



# Comprehensive Analysis of Code and Data Availability in Biomedical Research

*Dhrithi Deshpande*

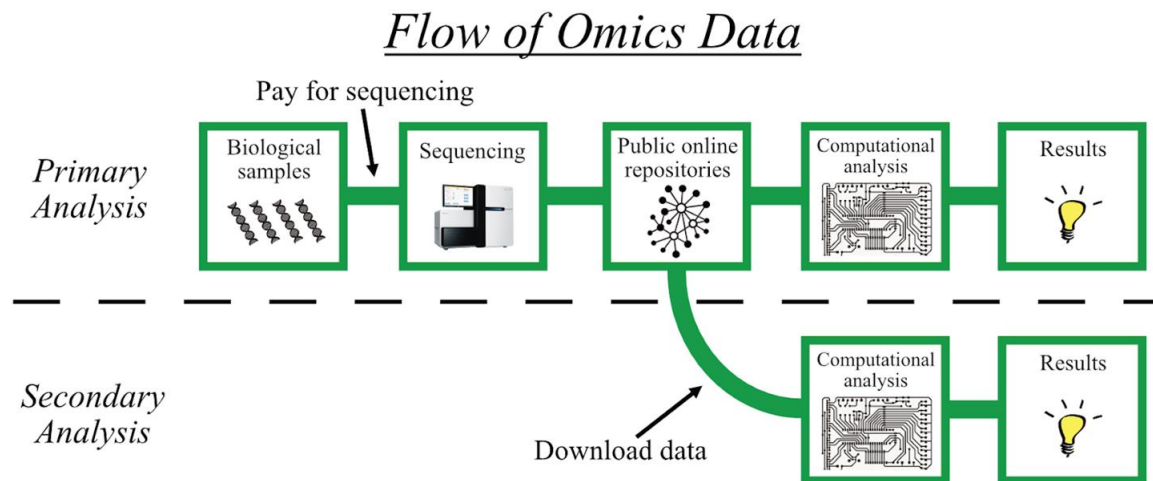
*The Mangul Lab, University of Southern California*

*ABACBS – November 2020*



# Why is sharing code and data important?

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





# 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.

**PLOS MEDICINE**

BROWSE PUBLISH ABOUT SEARCH Q

advanced search

Introduction  
Minimal Data Set Definition  
Acceptable Data Sharing Methods  
Acceptable Data Access Restrictions  
Unacceptable Data Access Restrictions  
FAQs  
PLOS Data Advisory Board  
Give Feedback

## Data Availability

### Introduction

**PLOS journals require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, authors must indicate how others may obtain access to the data.**

When submitting a manuscript, authors must provide a Data Availability Statement describing compliance with PLOS' data policy. If the article is accepted for publication, the Data Availability Statement will be published as part of the article.

Acceptable data sharing methods are listed below, accompanied by guidance for authors as to what must be included in their Data Availability Statement and how to follow [best practices in research reporting](#).

PLOS believes that sharing data fosters scientific progress. Data availability allows and facilitates:

- Validation, replication, reanalysis, new analysis, reinterpretation or inclusion into meta-analyses;
- Reproducibility of research;
- Efforts to ensure data are archived, increasing the value of the investment made in funding scientific research;
- Reduction of the burden on authors in preserving and finding old data, and managing data access requests;
- Citation and linking of research data and their associated articles, enhancing visibility and ensuring recognition for authors, data producers and curators.

Publication is conditional on compliance with this policy. If restrictions on access to data come to light after publication, we reserve the right to post a Correction, an Editorial Expression of Concern, contact the authors' institutions and funders, or, in extreme cases, retract the publication.

**nature research**

View all Nature Research journals Search Q Login

nature > nature research > editorial policies > reporting standards and availability of data, materials, code and protocols

## Reporting standards and availability of data, materials, code and protocols

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Research journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript.

