Paper Title

A comprehensive analysis of code and data availability in biomedical research

Authors and Affiliations

Dhrithi Deshpande

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

dhritideshpande@gmail.com

ORCID: <https://orcid.org/0000-0003-3794-7364>

Aditya Sarkar

School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, North Campus, Kamand, Mandi, Himachal Pradesh, 175005, India

asnov2k@gmail.com

ORCID: <https://orcid.org/0000-0002-4496-8289>

Ruiwei Guo

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

ruiweigu@usc.edu

ORCID: <https://orcid.org/0000-0002-6978-8735>

Andrew Moore

Daniel J. Epstein Department of Industrial and Systems Engineering, Viterbi School of Engineering, University of Southern California

mooreand@usc.edu

ORCID ID: <https://orcid.org/0000-0003-1457-4593>

Nicholas Darci-Maher

Computational and Systems Biology, University of California, Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA

niko.darcimaher@gmail.com

ORCID: <https://orcid.org/0000-0002-8124-030X>

Serghei Mangul

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA

serghei.mangul@gmail.com

ORCID: <https://orcid.org/0000-0003-4770-3443>

Abstract

Modern data-driven research increasingly depends on quantitative analysis, yet effective mechanisms ensuring data and analysis transparency and reproducibility are yet to be developed and adopted widely. The importance and benefits of sharing research products has been recognized widely by the scientific community. In biomedical research, it is not only imperative

to publish a detailed description of the study design, methodology, results and interpretation, but there is a pressing need to make all the research products publicly available, shareable, well documented to increase transparency and reproducibility. Current efforts in sharing research products mostly rely on individual researchers and widely but variably enforced by these individuals and research organizations. However, an increasing body of evidence in recent years also points to a growing problem of reproducibility across scientific disciplines, i.e. published results often contain analyses that are non replicated due to lack of documentation, code and data required to reproduce the analysis. Our results indicate that only 36% of the scientific manuscripts published in prominent biomedical journals share raw data and 9% of the papers share code. We hope that our analysis informs and exhorts the biomedical community to design effective strategies to be widely adopted by the researchers to improve the current scenario of transparency and reproducibility of data-driven biomedical research.

Keywords

Analytical code, raw data, transparency, rigor, reproducibility

1. Introduction

The true spirit of science is where every new discovery is based on strengthening existing work^{1,2}. Often, new discoveries can often be made uniquely based on shared data³. In data-driven biomedical research, it is not only imperative to publish a detailed description of the study design, methodology, results and interpretation, but, there is a pressing need to make all the research products underlying biomedical data including code and raw data used in corresponding

analyses publicly available, well documented and reproducible⁴. As more researchers are getting involved in data driven research, the biomedical community has recognized and advocated the importance of sharing data and code used for analysis.

The importance and benefits of sharing research data has been emphasized by many scientific communities, peer reviewed journals, academic societies, funding agencies, etc. The current mechanism includes providing various checklists for researchers and best practice guidelines. At a joint workshop held at the NIH in June 2014, editors from various journals met to discuss what is needed and how to enhance reproducible, robust, and transparent science. As an outcome, the principles and guidelines for reporting preclinical research were developed, which included the recommendation that journals require that all of the data supporting a paper's conclusion must be made available as part of the review process and after publication, these datasets be deposited to public repositories⁵. These principles have even been endorsed by the representatives of over 70 journals, associations, and scientific societies. Funding agencies and the research community in recognition of the needs of higher standards, today, have a resolve for open and reproducible research practices⁶. Despite these efforts, checklists and author guidelines which have been developed, research shows that the reporting of data supporting the inference is inadequate across the majority of research publications⁷⁻⁹. A poll conducted by two prominent biomedical journals in 2019 also showed that the majority of the researchers suggest that the largest change they want to see in support of Open Science is an increased open data and protocol sharing^{10,11}. The poll also revealed that researchers believe committing to open protocol and data sharing can eliminate the hindrance faced by other researchers for secondary analysis and avoid reprinting of similar data¹².

