Comprehensive Analysis of Code and Data Availability in Biomedical Research

Dhrithi Deshpande

The Mangul Lab, University of Southern California

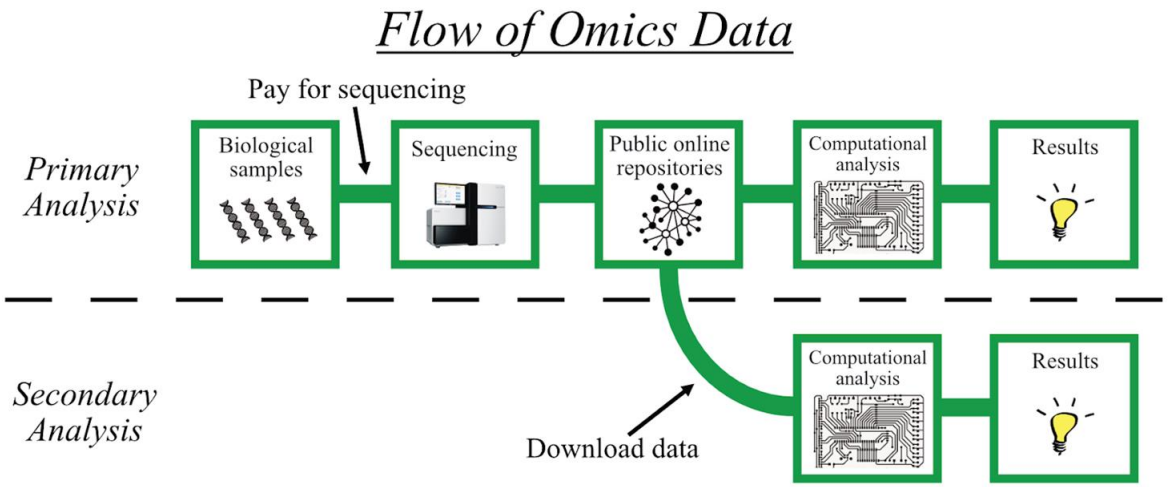
ABACBS – November 2020

USC School of Pharmacy

University of Southern California

Why is sharing code and data important?

- Secondary analysis
- Improve reproducibility
- Enhance robustness of biomedical research
- Increase efficiency