After publication, readers who encounter refusal by the authors to comply with these policies should contact the chief editor of the journal. In cases where editors are unable to resolve a complaint, the journal may refer the matter to the authors' funding institution and/or publish a formal statement of correction, attached online to the publication, stating that readers have been unable to obtain necessary materials to replicate the findings.

Editorial policies  
Authorship  
Competing interests  
Confidentiality  
Plagiarism and duplicate publication  
Image integrity and standards  
Preprints & Conference Proceedings  
Peer-review policy  
Reporting standards and availability of data, materials, code and protocols  
Ethics and biosecurity  
Corrections, Retractions and Matters Arising  
Press and embargo policies



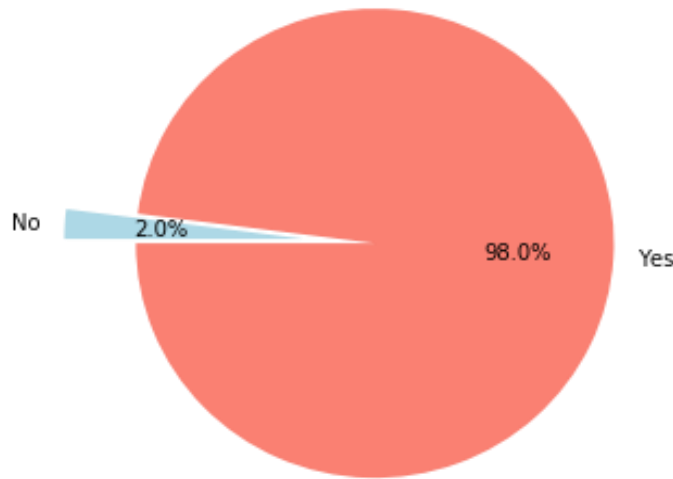
# What can we do about this?

- 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?

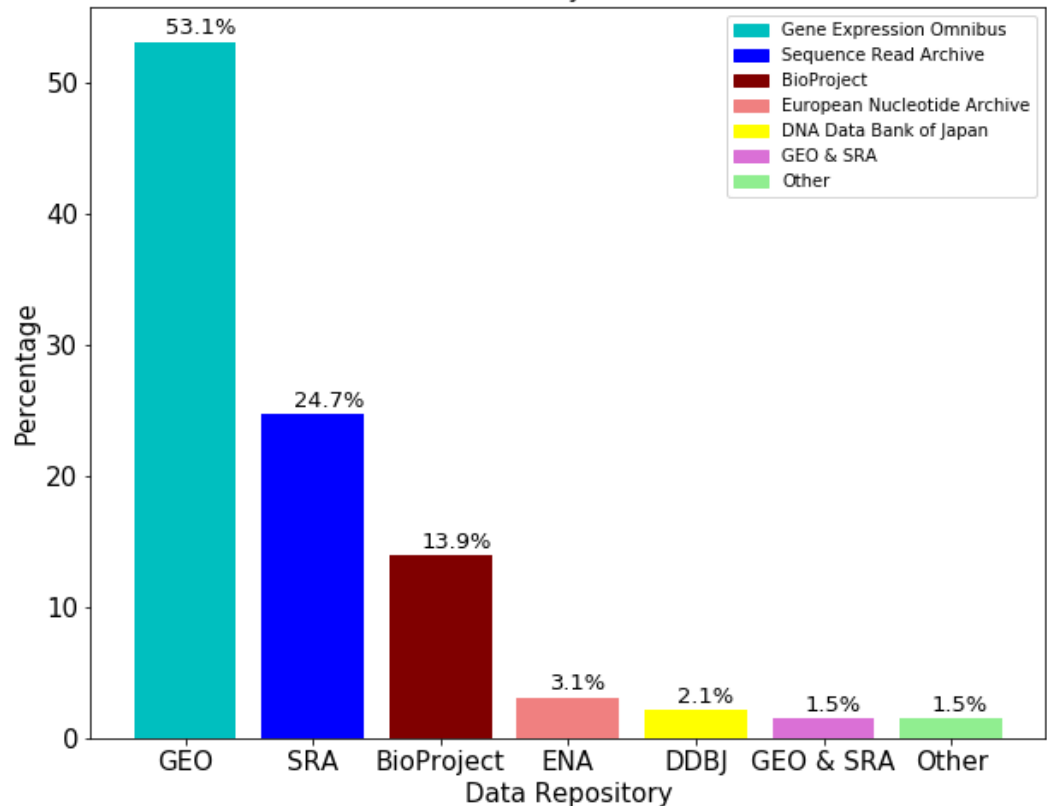
# Raw Data Availability across 200 papers from 11 biomedical journals