In biomedical research, there are different types of research data which are generated during the course of a research project we frequently come across - raw data, metadata, summary data, etc. Raw data is often defined as the primary data that is produced by the instrument (for example, genomics data produced by sequencing machines)¹³. Raw data is the data that needs to be cleaned and analyzed using bioinformatics methods before being able to interpret and obtain clinically meaningful or biologically relevant conclusions or research products. Despite sharing raw data being of crucial importance to ensure reproducibility of published results, it does not suffice. For any raw data to be consequential, data analysis is required. Biomedical research is becoming increasingly computationally intensive and data-driven¹⁴; scientists are using code to describe and process data, and summarize their results using effective visualisation methods, implying that code availability in a published article is of paramount importance. Sharing both raw data and analytical code (here on, referred to as data and code respectively) used in research, and making the study protocol or pipeline used easily discoverable can facilitate computational reproducibility and enhance robustness of biomedical research⁴. Most importantly, when data and code are shared along with the data analysis, other researchers can curate, examine and re-analyze the data in ways that further its value, known as secondary analysis; making novel biological discoveries and fostering the efficiency of biomedical research in terms of enabling fellow researchers to view data analysis pipelines and build on existing methods¹⁵. In general, data and code sharing promises the true spirit of science where every new discovery is based on strengthening existing work^{1,2}.

Across all disciplines, research projects are becoming more data-driven as well as computationally intensive. Transparency in terms of sharing data and code used for analysis, will enable us to effectively manage, structure and carry out the computational analysis in research projects. Well structured data and code is not only beneficial for reproducibility but also for researchers themselves. Well organized data and code will also allow swift installation and rerun of analyses to equip visualization, in addition to making the analyses robust against code errors and mistakes. In contrast to sharing data, which is widely enforced by scientific journals and research organizations, there is limited guidance on code sharing which, in addition to data, represents an essential component of reproducible and rigorous research. Multiple efforts have been made in the recent past, to make research data open and FAIR - Findable, Accessible, Interoperable, and Reusable¹⁶, although the impact of such efforts remain unclear.

In this study, we have performed a systematic analysis of the code and data sharing practices across 480 manuscripts from 8 prominent biomedical journals published between 2016 - 2021. The results of our study indicate that raw data sharing practices have increased from 6.7% in 2016 to 36.2% in 2021 and code sharing practices have increased appreciably from 0.8% in 2016 to 9% in 2021.

2. Results

2a. Raw data and analytical code availability in biomedical research

We have appraised the proportion of articles sharing data and code. The manuscripts published between 2016 and 2021 from 8 prominent biomedical journals were retrieved from

PubMed Central open access corpus (Details provided in the Methods Section). From these 12,603 manuscripts, 10 manuscripts from each of the 8 journals were randomly selected from each year between 2016 - 2021. Results of this survey of data sharing practices across the manuscripts conducting primary analyses indicated that 36.2% of the biomedical research articles share raw data (**Figure 1a**). Research papers conducting secondary analyses were not taken into consideration for raw data availability. Out of the 36.2% of the papers, our assessment indicated that the data used for analysis was shared most abundantly on GEO (40.2%) and SRA (16.5%) combined. Other repositories and webpages constituted 43.3% (**Figure 1b**). Further, 32.83% of the articles that shared raw data, also had a data availability statement mentioned in the manuscript.

Our analysis of the code sharing practices across the 480 articles indicated that 91.04% of the articles were not accompanied by code (**Figure 1c**).

In accordance with the FAIR guiding principles¹⁷, fundamentally, the code must first be findable and accessible to be reused for analysis. In our study we considered code and data to be easily findable if the article included a code and/ or a data availability statement. Overall, 42.7% of the articles had a code availability statement. With these papers with a code availability statement, 7.29% were actually accompanied with code and 1.66% of the papers did not have a statement but shared code. Code availability was otherwise mostly stated either in the main text of the manuscript or the data availability statement, and in some instances, in the results or supplementary section.

