PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Teaching students to R^{3}eason, not merely to solve problem sets: The role of philosophy and visual data communication in accessible data science education

Ilinca I. Ciubotariu[1,2], **Gundula Bosch**[2]*

1 Department of Biological Sciences, Purdue University, West Lafayette, Indiana, United States of America, **2** Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, R^3 Center for Innovation in Science Education, Baltimore, Maryland, United States of America

* gbosch2@jhu.edu

Citation: Ciubotariu II, Bosch G (2023) Teaching students to R^{3}eason, not merely to solve problem sets: The role of philosophy and visual data communication in accessible data science education. PLoS Comput Biol 19(6): e1011160. https://doi.org/10.1371/journal.pcbi.1011160

Editor: Patricia M. Palagi, SIB Swiss Institute of Bioinformatics, SWITZERLAND

Published: June 8, 2023

Funding: In this work, G.B. was supported in part by the National Institute of Allergies and Infectious Diseases (award number R25AI159447). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Much guidance on statistical training in STEM fields has been focused largely on the undergraduate cohort, with graduate education often being absent from the equation. Training in quantitative methods and reasoning is critical for graduate students in biomedical and science programs to foster reproducible and responsible research practices. We argue that graduate student education should more center around fundamental reasoning and integration skills rather than mainly on listing 1 statistical test method after the other without conveying the bigger context picture or critical argumentation skills that will enable student to improve research integrity through rigorous practice. Herein, we describe the approach we take in a quantitative reasoning course in the R3 program at the Johns Hopkins Bloomberg School of Public Health, with an error-focused lens, based on visualization and communication competencies. Specifically, we take this perspective stemming from the discussed causes of irreproducibility and apply it specifically to the many aspects of good statistical practice in science, ranging from experimental design to data collection and analysis, and conclusions drawn from the data. We also provide tips and guidelines for the implementation and adaptation of our course material to various graduate biomedical and STEM science programs.

This is a *PLOS Computational Biology* Methods paper.

Introduction

In the past decades, there has been a growing acknowledgement of the need to improve the practice of science through better education in research integrity: Responsibility, Rigor, and Reproducibility—the 3 "R" core norms of good science [1–5]. For instance, at the Johns Hopkins Bloomberg School of Public Health, in the R^3 Center for Innovation in Science Education

(R^3ISE), we produce graduate programs and curricula focusing on those core principles by including formal training in critical thinking across research disciplines, ethics in science and society, the nature and sources of mistakes, and responsible science communication [5–7]. Among the tenets of good research conduct fall good statistical practice, which encompasses the full data pipeline from experimental design and conduct to data collection and curation, data analyses, interpretation of results in context, full reporting, and appropriate quantitative and logical understanding [8,9].

Earlier previously published work has emphasized the separation between statistics and biology course work at the undergraduate level [10], and biology students lacking statistical training [11] can lead to statistical errors made by biologists at the professional stage, including statistical errors in high-impact journals [12,13]. There were many calls to change statistical training broadly with approaches like including problem solving [14], incorporating statistics into biology curricula [12,15], and developing data science initiatives for life science education [16]; however, this effort was concentrated at the college level.

At the graduate level, in our experiences [5] and those of others [17,18], there exist learning gaps in statistical training in the sciences, and this is often the first time these students receive advanced training in statistics [19]. For instance, in numerous seminars, committee meetings, and oral exams that we and colleagues attended (Personal Communications, Gundula Bosch), we frequently observed students uncritically operating statistical software, proving incapable of providing sound rationales for the application of a particular method or just making sense of statistical data. Thus, our motivation is to train students who needed to learn the basics of data science in their field to properly reason, not merely to solve problem sets. This has led us to develop a course for biomedical graduate learners, where we adapted our pedagogy approach in quantitative reasoning to include philosophical elements in statistics through exercises in reasoning and error reduction in statistical logic. Theoretically, this course could also be of particular utility to advanced-level undergraduate students interested in STEM careers, although an in-depth knowledge of public health and biomedical sciences is highly encouraged because many of the manuscripts we use as examples and resources in the course require a bit of background knowledge and topic understanding. Nonetheless, the foundational course material could be adapted and taught for other cohorts of interest.

This commentary provides information about our general approach and tips for other educators on how to adapt it to develop their own course modules in a similar fashion (Table 1). Our overall goal is to help improve research integrity through teaching good research practices, while including material centered around errors and fallacies that teach not only technical skills but also aspects of applied logic and ethics training, as well as fundamentals of responsible communication and data visualization.