Data availability



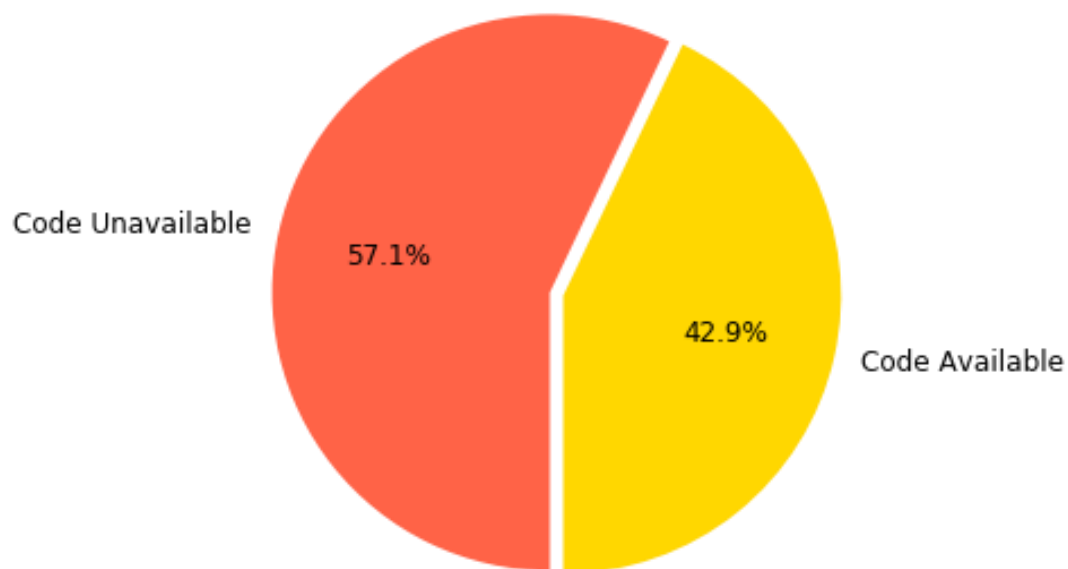
Where do they share data?



