A study of data sharing by authors published in scientific journals

Richard M. Wallace IV[1], Kelly M. Elkins[1]*

*[1]Chemistry Department, Forensic Science Program, Towson University, 8000 York Road, Towson, MD 21252, *corresponding author: kmelkins@towson.edu*

Abstract: Generating a research idea and a hypothesis, planning, and gathering materials for an experiment, seeking institutional review board (IRB) and Institutional Animal Care and Use Committee (IUCAC) approval, obtaining funding and support, conducting an experiment and analyzing data, and reporting conclusions and sharing the results with the community are all parts of scientific research. As educators, we often focus on the early steps when conducting research with undergraduates and teaching course-based undergraduate research (CURE) courses. In a classroom setting, we shifted the focus to how research is authored and reported and how data is shared. Journals and funding agencies have instituted new policies for authors regarding data reporting in databases and supplementary information in recent years. In this classroom exercise CURE project, students learned how to conduct a study with IRB approval, write professional queries, collect and analyze data, and report results, as well as important information about the scientific publishing enterprise and data sharing. One of the primary goals of the scientific community is the sharing and spread of information to further advance research and our understanding of the world. Therefore, sharing data between researchers and authors is paramount to the success of our community and to the education of forensic science students. It is as important to teach students the tenets of the scientific method as the process and ethics of sharing data.

Keywords: Data sharing, research, ethics, sequencing, international, CURE, project-based learning, data sharing

.

Introduction

The forensic community is guided by numerous codes of conduct and standards of practice. Ethical decision making and the impacts of moral beliefs are relevant to every field of study and job or professional practice. Whilst the two concepts can often be confused, morals are values of right versus wrong that a community believes and adheres to whereas ethics in the Merriam- Webster online dictionary refers to the "correct behavior within a relatively narrow area of activity" [1]. Ethics guide the actions of the scientific community. One of the primary goals of the scientific community is conducting original research including discovering how the world works and creating new technologies to improve our lives. For scientists to gain acceptance of their work, be credited, and get new technologies in the hands of everyday people, they must share and spread the information. Publicly funded science comes with it an ethical responsibility to report the findings to the public; publicly funded science, when published, often must be published open access so that the public can access the data and report free of charge. Government documents make this clear to grant seekers. When authors withhold data or fail to share data, it becomes difficult for other researchers to advance upon the work and may lead to others unnecessarily repeating experiments that have already been performed. Data sharing is a means to establish reproducibility. Additionally, the lack of shared data sets and research impedes meta-analysis and evaluation and comparison of software tools such as those used in probabilistic genotyping or for data visualization.

Fifteen years ago, in 2009, a study was conducted to determine how common data sharing was among scientists when requested data was requested of authors [2]. Only one in ten researchers shared the requested data in the study [2]. As educators, we often focus on the early steps of conducting research with undergraduates and teaching course-based undergraduate research (CURE) courses. We demonstrate and engage students in the scientific methods including teaching them to generate an idea and a hypothesis, plan and gather materials for an experiment, seek institutional review board (IRB) and Institutional Animal Care and Use Committee (IUCAC)

approval, as required, obtain funding and support, conduct an experiment and analyze data, and report conclusions and share the results via a poster or oral presentation to the community on campus or at a professional conference. While all of the above support students in developing essential skills, what is missing is an introduction to the publication, review, and data sharing process. Frequently there simply is not enough time in a quarter, semester, or degree program to mentor students through publishing papers and teach them about the discourse that follows once a report is published. In a classroom setting, an instructor shifted the focus to the ethics of conducting research, publishing, and data sharing. Journals and funding agencies have instituted new policies for authors regarding disclosure of contributions and requirements of data reporting in databases and supplementary information since the 2009 study [2] was published. Now several funding agencies and journals require that data is deposited in repositories and publicly shared as a condition of funding awards and publication. In a classroom exercise, students mimicked the prior study and queried authors about their study and asked them to share data that was not accessible in public repositories to analyze how data sharing trends have changed. The goal was to teach the ethics in research and publishing, how to conduct a research study, and the life and access of data post-publication to advance science.