Curriculum overview

Data science education from errors points of view. The lack of reproducibility of key scientific findings has been a growing concern across many disciplines, in tandem with retractions [20,21]. The reasons behind the reproducibility debate are multifaceted, ranging from insufficient training in rigorous research methods, to sloppy literature outputs, to outright misconduct [22–25]. Additionally, logical fallacies, statistical mistakes during data analysis and interpretation, inadequate reporting of statistical techniques, and erroneous communication are other important contributing factors. For instance, in fields like neuroscience, cell biology, and medicine (but certainly not limited to these disciplines), numerous papers were found to have used inappropriate statistical techniques to compare experimental effects [26], had invalid statistical analyses [27], or contained multiple statistical flaws [28,29].

Table 1. Foundational course materials and general tips for course modules.

Session title	Lecture content overview and error concepts addressed	Topics for activities and discussions
Responsible communication of statistical information along data science life cycle	• Importance of truthful and effective ways to communicating data science information in biomedicine. • Impact of misleading data science communication. • Ethical implications of effective data communication illustrated through examples from John Snow's historic cholera map [33].	• Provide an example of what you consider either effective or poor data communication from your own field of work, thereby using articles, the media, etc. to illustrate your argument (post the graph from the source into your response). Explain what the graph displays or what it should convey. • How would you communicate the main message from the graph to a non-scientist?
The philosophy behind sound hypothesis testing—more than just a statistical method	*p*-values: • The philosophical underpinnings of testable hypotheses and testing methods. Use and misuse of *p*-values in the biomedical data sciences. Importance of other metrics in this context, e.g., statistical power in relation to sample size, effect sizes, and their role in the *p*-value debate.	• Individual reflection: apply the principles from Fisher's work [44] to your research and design an experiment—formulate the appropriate elements of hypothesis testing and describe the experimental approach you would use to answer the proposed question. • Individual or group activity: find an article from your field of interest and critique the responsible use or misuse of the *p*-values, as well as the considerations of statistical power and sample size in the article.
Causal inference research and the role of chi-square tests	• Logical misconceptions in biomedical reasoning—evaluating statistical evidence in causation research.	• Individual problem sets on spurious (but not obvious) correlations [85]: applying the chi-square method to test whether a claim for causation is justifiable. • Group discussion evaluating a claim for causality in an exercise derived from real-word research: applying chi-square statistics to HBV infection data to check if sex and/or age play a causal or correlational role in HBV infection [86]. • Extra discussion: What implications could the statistical fallacy of false causality have, especially in the context of clinical research?
Investigating research (ir)reproducibility	• Explanation of ANOVA testing [87], pitfalls of variance testing [88], variance/ANOVA in the context of reproducibility [39], and applications of variance tests to replicability studies.	• Using examples from "omics" research or population-based vaccination studies [89–91], students learn to describe characteristics of well-defined statistical designs and parametric assumptions that data must satisfy for ANOVA variance tests before proceeding to do calculations. • Additional group discussions circle around identification of poorly selected nonparametric assumptions as well as potential pitfalls of ANOVA variance testing that could be harmful to research and reproducibility?
Big data interpretation—improving how data are visualized and what "data stories" can be accurately told through graphical representations	Errors and fallacies in reasoning using large data sets.	Individual (final) project work: Students: • Sign up for a provided data visualization example (graph) and critique it with respect to legibility, ability to convey a message effectively and clearly, and potential logical and statistical fallacies. • Improve the graph and provide a reasoning for your new visualization design and method. • Record a maximum 3-min "Ted Talk" style talk and present a "before and after" graph example from the above exercise for displaying a big data set. • Explain the context of the graph, thereby justifying their choice of graphical design update.

https://doi.org/10.1371/journal.pcbi.1011160.t001

Hence, the rationale to approach data science education from an error point of view stems from this literature-supported experience that mathematical and statistical mistakes make up for a significant portion of mishaps in scientific practice [30]. In the R3 program across all of our courses, we refer to "errors" in science broadly as mistakes that stem from conceptual misunderstandings, lack of good practice skills in hypothesis testing, data analysis and

management, as well as miscommunication [1]. This approach is broader than the statistical understanding of Type 1 or Type 2 errors, and although it allows us to cover those 2 specific errors in our quantitative reasoning course, we also cover other errors and mistakes in science that may occur in the statistical realm. For instance, through the literature, it is clear that there are a multitude of errors that could occur throughout the entire statistical process from the design of a study to data analysis, to the documentation of the application of statistical methods, to the presentation of study data, and finally to the interpretation of study findings. In other words, this could include a range of errors from failure to use and report randomization, to not providing test assumptions, to using presenting *p*-values as results and in interpretation [29].

Incorporating this lens in our course allows us to focus on science correction and create discussion for scientific advancement [31].