# Code Availability across 200 papers from 11 biomedical journals

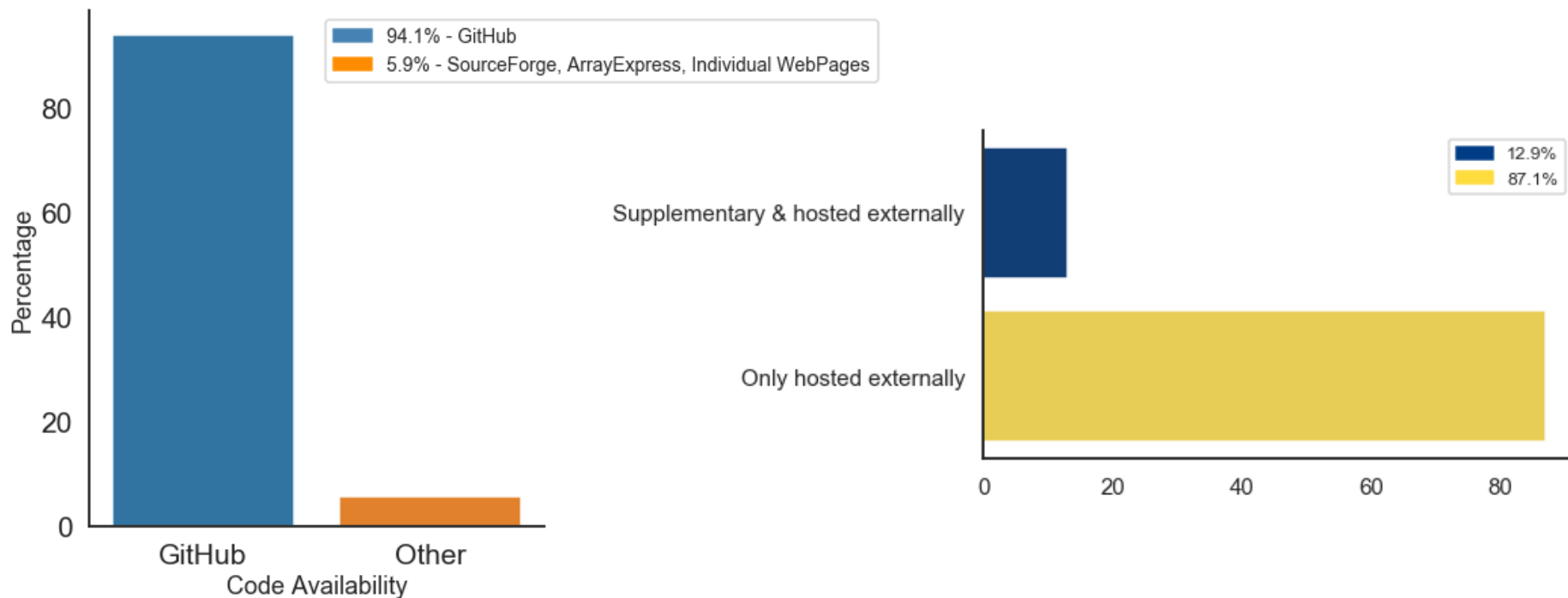


Code availability across 11 journals



What is the reason for not sharing code?