Methods

A classroom exercise was conducted modeled on a study conducted by Savage and Vickers in 2009 [2]. The papers selected in this study were selected by the 13 students enrolled in the course who chose to participate in the class project. Each student selected one paper. The papers of interest to the students were written within the past ten years and featured research related to protein sequencing, DNA sequencing, or a similar topic containing sequencing data including forensic applications. The students composed an email query as a group during class, made revisions as suggested by the instructor, and deployed their request on the same day using their university email addresses. In the request, the students identified themselves as Towson University (TU) students working with the instructor. Thereafter, responses were tracked over a 3-week period and sorted depending on whether the author shared the requested data and the reasons surrounding whether they did or not. If an author did not respond to the initial request, a follow-up email was sent as a second request. The instructor demonstrated the IRB approval process with the class. An IRB (TU protocol #2075) was approved and considered exempt for analysis of the data collected in this study.

The thirteen published papers were chosen from various professional, peer-reviewed journals. Of all the research papers surveyed, none included the full raw sequencing data in the paper or supporting data sections. The authors in this study published in a range of different journals and represented different labs and did not overrepresent any single journal or lab group; a broad range of scientists was represented. After the course ended, a student analyzed the class data and tabulated the journals' impact factors, the location of the lab where the research was conducted, the date the paper was published, how many emails were sent before a response was received, and the funding source.

Results

The course content included lectures on best practices in research, IRB, codes of conduct, authorship, what plagiarism is, rules for cropping gels and cleaning sequence data, publication, types of journals including non-profit and for profit and subscription and open access, as well as data sharing.

Emailed requests were sent out to thirteen corresponding authors with the results summarized in **FIGURE 1**. Of the thirteen surveyed, one author had no working email and thus could not be contacted, seven authors did not share their data after several requests, and five authors did share the requested sequence data within the three-week period. This computes to a 38% rate of data sharing.

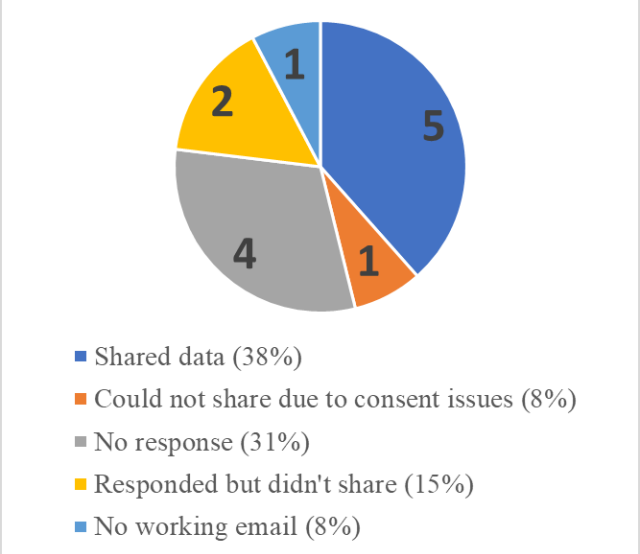

FIGURE 1 Results of data requests in this study.

Of the seven authors who did not share the requested data, one author responded that they were unable to share their data due to consent issues, two authors responded to our initial request email but did not share the requested sequence data, and four authors never responded to our queries. Of the two authors who responded, one initially shared data but it was not the data requested so a follow-

up email was sent but was unanswered. The other author responded saying they may have access to the data requested, but never responded to further follow up emails (**FIGURE 2**).

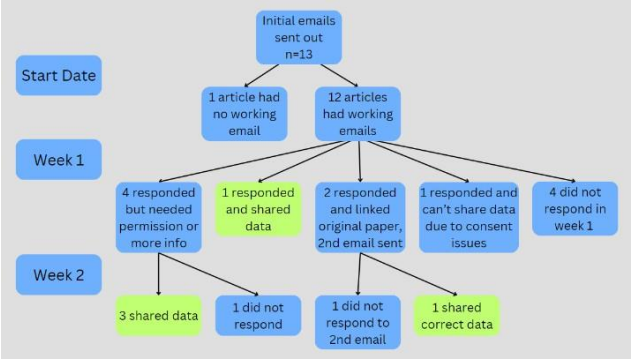

FIGURE 2 Timeline and summary of results from data requests (modeled after Savage and Vickers, 2009 [2])

The data requested was all published in peer- reviewed journals in the past ten years and primarily in the past 5 years. The impact factors of the journals ranged but the papers of interest for which we sought sequence data were primarily published in field specific journals with lower impact factors, but a couple of the journals were of the high impact with an impact factor of high teens and greater than 60.

The labs from which the research originated were scattered worldwide across 11 countries: The United States of America, Denmark, Poland, Finland, Egypt, India, People's Republic of China, Pakistan, Bangladesh, Japan and Australia.