Teaching approach. The format of the course is balanced between short-style lectures, fireside chats with experts in the field to bring some of these statistical concepts to life, and various discussions and activities (see Table 1 for more details). The activities are applications of the lectures in some form, but also have a few specific learning objectives, depending on the topic discussed at hand in a respective week. The students discuss the use of statistical tests and can understand the tests, although we don't place great emphasis on actual calculations. They also become versed in reading the scientific literature with examples of errors and statistical tests in practice and application. Further, we encourage the use of software like R, Tableau, or Excel for preparing the data visualization exercises (since this program is housed in the School of Public Health, we usually have students across departments and at different levels of expertise with programming languages and skills, so we provide some examples through the course material). Below, we describe our material in detail and present the activities and discussions we incorporate in our lessons.

Hypothesis testing and *p*-values. First, we begin with the integration of data science into the scientific life cycle [32] and connect this to effective data visualization and storytelling through graphical representations. We travel through time from John Snow's maps of the 1854 Broad Street Pump cholera outbreak [33] to the present day with examples of SARS-CoV-2 graphical summaries of research that were not the most effective in conveying important data and in turn, public health action [34–37]. We then introduce a hypothesis testing framework, which serves as a guide for all steps in research, from establishing a research question, identifying appropriate statistical tests, interpreting a test statistic outcome, and communicating one's findings.

Hypothesis testing conceived by Neyman and Pearson, in which a researcher considers null and alternative hypotheses to answer a scientific question, can lead to errors like Type 1 and Type 2, when a null hypothesis is rejected when true or when a null hypothesis is not rejected when false, respectively [38]. In our discussion of hypothesis testing, we emphasize the importance of planning an experiment and teach practice applications of common tests like *t* tests, ANOVA, and post hoc tests, thereby focusing on the reasoning aspect behind the applications. We also provide examples from the literature with respect to potential errors that might occur in the context of hypothesis testing (such as using unpaired tests for paired data, using unequal sample sizes for paired *t* tests), thereby emphasizing the "error lens."

Moreover, we highlight the need for clear and sufficient description in publication [39], as this information is key for the transparent reporting element of responsible science communication.

Interwoven into the theory of hypothesis testing from a problem-centric point of view is the *p*-value debate that provides a measure of the strength of evidence against a null hypothesis and which has stirred much controversy due to its improper use [40–42].

"*The p-value was never intended to be a substitute for scientific reasoning.*"–*Ron Wasserstein* [43]

We take a philosophical approach when teaching *p*-values, starting from the depths of Fisher's initial work in which he described their use as an index for measuring the strength of evidence against the null hypothesis [44]. Using an error lens, literature examples help us illustrate instances of less rigorous or improper research practice, and misuse of *p*-values and statistical inference [8,45–47]. We describe the recent state of the field as presented by statisticians through various pieces of scientific literature [8,46,48–52], with the argument that general misuse of statistical methods and *p*-values have played a large part in science's reproducibility crisis [8,47,53]. Further, we discuss the various perspectives [54] centered around the *p*-value debate, including the opinion that *p*-values may not be a useful metric even when used "correctly," the view that *p*-values should be used in tandem with other metrics such as confidence intervals, etc., the suggestion that the "gold-standard" of the *p*-value threshold should be redefined to 0.005 [8,50,55–57]. The *p*-value has been reinterpreted as an "observed error rate" and as measure of evidence; it is not used today as it was initially intended [42,58,59] and it shouldn't represent the be-all and end-all value of studies—although in many cases, it does.

"*Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields.*"–*Jessica Utts* [43]

Null results, publication bias, HARKing, p-hacking, and more. Related to hypothesis testing, we highlight the concept of "null results" or experiments that produce results that do not support a hypothesis made prior to conducting the study [60]. Many times, null findings do not end up published in the scientific literature, which represents publication bias or the trend of journals to prioritize positive findings for publication—one of the 4 horsemen of the reproducibility apocalypse [61–63].

For instance, a study of social science experiments found that majority of null results, in contrast to "strong" results which are usually published in top journals, remain unwritten [60,64]. Also called the "file drawer phenomenon" [65], this can be detrimental to the scientific community, as null results may not be empty of biological meaning and can be as informative as "statistically significant" results that lack replication [66].

This also accentuates the "pressure to publish" that affects the quality of rigorous research and can lead to instances of irreproducibility with practices like *p*-hacking, selective reporting, or data dredging [67]. This refers to performing statistical analyses continuously until nonsignificant results become significant and reporting only significant results [68,69]. Similarly, HARKing, or hypothesizing after results are known [70], is another questionable research practice that we emphasize in our teaching through fireside chats and case studies we have developed with short examples related to biomedical sciences. We also include other issues such as the multiple testing problem or considering a set of statistical inferences simultaneously.