# Why are there differences in where the code is shared?





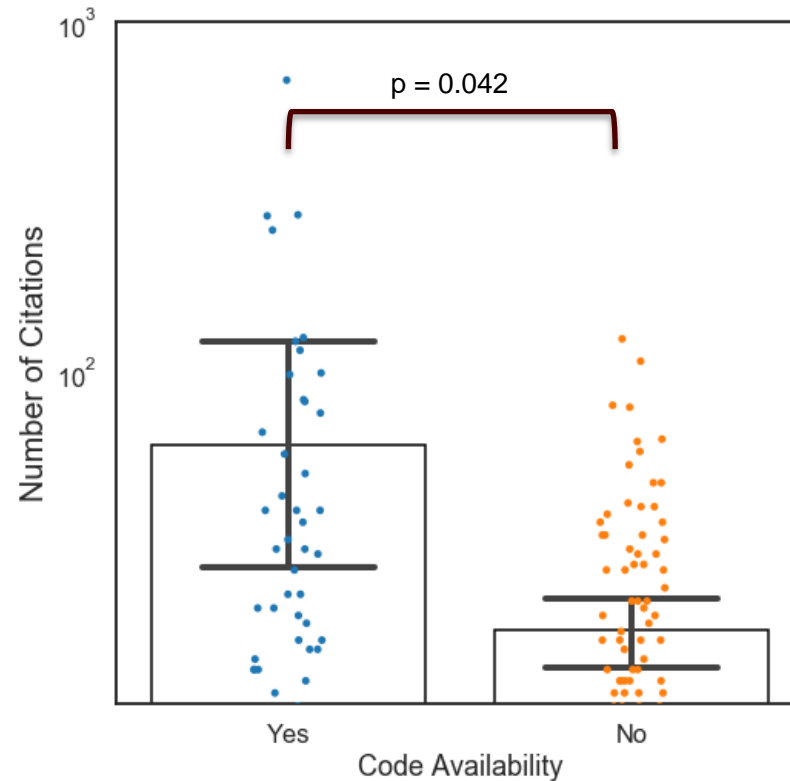


# Open Questions

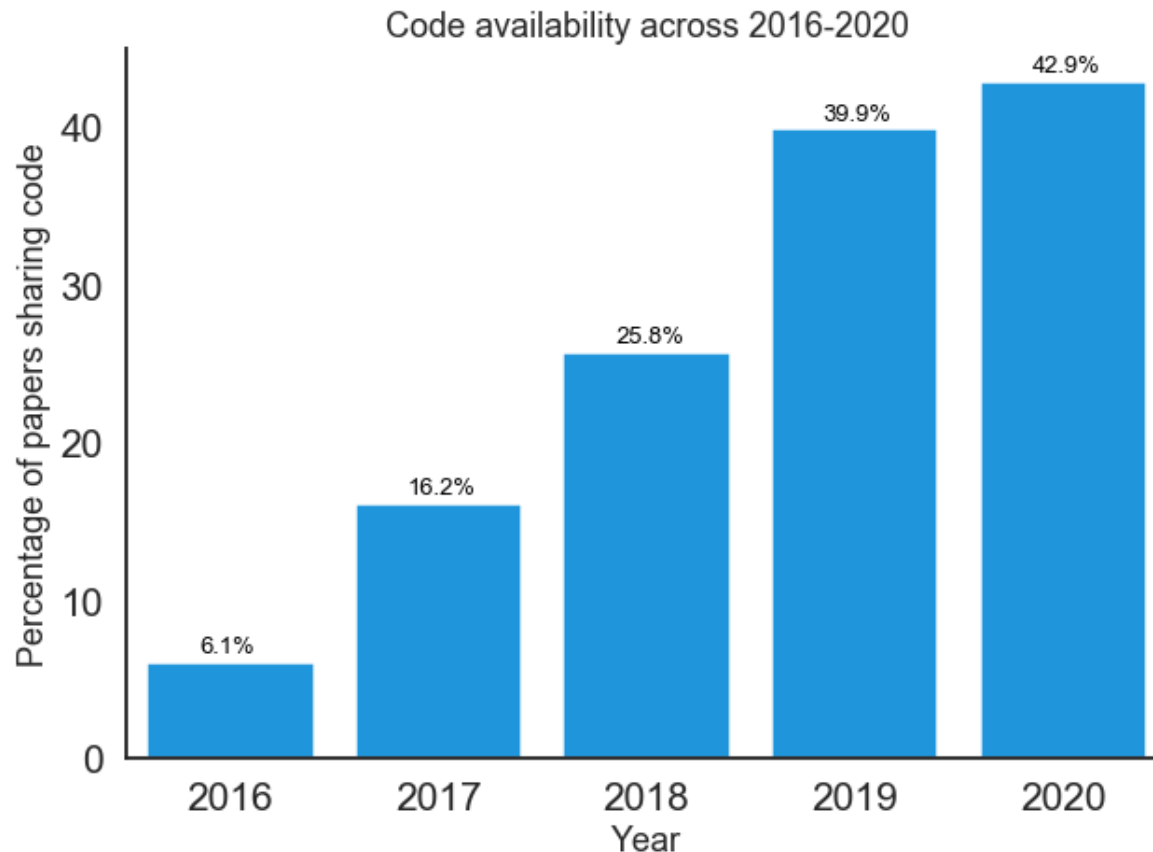
- Should GitHub or any other open-source repository be mandated?
- Platforms like GitHub do not accept large files
- GitHub, URLs - commercial - how reliable are commercial repositories?
- It is unclear how stable resources on the internet are.



# 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?