Of the 8.96% of the papers that did share code, the preponderance, 79.1% of the code was found to be archived on GitHub, 4.7% of the articles shared code in their supplementary materials, 4.7% on Zenodo and 11.6% were hosted on individual webpages (**Figure 1d**).

2b. Effect of journal mandates on sharing data and code

To understand why the differences in availability of the code and data occur, and the effects of journals mandate on data and code sharing, we investigated the policies enforced across the 8 journals. Six out of the eight journals (75%) of the journals have a mandatory data sharing policy and 2 out of the 8 (25%) encourage data sharing or mandate data sharing upon request. While, on the other hand, five out of eight (62.5%) journals mandate code sharing, three out of eight (37.5%) of the journals encourage code sharing or mandate code sharing upon request (**Figure 2**).

According to our analysis, data availability has appreciably improved from 6.7 to 36.2% and code availability has appreciably improved from a meagre 0.8% in 2016 to 9% in 2021 (**Figure 3a and 3b**), however, this is less than ideal. Additionally, we observed the association between sharing code and citations which indicates that articles that share code exhibit an increased number of citations compared with papers that do not share code.

3. Discussion

Our systematic study aims to raise awareness and encourage an open dialogue among journals, funding agencies, and researchers to address the perturbing situation of data and code availability which represents an encumbrance to the growth of biomedical research. It remains unclear as to whether it is the lack of technical skills or cultural beliefs of the researcher that is causing the impediment in conducting reproducible and transparent research⁹. Sharing data and code used for analysis can warrant for reproducibility of the scientific analyses and transparency but there is a paucity of actionable strategies to accelerate the sharing of research products to ensure this.

There is a lack of consensus as to what we as a scientific community can do that is actionable. If there are guidelines and procedures developed to ensure transparency, it is important to unriddle who the authority will be for enforcing this change, to verify and ensure compliance, and that the data and code shared is usable and reproducible. If code and data used for analysis are shared, the question that arises is if one will be able to implement the provided data and code (the same pipeline) involved as the original research article and ensure the same output is obtained using the same datasets (reproducibility) or if qualitatively similar results can be obtained using the same pipeline on different datasets (replicability). Replicability has a very important role, especially in biomedical research, because preliminary studies conducted often are used in downstream clinical research, potentially having a major impact in healthcare. Various studies that have been conducted indicate that reproducing and replicating existing work cannot be easily achieved^{18,19}.

It is an open question whether GitHub or other open-source repositories should be mandated for all the code to be shared. Although, platforms such as GitHub do not accept large files. In principle, the thought arises if it is expedient to rely on commercial platforms. These open source repositories and URLs are commercial, and hence, likely non-permanent, raising issues on their archival stability. Albeit, the advantages of open source repositories and URLs outweigh the disadvantages at present and is a practical option, we need to launch an open dialogue about what the long term implications could be - whether they are archivally stable, URLs may become unstable over time, the web pages may not be maintained; but, the generality is commercial.

Our study does not assess the installability and reproducibility of the code at the current stage, however, normally, when code and data is shared, there is no conformity if one can install the same pipeline with the exact same dependencies and make sure it produces the same output. It is alarming that multiple studies indicate that it is not very easy to achieve this, and have reported that upto 50% of published research data is not reproducible^{19,20}. The consequences of published, peer-reviewed data not being reproducible are copious. A survey conducted by Nature with 1576 researchers also showed that researchers have on numerous occasions failed to reproduce another published experiment and their own experiments as well¹⁰. In biomedical research, irreproducible experiments or reproducibility errors could lead to massive real-time implications in clinical research and healthcare.