Discussion

These results show a marked increase from the previous study conducted in 2009 where only 10% of respondents shared their data compared to our recorded 38% as shown in **FIGURE 1**. This is an interesting finding as we discuss below.

The classroom exercise was modeled on a previous study published in the *Public Library of Science (PLoS) One* journal in 2009 [2]. The sample size is admittedly small but slightly greater than the number of papers surveyed in the original study (10) [2]. Notably, the Savage and Vickers study in 2009 [2] marked an apparent decrease in data sharing in comparison to a previous study from 2006 [3] in which 27% of authors queried shared their data, although this could be attributed to variation in sample size as the 2006 study surveyed 249 studies whereas Savage and Vickers surveyed 10 [2,3]. A 2021 study found that scientists reported sharing their data upon request roughly 40-60% of the time [4]. Although this comes from personal survey response rather than concrete proof of shared data, it is interesting to note that our results are congruent with the most recent 2021 study [4]. Again, the sample size disparity (our sample size was small) could be a possible explanation for the variation in percent sharing. Taken as a whole, the recent studies display a general increase in data sharing over time (**TABLE 1**).

TABLE 1 Results of selected data sharing studies over the past several years

Year	% Data Shared	Author(s)
2006	27%	Wicherts et al.
2009	10%	Savage & Vickers
2021	40-60%	Hrynaszkiewicz et al.
2023	38%	This study

The corresponding authors contacted had published in well-known journals published by widely regarded as reputable publishers including the American Chemical Society (ACS), Elsevier, and Springer. Each published journal has a set of guidelines specifically regarding data sharing policies. For example, ACS journals strongly endorse data sharing and making data open access upon publication and defines data as "materials and information used in the experiments that enable the validation of the conclusions drawn in the article" [5]. ACS also employs different data policy levels depending on which specific journal the author intends on publishing under for specific data sharing requirements and requires all authors to include a Data Availability Statement that describes the availability of the data in the publication. The journals the authors were queried published in were level 1 journals meaning that data sharing is not required and only the Data Availability Statement needs to be provided as a condition of publication at the time of this writing. Some of the publications queried were published in Elsevier journals which, at the time of this writing, encourage authors to share their data and interlink data into the paper when appropriate but do not require it. One of the journals had an expanded statement that included sharing of code, software, models, methods, and other materials the author used to validate their findings. A couple of the papers were published in a Springer journal which encourages authors to share their data and encourages the inclusion of a statement of data availability but requires neither.

The final journal that a queried author published in listed no guidelines for data sharing policies on their website. The journal staff was contacted on whether they had any data sharing policies that were unlisted and, although they responded to the email, they did not give any specific response.

The fact that data sharing is not a requirement by any of the journals in which the papers were published could be a reason for the lower data sharing rate than the 2021 study [4]. However, this study and the 2021 study [4]

shows marked improvement in data sharing since the 2006 and 2009 studies [2, 3] in only a short three-week period. This demonstrates a shift in research culture, both when the data was self-reported and independently collected, possibly due to newer grant funding and publication guidelines that have emerged in the past fifteen years. Nevertheless, there is still work to be done. It is of concern that so many prominent publishers have not yet considered data sharing a necessary part of the research and dissemination process. Understandably, while some studies' raw data cannot be shared due to case or medical data confidentiality and privacy concerns or dual-use research concerns, most studies do not fall into these restrictions.

Moving forward, journals should strive to strengthen data sharing policies and require specific data sharing and practice including potential accountability measures such as risk of not publishing accepted papers if the data is not disseminated. As the journals the authors published in did not have data sharing requirements and we received data from 38% of authors, it is apparent that several of the authors feel strongly that data sharing is important and there has been growth in willingness to share data over the past fifteen years.

The project had several educational outcomes. The students actively collaborated on the research study in class and investigated a topic they had learned about in class (e.g., data sharing). The students gained practice with professional communication and writing. The data served as the basis for a research project from one student who performed the data analysis, created the charts and graphs, and wrote the first draft of this paper. Two students presented this project at a campus presentation for a regional meeting and at a national professional association meeting; one student wrote the abstract for the presentations while the other student performed the data analysis and wrote the draft of this manuscript, as described.