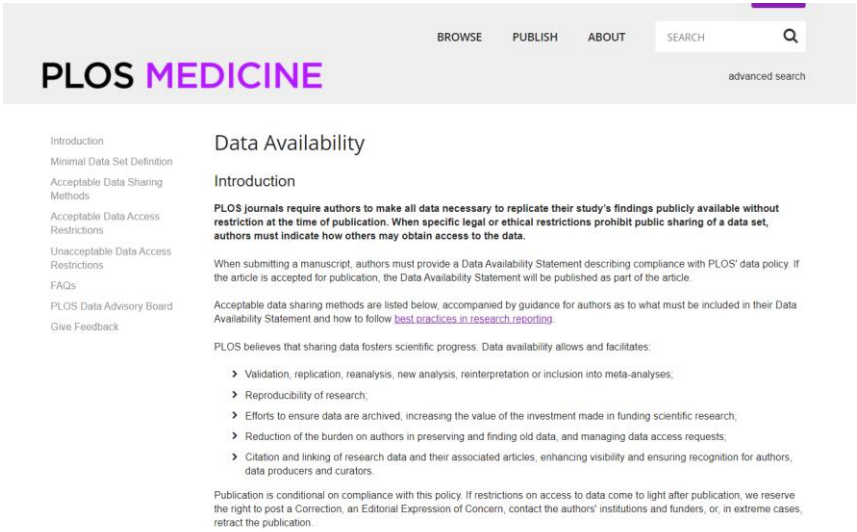


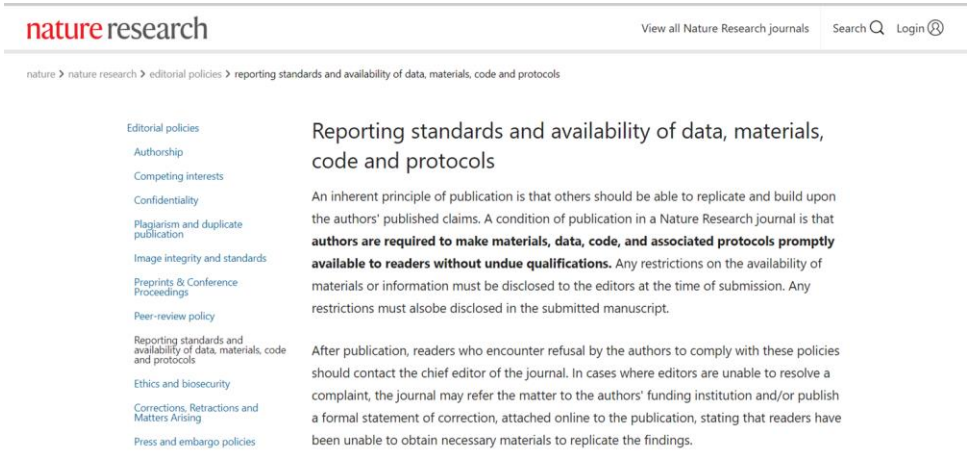
NIH National Institutes of Health
Turning Discovery Into Health

Principles and Guidelines for Reporting Preclinical Research

NIH held a joint workshop in June 2014 with the Nature Publishing Group and Science on the issue of reproducibility and rigor of research findings, with journal editors representing over 30 basic/preclinical science journals in which NIH-funded investigators have most often published. The workshop focused on identifying the common opportunities in the scientific publishing arena to enhance rigor and further support research that is reproducible, robust, and transparent.

- Some of the publishers such as Nature, PLOS, the Royal Society, etc have implemented these data sharing requirements.