Current efforts mostly rely on individual researchers although many journals have taken the initiative to ensure code and data availability. In these journals, there are two main

ambiguities in the data and code sharing policies. Firstly, some journals mention that all underlying data used for analysis and/or all supporting data must be made available. It is unclear as to whether ‘supporting data’ and ‘underlying data used for analysis’ includes only raw data or raw data and software source code, algorithms, and analytical code. Secondly, for those journals that have mandatory or encouraged code sharing policies, journals only state that sharing software source code or custom code is necessary. It is unclear if the policy includes that all analytical code used must be made available. We have identified the substantial efforts of the biomedical community in the reference section^{4,21–24} which include effective strategies and best practices in enhancing reproducibility and rigor in research. We as a community need to establish dialogue with the interested stakeholders to identify effective strategies to mitigate perceptual and technical barriers in reproducible and transparent data-driven research²⁵. This will foster scientific progress across a variety of biomedical disciplines.

Figures

Figure 1. Availability of data (a) and code (c) across the 200 biomedical articles classified according to the data repository (b) and code repository (d) on which the supporting information was shared.

Figure 2. Data and code sharing policies of journals compared to the actual availability of code and data (a) Data sharing policies of journals (b) Code sharing policies of journals

Figure 3. Trend in sharing data and code availability (a) Improvement in the availability of data from 2016-2021 (b) Improvement in availability of code from 2016 - 2021 (c) Association between code availability and number of citations

Figure 4. Comparison of presence of data and code availability statements and actual availability of data and code

4. Methods

Download and prepare scientific publications for analysis

We downloaded the full text of all publications from the PubMed Central (PMC) open access corpus in Extensible Markup Language (XML) format. These publications were published in 8 unique journals between the years 2016 and 2021. We included the commercial use and non-commercial use subsets. Each publication had a unique PMC identification number (PMC ID). The XML file for each publication was stored in a directory named according to the journal that publication was published in. To avoid ambiguity downstream when working with comma separated value (CSV) files, we deleted commas in these journal directory names.

Filter downloaded data for desired journals

After downloading the entire open access corpus of publications from PMC, we kept only the directories named after the following 8 journals: Nature Biotechnology, Genome Medicine, Nature Methods, Genome Biology, Bioinformatics, BMC Bioinformatics, Nucleic Acid

Research, and Nature Genetics. This brought the entire set of publications down to 12,603 publications.

Extract publish dates of selected publications

We created a key file with the paths to every XML file representing a publication. We then used a Python script utilizing the package ElementTree to systematically extract publication metadata from these XML files. To find the date the publication was published, our script considered each publication date that was listed in a publication's file (there were often multiple). When the year was missing from a date, we disregarded the date completely. When the month was missing, we recorded the publication month as December. When the day was missing, we recorded the publication day as the last day of the publication month. After completing this process with each listed date, the earliest converted date was recorded and associated with our record of the publication.

Analysis of extracted information

We also recorded various parameters such as the raw data availability, where the data is shared, the analytical code availability, code availability in supplementary files, code availability in repositories or webpages, presence of code and data availability statements, and the number of citations for each paper. All the parameters recorded were based on what the authors have mentioned in the publication. Amongst the 480 papers, we ignored the papers that were conducting secondary analyses while looking into raw data availability. When we looked into the raw data availability in each article, we assigned 'Yes' and 'No' to indicate availability and unavailability of raw data. We also recorded whether the data was shared on 'GEO', 'SRA' or

other repositories such as Bioproject, ENCODE, DDBJ, GitHub, ArrayExpress, European Nucleotide Archive (ENA). Those articles which shared data on either DDBJ or ENA were included under ‘SRA’. Some articles had data shared or obtained from multiple datasets for different parts of their study and these articles were included under the category ‘Multiple’. The category ‘Multiple’ not only refers to articles which have split their data on multiple repositories but also could refer to those which have duplicated datasets on different repositories.

Similarly, we recorded the code availability in all the 480 articles and assigned ‘Yes’ or ‘No’ to indicate availability and unavailability of analytical code. The goal of this study was not to check the completeness of analytical code but the presence and incentive of researchers sharing code. We specifically looked for code used for analysis (not to be confused with software or source code for algorithms and tools) and considered code availability as a ‘Yes’ if any analytical code was shared. For those articles that shared code, we recorded if the code was hosted on GitHub, shared in the supplementary materials or hosted other repositories such as Zenodo or a webpage. Some articles that shared code in both supplementary materials and on Github were included under GitHub. Verifying the completeness and/or accuracy of the code shared was beyond the scope of this study.

