errors

(SEM), the poor graphical display (with the ubiquitous use of bar graphs) or omnipresence of sole p-values are all features of data/method reporting worth being quantified in future studies. Furthermore, beyond presentation and reporting, insufficiencies in experimental design and statistical analysis are commonplace and notably jeopardize the validity of the preclinical life science literature.

In conclusion, this work provides a rigorous documentation of important flaws in reporting in preclinical sciences. It prompts for both an active enforcement of new simple rules against this misuse of statistics and for similar future studies that quantify other statistical features in the literature.

Acknowledgements

The author would like to thank Prof Penny Reynolds for giving him the opportunity to contribute to this session, and all panelists of the session for their remarkable contribution.

References

1. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol.* 2010;8(6):e1000412.
2. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. *Adv Physiol Educ.* 2007;31(4):295-298.
3. Leung V, Rousseau-Blass F, Beauchamp G, Pang DSJ. ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One.* 2018;13(5):e0197882.
4. Gosselin RD. Statistical Analysis Must Improve to Address the Reproducibility Crisis: The ACcess to Transparent Statistics (ACTS) Call to Action. *Bioessays.* 2020;42(1):e1900189.
5. Landis SC, Amara SG, Asadullah K, et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature.* 2012;490(7419):187-191.
6. Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. Why we need to report more than 'Data were Analyzed by t-tests or ANOVA'. *Elife.* 2018;7.