In our course, we provide existing proposals and available solutions to addressing these issues such as the recommendation of "results blind evaluation" of manuscripts [71]. Recently, some journals are moving in the direction of welcoming the publication of negative findings and recognize their value in addressing important questions while being methodologically robust [61]. In another effort to combat this problem, researchers can preregister their studies or write registered reports [72], which allows for preconditioned acceptance after an initial peer review prior to data collection, thus emphasizing study design and sound methodologies rather than "significant" results. Not surprisingly, this has so far been shown to reduce publication bias and increase replication studies [73,74].

We teach that *p*-values cannot be interpreted in isolation, and statisticians have offered suggestions in this direction for many years; for instance, instead of thresholding, context is

needed for interpretation with other robust measures such as sample size, effect sizes, and confidence intervals [51,52,75–77].

Another example of a session in which we relate content material to potential consequences of error or misuse is a discussion of chi-square test for independence, Pearson correlations, and the logical fallacy of false causality. *Cum hoc ergo propter hoc* indicates that tests which are meant to present relationships between variables or correlations should not be used to establish a cause-and-effect relationship, and we ask students to provide examples from the scientific literature.

The overarching debate over the *p*-value and other elements clearly shows that institutional changes are needed, including better training and educational resources that highlight reasoning and error reduction, and this is what our course aims to accomplish [9,53,78].

Responsible communication—Essential to scientific reasoning

As part of the R3 program, we have previously highlighted the importance of rigorous research and presented our framework for responsible scientific communication training as means by which research integrity can be improved [7]. Responsible science communication, which includes value-based principles built on established ethical principles, such as honesty, stewardship, objectivity, openness, accountability, and fairness, encompasses processes from the lab bench to the dissemination of scientific work [7]. By including the philosophical elements of reasoning and error reduction in statistical training, we also realized that now more than ever, visualization is a form of responsible scientific communication; hence, we included it as a connecting thread through our lectures.

Throughout the SARS-CoV-2 pandemic, it has become much clearer that responsible scientific communication is necessary to translate research into public health practice and this includes accurate representation and display of data. Data display is a form of visual communication, and this tool is useful when reaching out to the public and conveying research findings to bring data to life. There have unfortunately been many examples of poor graphical representation of data, ranging from inappropriate graph choice to wrongly scaled axes, and to nondigestible information. We reflect on these in the course and ask students to reconstruct data display in new form through guidelines we cover in lectures.

To this end, one important exercise we highlight throughout the course (see Table 1) is the ability to responsibly communicate data to both one's research colleagues with STEM-focused training and non-scientists. We ask our students to pick specific examples of both effective and ineffective presentation of data from the literature and briefly explain the main message of the selected plots. This exercise, while challenging, is rewarding because it allows the students to reflect on how to improve both their data representation and their communication—parts that are both essential for furthering science and its results.

The themes of responsible science communication and data visualization represent 2 examples of continuity elements which we revisit in multiple course sessions, and we believe highlighting an exercise throughout the course in various forms and contexts prepares the students to profoundly grasp a concept such as hypothesis testing. Another example we review is that of *p*-values, through older studies that have misused this metric and both the current biomedical literature (see Table 1 for details).

Course learning outcomes: Results and reflections

Prior to the implementation of this reformed quantitative reasoning course with an error-focused lens, and with particular emphasis on visualization and communication competencies, we offered a course that was much more focused on rote statistical tests and concepts. As we

collected students' impressions of the 2 courses offered a year apart, we have a comparative "before and after" assessment and can summarize findings from a combination of student evaluation and surveys from the course prior to its new approach and after implementation with the new lens. We can thus provide a better sense of the outcome of students' experiences following the implementation of this didactic approach to a cohort of graduate students in the R3ISE program. Many students claimed that the previous course did not improve their confidence, while based on the course evaluation, the students appreciated this new lens on data visualization, communication, and interpretation, claiming that the course "looked at stats in a different way" and "helped [us] critically assess data [while] providing foundational knowledge of statistical analysis and data visualization". Further, when asked whether the students of this new pedagogical approach felt that they achieved the course objectives as designated in the syllabus, there was improvement (from a mean of 3.3 to 3.9, on a scale of 1 to 4, with 1 representing "pool" and 4 representing "excellent"), and similarly when asked if they gained more expertise in the targeted skills of reasoning and data communication (3.1 increased to 3.9). Additionally, based on students' performances on the various assignments described in Table 1, it was clear that this new approach made the material more engaging and digestible for students, who performed better and were able to answer the questions correctly 90% of the time as opposed to 74%. For this new course, the students indicated that examples were "practical" and "reproducible," which strengthens our argument for good scientific practices and application. Overall, students claimed that the course was "useful" and this was the first time the concepts covered were "used and interpreted in a meaningful way." Further, multiple students "appreciated the depth" of the course, especially the "experience with visualization and analyzing biology-related data." Based on these observations, the strong focus on philosophy and visualization/communication was effective in achieving the main targets of the course and lastly, we copy some direct points from the student survey on achieved objectives:

- *I gained a better outlook on statistical significance and biological relevance.*
- *It greatly improved my quantitative reasoning skills.*
- *[I am] definitely more appreciative of statistics and reasoning now. This course allowed me to understand and truly reflect on the "why" behind each statistical test/term.*
- *[The course] made me think more critically about the tests I can perform on data.*
- *[The course] Made me more wary of data manipulation to fit a hypothesis.*
- *[I] Understand the p-value debate and how context is ultimately important for study results in biomedical science.*
- *[The course] Made me more aware of the different statistical tools available to me and their potential uses and shortcomings.*
- *Now [I] am more confident in my quantitative reasoning abilities and have skillset to use in my upcoming research.*

Surely, we recognize that our program (including this revised course) is still in its young stages, although based in design and structure on established educational theory, and therefore more data collection on program effectiveness is ongoing. We aim to continue collecting data and further understand the impact of such a course on students' performances in acquiring the target competencies and even plan to compare with other graduate training programs based on measurable outcomes.

Furthermore, in addition to making this data publicly available, our goal is also to make our course materials and resources (such as modules, lesson plans, and fireside chats) available for other institutions to adapt to their biomedical science training programs (see our corresponding GitHub repository: https://github.com/JHU-R3ISE).

Conclusions

More work is needed to enhance quantitative and statistical training education for graduate students in the science disciplines [79]. Here, we described and propose our approach to tackling this issue of deficit in quantitative reasoning training by means of a course with an error-focused lens that can be molded and adapted to particular student cohort needs.

> *When training scientists in the use of quantitative methods, I and others often feel pressure to teach the standard approaches that peers and journals expect rather than to expose the problems.–Steven Goodman* [80]

In a sense, we have implemented the familiarly known Bloom's Taxonomy [81,82] for teaching, encompassing major categories of comprehension through our lectures of error concepts, application of learned material to students' specific research contexts, analysis through abstracted case examples, evaluation by means of critiquing provided cases, and creation through a final project bringing together all elements in a Ted-style talk. We provide tips for planning and implementation for other educators so such a course can be integrated into graduate science curricula.

Real-world data sets and cases of problems can be easily extracted from resources like Retraction Watch (retractionwatch.com) or other publicly available databases as the basis for case scenarios with problems and discussion questions (Table 1).

We recognize that other trainers in statistics may have described similar approaches in the past, especially with regards to hypothesis testing and *p*-values; for instance, books such as *Common Errors in Statistics (and How to Avoid Them)* [83] and *Statistics Done Wrong*: *The Woefully Complete Guide* [84] address popular mistakes in data collection, discuss statistical blunders in modern science, and provide information on appropriate statistical technique, and are thus excellent resources for our class. To our knowledge, there has not been any published comprehensive course approach, and guide format such as this one, and so we aim for such a course to prevent gaps in quantitative reasoning and enable our students to feel prepared in their understanding of statistical test applications and visualization and communication, which we argue are key elements of responsible science communication. The specific focuses on the philosophical bases of elements like hypothesis testing, in combination with data visualization and responsible communication which we implement in our course, is what sets this work apart from the traditional approach to teaching data science material. We believe this work could be taken as a foundational approach and adapted to specific departments for teaching data science across STEM-focused disciplines to lab scientists, public health practitioners, etc.

Acknowledgments

We are very appreciative of the suggestions of colleagues and collaborators in the expanding R^3ISE network.

Author Contributions

Conceptualization: Ilinca I. Ciubotariu, Gundula Bosch.

Funding acquisition: Ilinca I. Ciubotariu, Gundula Bosch.

Investigation: Ilinca I. Ciubotariu.

Resources: Gundula Bosch.

Supervision: Gundula Bosch.

Writing – original draft: Ilinca I. Ciubotariu.

Writing – review & editing: Ilinca I. Ciubotariu, Gundula Bosch.

