Reproducibility Literature Review

Among statisticians at large, there is a reproducibility crisis.[1] In published articles from reputable journals, reviewers intending to reproduce the results of study face obstacles that authors fail to clear.[2] The availability of good data to reviewers creates one of the biggest challenges in reproducibility.[3] Authors may not share data sets (even upon reasonable request) or reviewers may struggle to recreate/recover the data.[3,4] From there, many data sets have incorrect or inconsistent data reporting which can further complicate the replication process.[3] Without identical data to those of the study, it is less likely that the published results can be achieved, harming the credibility of the work. As such, whether due to data inconsistency or simply poor statistical practices, some studies' claims of significance fail upon further review.[2] While much of the literature studying the validity of these claims are in psychological research, it stands to reason that these issues are relevant in all fields of research.[2]

One of the greatest ways to address this is in the classroom. Training statisticians in reproducibility from the start of their work can create better practice in the field. With properly trained researchers, the effort into properly reporting methods can maximize the reproducibility of published research studies in the future.[1] As teaching reproducibility earlier in statistics education is becoming standard, it is important to understand the current ways it is addressed and areas to expand upon.[1]

A three pronged approach advocated in Dogucu & Centinkaya-Rundel (2022) proposes teaching reproducibility with reproducible research, teaching reproducibility, and reproducible teaching.[1,5] This pedagogy involves expanding upon each of the workflows for research, learning, and teaching.[5] Research and learning reproducibility is somewhat integrated already, with opportunities for research in labs or faculty-advised projects as well as instructors introducing students to technology and resources used to create a reproducible workflow.[1]

The idea of teaching reproducibility is less standard in practice than the other two. However, it can have great impact on student success. Teaching reproducibility provides an opportunity to teach by example, exposing students to the instructor's own best practices in reproducibility which they might not otherwise witness.[5] It furthermore seeks to improve collaboration between instructors to learn and advance statistics education.[5] The core pillars of this principle are computational reproducibility, documentation, and openness. Computational reproducibility entails integrating plain text with code (literate programming), keeping raw data and tracking steps taken, file organization, and version control.[5] For documentation, instructors should write plain text files to track the software (and versions), steps for file reproduction, codebook (explaining data structure), and a style guide to aid in

organization.[5] Finally, materials should be shared as openly as possible; instructors should provide clear licensing and a site host to share with the public.[5]

To facilitate teaching and learning reproducibility, tools like R Markdown/Quarto or Jupyter notebooks effectively unify notes with R code.[1] Using these notebook files allows students to bypass struggles with pure coding and computational environments (some difficulties running R packages on different operating systems).[6] By allowing instructors to integrate step-by-step instructions with working code, students can better step through code and see how to replicate solutions.[6] Much of the research done on these sort of technologies uses R, but identical products exist for a wide range of programming languages, and the impacts they bring to education are likely similar.

Integrating GitHub, either on its own or through GitHub Classroom can further teach students the necessity of a reproducible workflow. Version control systems like Git allow researchers to track their workflow and trace errors, while maintaining the openness and clarity needed in a reproducible workflow.[7] Given the importance of GitHub as an online host for open source projects and data sharing, it is a crucial platform to be familiar with in research. The study of interest links a GitHub repository of all data and code used in their project, allowing for easy access to attempt to reproduce their work. Broader adoption of such a practice will benefit statisticians at large, and the greatest way to start is with integration at the educational, training level.[8]

At the present, specifically addressing reproducibility in the classroom is not treated as a priority. Out of the USnews' top ten biostatistics programs in the US,[9] not one mentions reproducibility as a topic in the most recent publicly available syllabus or course description.[10–19] Of course, some of these may mention it as an important principle throughout or address it in a later course, but none appear to contain a specific lesson addressing it. Introducing it as a topic on par with hypothesis testing, confidence intervals, or regression would serve to bring reproducibility to the forefront of the minds of future statisticians and aid in solving the reproducibility crisis.

Due to the lack of curricula on this topic, it is difficult to design a lesson based on prior utilized material. However, a Master's course at the Universite Paris-Saclay, the Reprohackathon provides students with an overview of challenges in reproducibility.[20] Part of this involves detailing the best practice of developing a workflow to achieve reproducibility.[20] Additionally, the course includes a data analysis projects where students reanalyze data from a previously published study, similar to the goals of this research project.[20] While this project is intended to be nearly semester-long, and thus would not be an exact match for an introductory biostatistics course, there are certainly ways to adapt it; instructors could design a study with fairly simple analysis designed to be done in a week of

classwork or provide code with errors that students need to address. Regardless, this model of a project to teach reproducibility certainly trains students on how to maximize their reproducibility in the future.