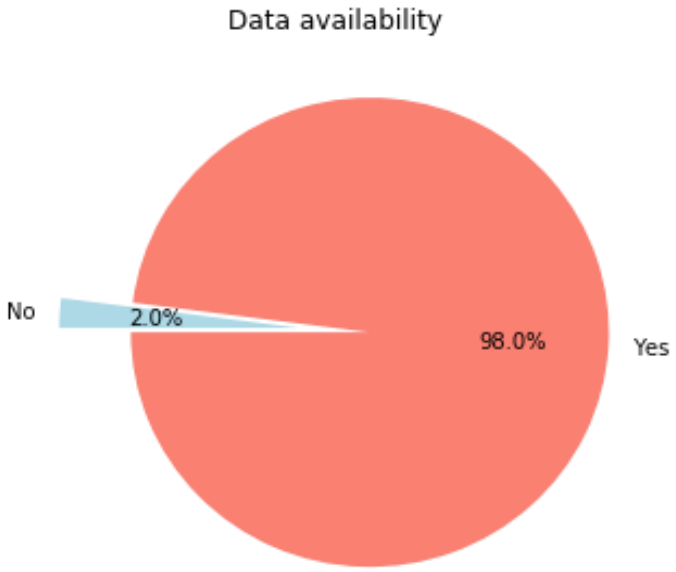


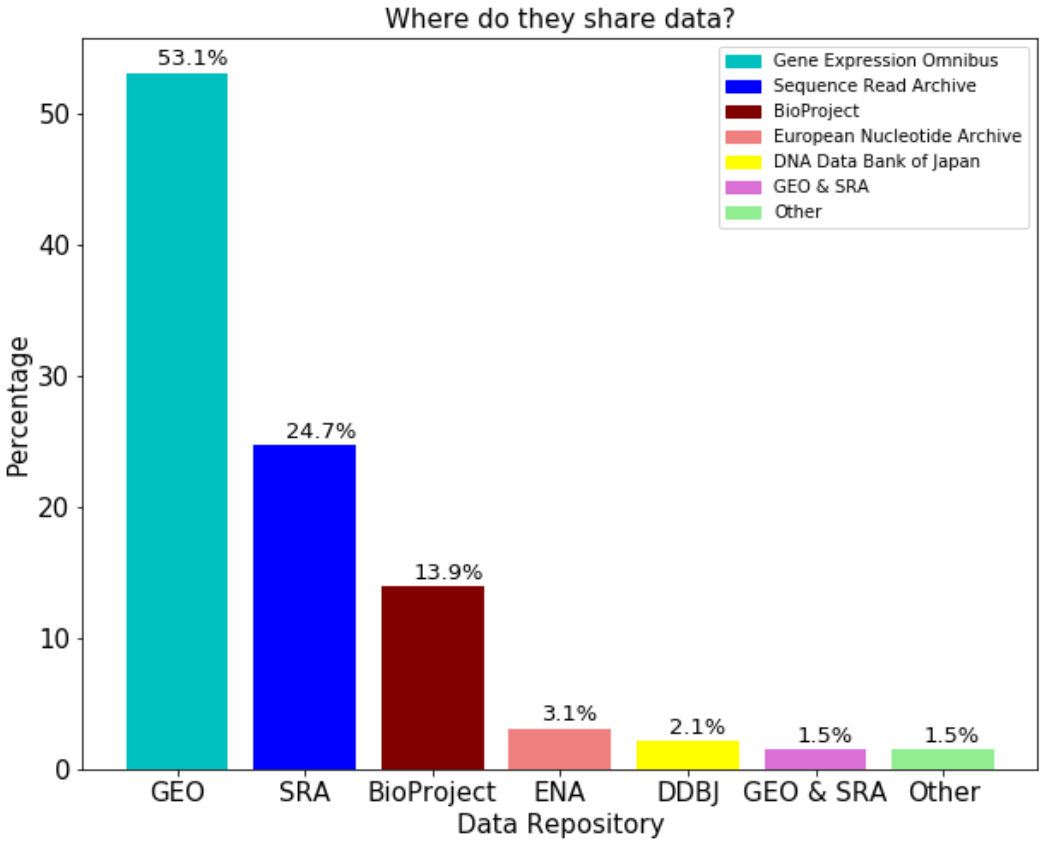


What can we do about this?

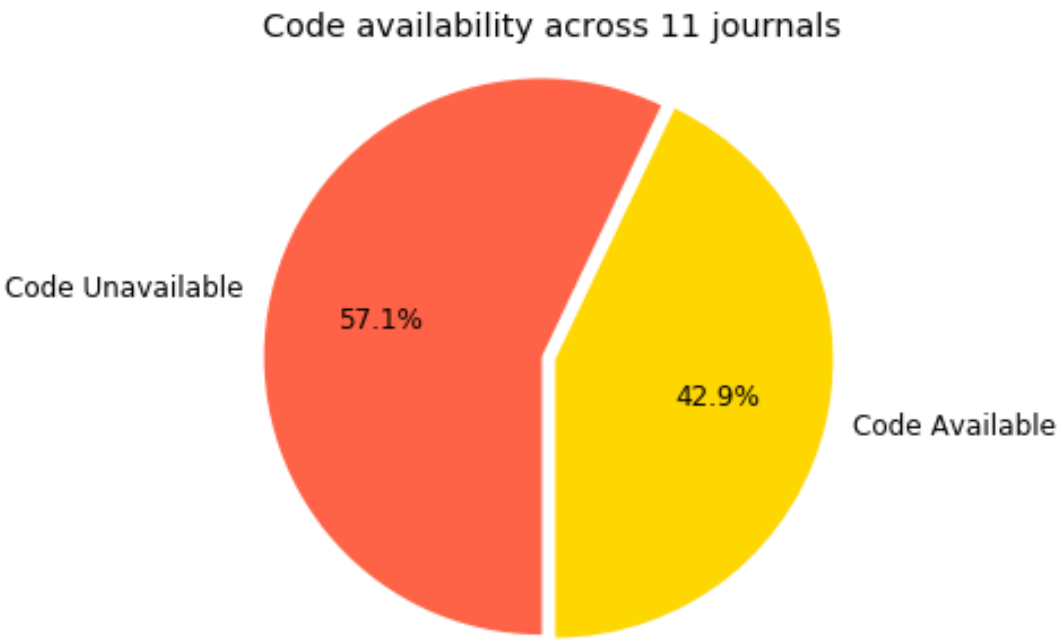
- Is it possible to develop procedures to ensure transparency and reproducibility?
- Who is responsible for enforcing this change?
  - Many journals have mandated data requirement
- Individual researchers or Journals?
- What roles should institutions play?