References

1. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. Nat Hum Behav. 2017; 1(1):0021. https://doi.org/10.1038/s41562-016-0021 PMID: 33954258
2. Lorsch J, Gammie A, Singh S. Catalyzing the Modernization of Graduate Education. Microbe Magazine. 2016; 11:96–97.
3. Leshner AI. Rethinking graduate education. Science. 2015; 349(6246):349. https://doi.org/10.1126/science.aac9592 PMID: 26206902
4. Yamamoto KBB, Cech T, Charo RA, Fishman M, Horvitz HR, Hyman S. 2016. A vision and pathway for the NIH. 2016. National Institutes of Health, Bethesda, MD.
5. Bosch G, Casadevall A. Graduate Biomedical Science Education Needs a New Philosophy. mBio. 2017; 8(6):E01539–E01517. https://doi.org/10.1128/mBio.01539-17 PMID: 29259084
6. Bosch G. Train PhD students to be thinkers not just specialists. Nature. 2018; 554:277–277.
7. Ciubotariu II, Bosch G. Improving research integrity: a framework for responsible science communication. BMC Res Notes. 2022; 15(1):177. https://doi.org/10.1186/s13104-022-06065-5 PMID: 35570294
8. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016; 70(2):129–133.
9. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015; 520(7549):612–612. https://doi.org/10.1038/520612a PMID: 25925460
10. A'Brook R, Weyers JDB. Teaching of statistics to UK undergraduate biology students in 1995. J Biol Educ. 1996; 30(4):281–288.
11. Horgan GW DA Elston, Franklin MF, Glasbey CA, Hunter EA, Talbot M, et al. Teaching Statistics to Biological Research Scientists. J R Stat Soc. 1999; 48(3):393–400.
12. Metz AM. Teaching statistics in biology: using inquiry-based learning to strengthen understanding of statistical analysis in biology laboratory courses. CBE Life Sci Educ. 2008; 7(3):317–326. https://doi.org/10.1187/cbe.07-07-0046 PMID: 18765754
13. Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AHS. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC Res Notes. 2011; 4(1):304. https://doi.org/10.1186/1756-0500-4-304 PMID: 21854631
14. Stuart M. Changing the Teaching of Statistics. J R Stat Soc. 1995; 44(1):45–54.
15. Colon-Berlingeri M, Burrowes PA. Teaching Biology through Statistics: Application of Statistical Methods in Genetics and Zoology Courses. CBE—Life Sciences. Education. 2011; 10(3):259–267. https://doi.org/10.1187/cbe.10-11-0137 PMID: 21885822
16. Robeva RS, Jungck JR, Gross LJ. Changing the Nature of Quantitative Biology Education: Data Science as a Driver. Bull Math Biol. 2020; 82(10):127. https://doi.org/10.1007/s11538-020-00785-0 PMID: 32951075
17. Serão NVL, Petry AL, Sanglard LP, Rossoni-Serão MC, Bundy JM. Assessing the statistical training in animal science graduate programs in the United States: survey on statistical training. J Anim Sci. 2021; 99(5). https://doi.org/10.1093/jas/skab086 PMID: 33738494
18. Oster RA, Devick KL, Thurston SW, Larson JJ, Welty LJ, Nietert PJ, et al. Learning gaps among statistical competencies for clinical and translational science learners. J Clin Transl Sci. 2021; 5(1):e12.
19. Zitomer RA, Karr J, Kerstens M, Perry L, Ruth K, Adrean L, et al. Ten simple rules for getting started with statistics in graduate school. PLoS Comput Biol. 2022; 18(4):e1010033. https://doi.org/10.1371/journal.pcbi.1010033 PMID: 35446846
20. Casadevall A, Fang FC. Reproducible science. Infect Immun. 2010; 78(12):4972–4975. https://doi.org/10.1128/IAI.00908-10 PMID: 20876290

21. Fang FC, Casadevall A. Retracted science and the retraction index. Infect Immun. 2011; 79(10):3855–3859. https://doi.org/10.1128/IAI.05661-11 PMID: 21825063

22. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2(8):e124–e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

23. Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications. Proc Natl Acad Sci U S A. 2012; 109(42):17028–17033. https://doi.org/10.1073/pnas.1212247109 PMID: 23027971

24. Flier JS. Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy. Mol Metab. 2016; 6(1):2–9. https://doi.org/10.1016/j.molmet.2016.11.006 PMID: 28123930

25. Munafò MR, Chambers C, Collins A, Fortunato L, Macleod M. The reproducibility debate is an opportunity, not a crisis. BMC Res Notes. 2022; 15(1):43. https://doi.org/10.1186/s13104-022-05942-3 PMID: 35144667

26. Nieuwenhuis S, Forstmann BU, Wagenmakers E-J. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci. 2011; 14(9):1105–1107. https://doi.org/10.1038/nn.2886 PMID: 21878926

27. Lazic SE, Clarke-Williams CJ, Munafò MR. What exactly is 'N' in cell culture and animal experiments? PLoS Biol. 2018; 16(4):e2005282. https://doi.org/10.1371/journal.pbio.2005282 PMID: 29617358

28. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med. 1987; 317(7):426–432. https://doi.org/10.1056/NEJM198708133170706 PMID: 3614286

29. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research—a review of common pitfalls. Swiss Med Wkly. 2007; 137(3–4):44–49. https://doi.org/10.4414/smw.2007.11587 PMID: 17299669

30. Casadevall A, Steen RG, Fang FC. Sources of error in the retracted scientific literature. Faseb J. 2014; 28(9):3847–3855. https://doi.org/10.1096/fj.14-256735 PMID: 24928194

31. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: Errors, underlying themes, and potential solutions. Proc Natl Acad Sci U S A. 2018; 115(11):2563–2570. https://doi.org/10.1073/pnas.1708279115 PMID: 29531079

32. Ezer D, Whitaker K. Data science for the scientific life cycle. eLife. 2019; 8:e43979. https://doi.org/10.7554/eLife.43979 PMID: 30839275

33. Snow J. On the mode of communication of cholera. 1855.

34. CDC. COVIDView Summary ending on April 18, 2020. 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/04242020.html.