Educating students on reproducibility early in their statistical studies is important to begin correcting the irreproducibility of many published studies in the future. By making all aspects of the classroom reproducible, students can learn by example and gain a broader understanding of how and why to maintain a fully reproducible workflow.

References:

1. Dogucu M. Reproducibility in the Classroom. *Annu Rev Stat Its Appl*. 2025;12(Volume 12, 2025):89-105. doi:10.1146/annurev-statistics-112723-034436

2. Artner R, Verliefde T, Steegen S, et al. The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychol Methods*. 2021;26(5):527-546. doi:10.1037/met0000365

3. López-Nicolás R, Lakens D, López-López JA, et al. Reproducibility of Published Meta-Analyses on Clinical-Psychological Interventions. *Adv Methods Pract Psychol Sci*. 2024;7(1):25152459231202929. doi:10.1177/25152459231202929

4. Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J Clin Epidemiol*. 2022;150:33-41. doi:10.1016/j.jclinepi.2022.05.019

5. Dogucu M, and Çetinkaya-Rundel M. Tools and Recommendations for Reproducible Teaching. *J Stat Data Sci Educ*. 2022;30(3):251-260. doi:10.1080/26939169.2022.2138645

6. Kumwichar P. Enhancing Learning About Epidemiological Data Analysis Using R for Graduate Students in Medical Fields With Jupyter Notebook: Classroom Action Research. *JMIR Med Educ*. 2023;9:e47394. doi:10.2196/47394

7. Ram K. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol Med*. 2013;8(1):7. doi:10.1186/1751-0473-8-7

8. Fiksel J, Jager ,Leah R., Hardin ,Johanna S., and Taub MA. Using GitHub Classroom To Teach Statistics. *J Stat Educ*. 2019;27(2):110-119. doi:10.1080/10691898.2019.1617089

9. Best Biostatistics Programs in America. Accessed June 25, 2025. https://www.usnews.com/best-graduate-schools/top-science-schools/biostatistics-rankings

10. Ayton S. Introduction to Biostatistical Methods Course Description. Published online October 30, 2024. Accessed June 25, 2025. https://www.publichealth.columbia.edu/academics/course-directory/archived-course-directory

11. Introduction to Biostatistics Course Description. Accessed June 25, 2025. https://catalog.registrar.ucla.edu/course/2023/biostat100a?siteYear=2023

12. Introduction to Biostatistics Course Description. Accessed June 25, 2025. https://extension.berkeley.edu/search/publicCourseSearchDetails.do?method=load&courseId=41598

13. Monaco J. Syllabus for BIOS500H: Introduction to Biostatistics. Published online Fall 2024. Accessed June 25, 2025. https://uncch.instructure.com/courses/66306/pages/syllabus?module_item_id=902523

14. Thornton T. Syllabus for BIOST 514/517: (Applied) Biostatistics I.

15. Jiang H. Syllabus for BIOSTAT 600: Introduction to Biostatistics. Accessed June 25, 2025. https://sph.umich.edu/research-education/courses/syllabi/BIOSTAT600.pdf

16. Liublinska V, Haneuse. Syllabus for HST-190: Introduction to Biostatistics.

17. German J. Syllabus for MTR 6000: Introduction to Biostatistics. Published online Fall 2024.

18. Shore M. Syllabus for PubH 6450: Biostatistics I. Published online Spring 2018. Accessed June 25, 2025. https://www.sph.umn.edu/sph/wp-content/uploads/syllabi/2018/spring/pubh-6450.pdf

19. Varadhan R. Methods in Biostatistics I | Johns Hopkins | Bloomberg School of Public Health. Accessed June 25, 2025. https://publichealth.jhu.edu/course/42107

20. Cokelaer T, Cohen-Boulakia S, Lemoine F. Reprohackathons: promoting reproducibility in bioinformatics through training. *Bioinformatics*. 2023;39(Supplement_1):i11-i20. doi:10.1093/bioinformatics/btad227