Raw Data Availability across 200 papers from 11 biomedical journals



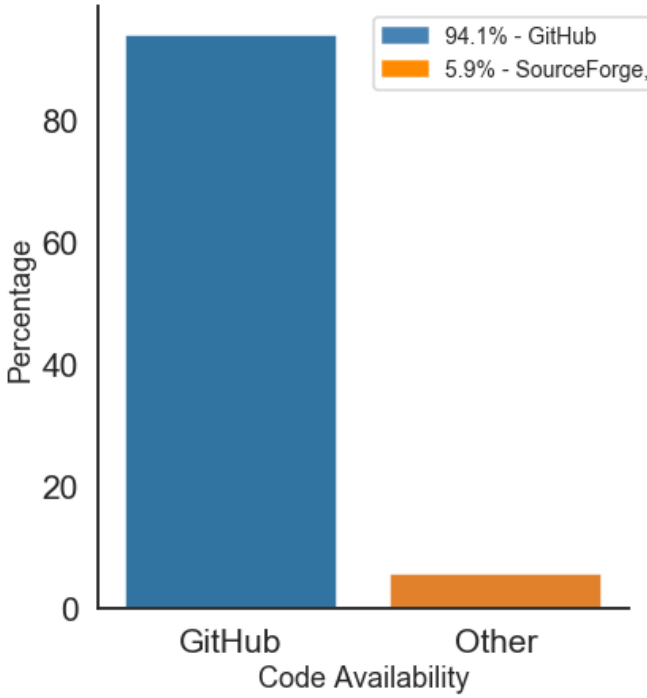


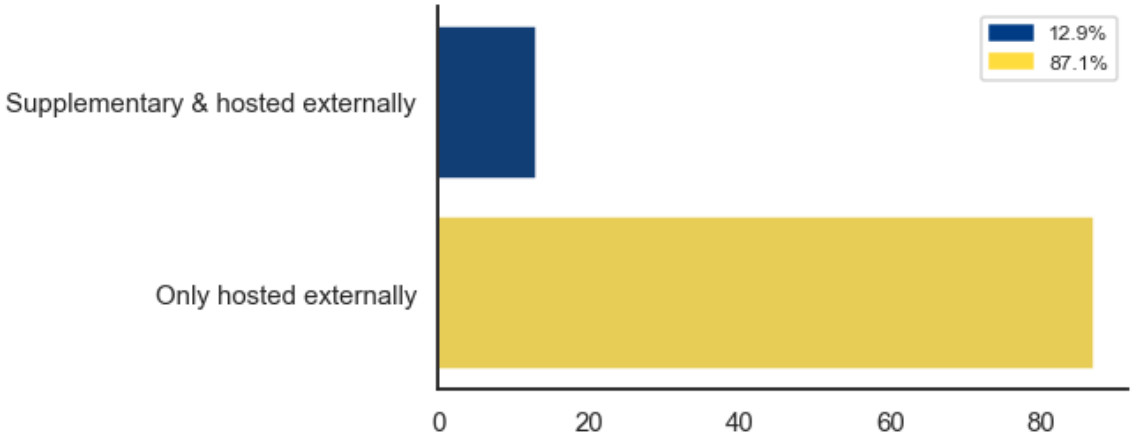
Code Availability across 200 papers from 11 biomedical journals



What is the reason for not sharing code?

Why are there differences in where the code is shared?