The concept that researchers could ask for additional details or data that was not provided but on which the study conclusions were based was new to the students as well as the previous research on data sharing. In performing this study, the students learned many new skills. The students learned about the role of the corresponding author (author) and explicitly how to contact them and what responses they might receive. Most of the students struggled to write an email to the author and how to ask for the data they sought. Instruction and guidance were provided by the instructor on professional writing. Each week when the responses were received, they were shared and discussed.

The course instructor has experience creating and teaching CURE courses and this was not traditionally a course with a CURE component. Nonetheless, it demonstrates an example of a CURE class project in a regular lecture course on ethics required for forensic students at our university. While the students did not conceive of the project, they did identify research that they were genuinely interested in and selected the papers with the instructor's approval. The students sought the data and collected and recorded the responses from the authors.

As for many undergraduates that have been exposed to research in CURE courses, one student from the class enrolled in undergraduate research in subsequent semesters and continued the work as part of a capstone experience. This was the student's first exposure to research and the student was not planning to perform research. Thus, the course had the outcome of increasing student exposure and continuing research. A few of the students enrolled in the course were already conducting lab research with other faculty as at least a couple enrolled in additional research experiences after the course.

Because of the time and effort entailed in clarifying the data request, finding, and sharing research data, this study may not be suitable to be repeated often. However, it demonstrates how project-based research can be incorporated into a course that typically has not had a research component. It also serves to evaluate data sharing as demonstrated by this exercise and previous studies. Nevertheless, a moral question arises of whether these studies should be conducted at all. Here, some of the data requested in this study was of interest for research in the instructor's laboratory and some was of interest for research projects or papers students were working on for other courses. Other papers were those students were interested in. This may have lessened the moral issue or introduced bias toward certain data or topics. Data is shared to advance science and research and time is precious. As a practice, students and researchers should only make research data requests to support their undergraduate, graduate, or postgraduate research or ongoing faculty research.

Conclusion

The frequency of data sharing is an often-overlooked issue amongst scientists but remains to be one of paramount importance. Data sharing is a requirement for the adequate spread of information that the scientific community needs to continue to grow and evolve. In this classroom exercise, undergraduate students engaged in a research project, the first of such experiences for most of them. Notably, CUREs and project-based learning provided students a framework to practice hypothesis and data driven research, analysis, reporting and writing. Even though less than half of authors queried shared the relevant sequencing data upon request, the results do demonstrate a marked increase in comparison to the previous 2009 study. The percent sharing may have been higher if the data was recorded over a longer time period.

Data requests cost researchers time that is not supported by grants and is outside of their teaching and service responsibilities. While our sample size was small, our results are congruent with similar studies in recent years. Although the subject of some studies can make data sharing difficult or impossible, in general journals could attempt to implement stronger policies in an effort to improve data sharing trends. Likewise, funders need to provide a mechanism to support data sharing on websites and to the community after the original support has ended. So, whilst there is still room for improvement, the increasing trends in data sharing mark a positive outlook for future scientists and students who seek to engage in advancing research.

Acknowledgements

The authors thank the students in the class for their queries and sharing their data analyzed in this manuscript as well as two anonymous reviewers that provided extremely helpful suggestions for improving this manuscript.

References

1. Merriam-Webster. (n.d.). Ethics vs morals. In Merriam-Webster.com dictionary, https://www.merriam-webster.com/dictionary/moral#:~:text=While%20ethics%20can%20refer%20broadly,the%20ethics%20of%20genetic%20testing. (accessed June 6, 2024).
2. Savage CJ, Vickers AJ. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS ONE 2009;4 (9):e7078. https://doi.org/10.1371/journal.pone.0007078.
3. Wicherts JM, Borsboom D, Kats J, Molenaar D. The poor availability of psychological research data for reanalysis. Am Psychol 2006;61:726–728. https://doi.org/10.1037/0003- 066X.61.7.726.
4. Hrynaszkiewicz I, Harney J, Cadwallader LA. Survey of Researchers' Needs and Priorities for Data Sharing. Data Science Journal 2021;20:31. https://doi.org/10.5334/dsj-2021-031.
5. ACS Research Data Policy, https://publish.acs.org/publish/data_policy (accessed Nov 21, 2023).
6. Guide for authors - Forensic Science International, https://www.sciencedirect.com/journal/forensic-science-international/publish/guide-for-authors (accessed Nov 21, 2023).
7. International Journal of Legal Medicine. Springer. https://www.springer.com/journal/414/submission-guidelines#Instructions%20for%20Authors_Research%20Data%20Policy%20and%20Data%20Availability%20Statements (accessed Nov 21, 2023).