35. Federalist. These 12 Graphs Show Mask Mandates Do Nothing To Stop COVID. 2020. Available from: https://thefederalist.com/2020/10/29/these-12-graphs-show-mask-mandates-do-nothing-to-stop-covid/.

36. Foley KE. How bad Covid-19 data visualizations mislead the public. 2020. Available from: https://qz.com/1872980/how-bad-covid-19-data-visualizations-mislead-the-public.

37. Lee C., Yang T, Inchoco G, Jones GM, Satyanarayan A. The Data Visualizations Behind COVID-19 Skepticism. 2021. Available from: http://vis.mit.edu/covid-story/.

38. Neyman J, Pearson ES, Pearson K. IX. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A. 1933; 231(694–706):289–337.

39. 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:e36163. https://doi.org/10.7554/eLife.36163 PMID: 30574870

40. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol. 2008; 45(3):135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003 PMID: 18582619

41. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med. 1999; 130(12):995–1004.

42. Goodman SN. p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate. Am J Epidemiol. 1993; 137(5):485–496. https://doi.org/10.1093/oxfordjournals.aje.a116700 PMID: 8465801

43. ASA. American Statistical Association releases statement on statistical significance and p-values. 2016. Available from: https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf.

44. Fisher RA. Statistical Methods for Research Workers, in Breakthroughs in Statistics: Methodology and Distribution. Kotz S, Johnson NL, editors. New York: Springer; 1992. p. 66–70.

45. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens. 2011; 24(1):18–23. https://doi.org/10.1038/ajh.2010.205 PMID: 20966898

46. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016; 531(7593):151–151. https://doi.org/10.1038/nature.2016.19503 PMID: 26961635

47. Peng R. The reproducibility crisis in science: A statistical counterattack. Significance. 2015; 12(3):30–32.

48. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015; 12(3):179–185. https://doi.org/10.1038/nmeth.3288 PMID: 25719825

49. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. R Soc Open Sci. 2017; 4(12):171085. https://doi.org/10.1098/rsos.171085 PMID: 29308247

50. Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature. 2019; 567:305–307.

51. Halsey LG. The reign of the p value is over: what alternative analyses could we employ to fill the power vacuum? Biol Lett. 2019; 15(5):20190174. https://doi.org/10.1098/rsbl.2019.0174 PMID: 31113309

52. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon Statistical Significance. Am Stat. 2019; 73(sup1):235–245.

53. Goodman SN. Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. Am Stat. 2019; 73(sup1):26–30.

54. Di Leo G, Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur Radiol Exp. 2020; 4(1):18. https://doi.org/10.1186/s41747-020-0145-y PMID: 32157489

55. Hurlbert SH, Levine RA, Utts J. Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. Am Stat. 2019; 73(sup1):352–357.

56. Benjamin DJ, Berger J, Johannesson M, Nosek BA, Wagenmakers E, Berk R, et al. Redefine statistical significance. PsyArXiv. 2017.

57. Ioannidis JPA. The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance. JAMA. 2019; 321(21):2067–2068. https://doi.org/10.1001/jama.2019.4582 PMID: 30946431

58. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ (Clinical research ed.). 2001; 322(7280):226–231. https://doi.org/10.1136/bmj.322.7280.226 PMID: 11159626

59. Nuzzo R. Scientific method: Statistical errors. Nature. 2014; 506(7487):150–152. https://doi.org/10.1038/506150a PMID: 24522584

60. Mervis J. Why null results rarely see the light of day. Science. 2014; 345(6200):992–992.

61. The importance of no evidence. Nat Hum Behav. 2019; 3(3):197–197. https://doi.org/10.1038/s41562-019-0569-7 PMID: 30953022

62. Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics. 2012; 90(3):891–904.

63. Bishop D. Rein in the four horsemen of irreproducibility. Nature. 2019. https://doi.org/10.1038/d41586-019-01307-2 PMID: 31019328

64. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science. 2014; 345(6203):1502–1505.

65. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979; 86(3):638–641.