Additionally, we also recorded whether the articles had data availability statements and code availability statements mentioned in a separate section for easy accessibility. We assigned a ‘No’ if the article did not have a statement but mentioned the availability of data or code amidst the main text. We also recorded the number of citations the articles had on Google Scholar as of 15 November, 2021, and the policies of the ten journals on sharing code and data. We classified

the code and data sharing policies of the journals as ‘Mandatory’, ‘Encouraged/Mandatory’ if it was unclear, and ‘No policy’ for those that did not mention any guidelines for sharing code or data.

5. References

1. Crow, M. M. & Tananbaum, G. We Must Tear Down the Barriers That Impede Scientific Progress. *Scientific American* (2020).
2. Elliott, K. C. & Resnik, D. B. Making Open Science Work for Science and Society. *Environ. Health Perspect.* **127**, 75002 (2019).
3. Gewin, V. *et al.* Data sharing and the future of science. *Nat. Commun.* **9**, 1–2 (2018).
4. Brito, J. J. *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *Gigascience* **9**, (2020).
5. Website.
<https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research>.
6. Chen, X. *et al.* Open is not enough. *Nat. Phys.* **15**, 113–119 (2018).
7. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *Osteoarthritis Cartilage* **20**, 256–260 (2012).
8. Baker, D., Lidster, K., Sottomayor, A. & Amor, S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol.* **12**, e1001756 (2014).

9. Serghiou, S. *et al.* Assessment of transparency indicators across the biomedical literature: How open is open? *PLoS Biol.* **19**, e3001107 (2021).
10. Website.
<https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>.
11. Building the foundation for future research through Open data, code and protocols.
<https://theplosblog.plos.org/2019/12/building-the-foundation-for-future-research-through-open-data-code-and-protocols/> (2019).
12. Buck, S. Beware performative reproducibility. *Nature* vol. 595 151 (2021).
13. Gatto, L. Chapter 9 Bioinformatics.
<https://uclouvain-cbio.github.io/WSBIM1207/sec-bioinfo.html> (2021).
14. Deardorff, A. Why do biomedical researchers learn to program? An exploratory investigation. *J. Med. Libr. Assoc.* **108**, 29–35 (2020).
15. Duvallet, C. Data detectives, self-love, and humility: a research parasite’s perspective. *Gigascience* **9**, (2020).
16. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, (2016).
17. Schulz, S. Faculty Opinions recommendation of The FAIR Guiding Principles for scientific data management and stewardship. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2018) doi:10.3410/f.726216348.793543848.
18. Glenn Begley, C. & Ellis, L. M. Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
19. Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M. & Zwelling, L. A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate

- Findings from the Laboratory to the Clinic. *PLoS One* **8**, e63221 (2013).
20. Samota, E. K. & Davey, R. P. Knowledge and attitudes among life scientists towards reproducibility within journal articles. *Cold Spring Harbor Laboratory* 581033 (2019) doi:10.1101/581033.
  21. Barnes, N. Publish your computer code: it is good enough. *Nature* **467**, 753 (2010).
  22. Stodden, V. *et al.* Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016).
  23. Gezelter, J. D. Open Source and Open Data Should Be Standard Practices. *J. Phys. Chem. Lett.* **6**, 1168–1169 (2015).
  24. Cohen-Boulakia, S. *et al.* Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Gener. Comput. Syst.* **75**, 284–298 (2017).
  25. Cadwallader, L., Papin, J. A., Mac Gabhann, F. & Kirk, R. Collaborating with our community to increase code sharing. *PLoS Comput. Biol.* **17**, e1008867 (2021).

6. Supplementary Figure(s)

Figure S1. Association between code availability and number of citations