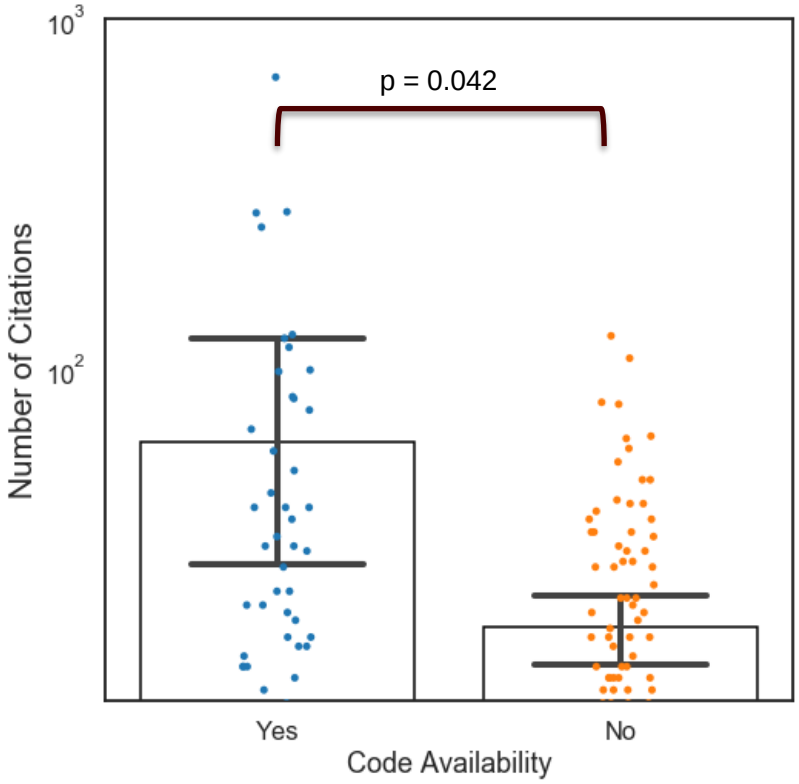




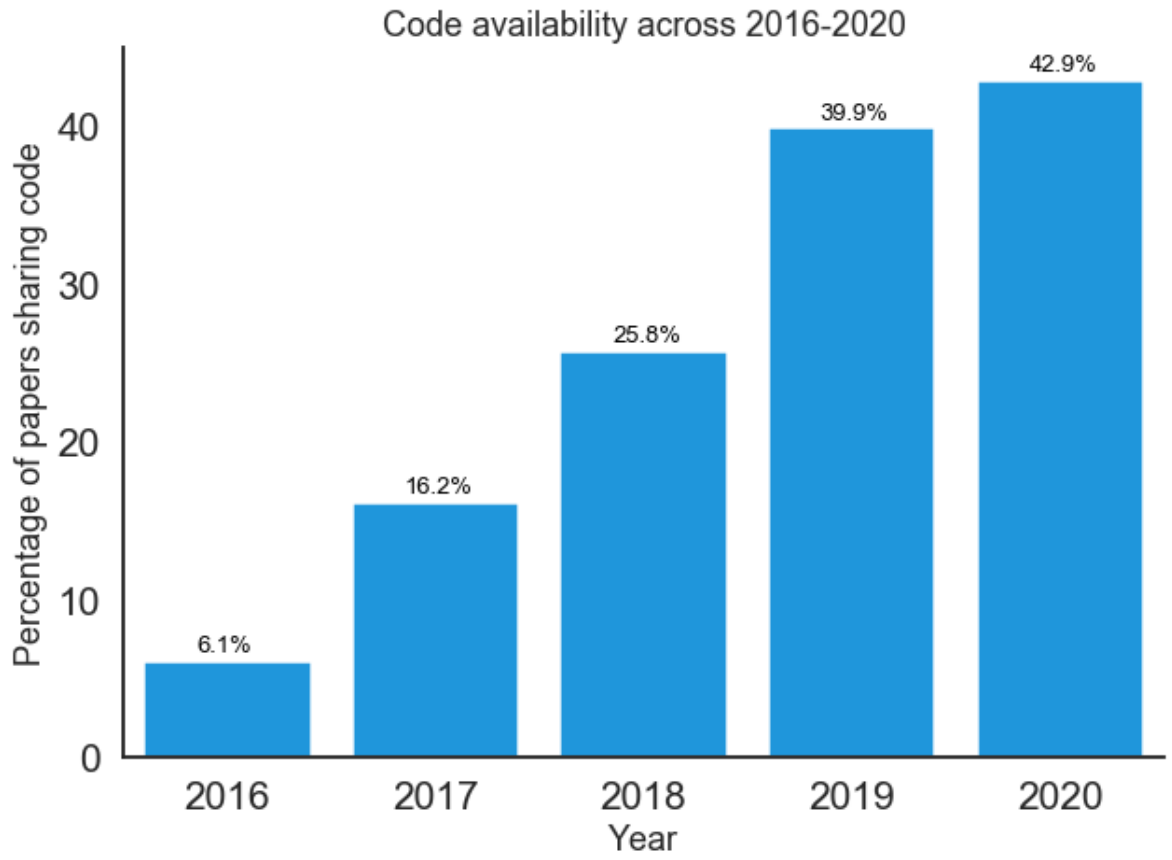
Open Questions

- Should GitHub or any other open-source repository be mandated?
- Platforms like GitHub do not accept large files
- GitHub, URLs - commercial - how reliable are commercial repositories?
- It is unclear how stable resources on the internet are.

Association between sharing code and citations



Has code availability improved over the years?



Conclusion

- For those research papers which share code, who will be the authority to verify and ensure that the code shared is usable and reproducible?
- Current efforts mostly rely on individual researchers. Many journals are taking initiatives to ensure code and data availability.
- We as a community should develop and adopt the best practices to address this problem of accessibility and reproducibility.

Acknowledgements

- Dr Serghei Mangul, University of Southern California, LA
- Ruiwei Guo, University of Southern California, LA
- Nicholas Darci-Maher, UCLA

Questions?