66. Sterling TD. Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa. J Am Stat Assoc. 1959; 54(285):30–34.

67. Sarewitz D. The pressure to publish pushes down quality. Nature. 2016; 533(7602):147–147. https://doi.org/10.1038/533147a PMID: 27172010

68. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The Extent and Consequences of P-Hacking in Science. PLoS Biol. 2015; 13(3):e1002106. https://doi.org/10.1371/journal.pbio.1002106 PMID: 25768323

69. Gadbury GL, Allison DB. Inappropriate Fiddling with Statistical Analyses to Obtain a Desirable P-value: Tests to Detect its Presence in Published Literature. PLoS ONE. 2012; 7(10):e46363. https://doi.org/10.1371/journal.pone.0046363 PMID: 23056287

70. Kerr NL. HARKing: Hypothesizing After the Results are Known. Pers Soc Psychol Rev. 1998; 2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4 PMID: 15647155

71. Locascio JJ. Results Blind Science Publishing. Basic Appl Soc Psych. 2017; 39(5):239–246. https://doi.org/10.1080/01973533.2017.1336093 PMID: 29780193

72. COS. Preregistration. 2018. Available from: https://www.cos.io/initiatives/prereg.

73. Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. PLoS Biol. 2019; 17(5):e3000246. https://doi.org/10.1371/journal.pbio.3000246 PMID: 31042704

74. Chambers CD, Tzavella L. The past, present and future of Registered Reports. Nat Hum Behav. 2022; 6(1):29–42. https://doi.org/10.1038/s41562-021-01193-7 PMID: 34782730

75. Replication Cumming G. and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. Perspect Psychol Sci. 2008; 3(4):286–300.

76. Betensky RA. The p-Value Requires Context, Not a Threshold. Am Stat. 2019; 73(sup1):115–117.

77. Brereton RG. The use and misuse of p values and related concepts. Chemometr Intell Lab Syst. 2019; 195:103884.

78. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". Am Stat. 2019; 73 (sup1):1–19.

79. Gross LJ, McCord RP, LoRe S, Ganusov VV, Hong T, Strickland WC, et al. Enhancing Quantitative and Data Science Education for Graduate Students in Biomedical Science. bioRxiv. 2021:2021.12.03.471108.

80. Leek J. Five ways to fix statistics. Nature. 2017; 551:557–559. https://doi.org/10.1038/d41586-017-07522-z PMID: 29189798

81. Anderson LW, Krathwohl DR, Bloom BS. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Editors, Anderson LW, Krathwohl D; contributors, Airasian PW. Complete ed. New York: Longman; 2001.

82. Armstrong P. Bloom's Taxonomy. 2010. Available from: https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/.

83. Good PI, Hardin JW. Common Errors in Statistics (And How to Avoid Them). Hoboken, New Jersey: John Wiley & Sons; 2012.

84. Reinhart A. Statistics Done Wrong: The Woefully Complete Guide. San Francisco, CA: No Starch Press; 2015.

85. Vigen T. Spurious Correlations. 2015. Available from: https://www.tylervigen.com/spurious-correlations.

86. Odetunmibi OA, Adejumo AO, Anake TA. A study of Hepatitis B virus infection using chi-square statistic. J Phys Conf Ser. 2020;1734(012010).

87. Ståhle L, Wold S. Analysis of variance (ANOVA). Chemometr Intell Lab Syst. 1989; 6(4):259–272.

88. Evans SJ. Uses and abuses of analysis of variance. Br J Clin Pharmacol. 1983; 15(6):629–48. https://doi.org/10.1111/j.1365-2125.1983.tb01544.x PMID: 6347228

89. Lehtinen M., Lagheden C, Luostarinen T, Eriksson T, Apter D, Bly AA, et al. Human papillomavirus vaccine efficacy against invasive, HPV-positive cancers: population-based follow-up of a cluster-randomised trial. BMJ Open. 2021; 11(12):e050669. https://doi.org/10.1136/bmjopen-2021-050669 PMID: 35149535

90. Noya O, Gabaldón Berti Y, Alarcón de Noya B, Borges R, Zerpa N, Urbáez JD, et al. A population-based clinical trial with the SPf66 synthetic Plasmodium falciparum malaria vaccine in Venezuela. J Infect Dis. 1994; 170(2):396–402. https://doi.org/10.1093/infdis/170.2.396 PMID: 8035026

91. Zangwill KM, Eriksen E, Lee M, Lee J, Marcy SM, Friedland LR, et al. A population-based, postlicensure evaluation of the safety of a combination diphtheria, tetanus, acellular pertussis, hepatitis B, and inactivated poliovirus vaccine in a large managed care organization. Pediatrics. 2008; 122(6):e1179–85. https://doi.org/10.1542/peds.2008-1977 PMID: 19047220