Reproducibility, Fabrication, and Falsification

Joanna F. DeFranco, Penn State Great Valley

Jeffrey Voas, IEEE Fellow

Research data validation is challenging and expensive. If not addressed, falsification and fabrication of data are encouraged, causing disinformation directed at the public.

Disinformation is not only rampant on the Internet and social media; it can also be found in scientific research publications. Occurrences in research are more apparent when research or study data are irreproducible. A Harvard researcher resigned after a fraud discovery occurred.[1] And one study concluded 33.7% of scientists surveyed admitted to questionable research practices at least once in their career.[2]

Source data validation is necessary for research—especially funded research. The cost of source data validation is estimated to be between 20% and 30% of an overall clinical trial budget.[3] However, what stops someone from simulating, tampering, or falsifying raw data to deliver a desired result to support a "desired" study hypothesis? If data can be easily fabricated and falsified, is source data validation worth the costs? In addition, falsification may not be the only problem here; withholding data is another.[4]

IEEE has attempted to help here by providing a utility for researchers, named "IEEE Dataport" (https://ieee-dataport.org/). This repository offers researchers free data uploads and access of up to 2 TB. This utility is not only beneficial by having research data stored at a trusted organization but data sets may also be connected to IEEE journal and magazine articles. This increases data and research visibility. The Dataport utility also assists researchers 1) in meeting funding agency data management requirements, 2) in facilitating possible collaboration opportunities with data set owners, and 3) by

Digital Object Identifier 10.1109/MC.2021.3055926
Date of current version: 17 November 2021

0018-9162/21©2021IEEE

EDITOR IN CHIEF **JEFFREY VOAS**
IEEE Fellow; j.voas@ieee.org

offering other benefits related to transparency. Most importantly, this offering should support reproducible research, a topic that *Computer* will discussing more in future issues. IEEE Dataport currently has almost 700,000 users and over 1,500 data sets.

It is important to keep in mind that data reproducibility can be a challenge because of improper research techniques where, for example, researchers look for data correlations until they find a bizarre outlier and then claim its statistical significance. Here, they could employ improper statistical techniques or change variables/combine data sets—thus invalidating the research/study data and its results.[5] This is another "pro" argument as to why to maintain and use a data repository because it also might spin off a learning community. In a learning community, outsiders not associated with the creation of the original data could request access to data sets to then test research outcomes and offer peer-reviewed improvements in a data owner's experimental techniques. It could also discourage data tampering and falsification.

Proper research data validation is paramount. Disinformation directed at the public through social media and questionable/debatable research results

IN THIS ISSUE

Computer receives a fair number of submissions relating to software development, software engineering, composing systems from components, testing and validation, and requirements elicitation. Because these topics remain a technical mainstay of the IEEE Computer Society, we even launched a new software engineering column in *Computer* this year.

This December 2021 issue features four articles that are related to software engineering and software composition and have been waiting for publication, and I'm pleased to finally release them to you.

In the first article, "Agile–CMMI Alignment: Contributions and To-Dos for Organizations," the authors posture that the Capability Maturity Model Integration (CMMI) and Agile can be aligned. They argue that the new CMMI V2.0 model provides an evolutionary capability roadmap to achieve business value while being flexible enough and applicable to approaches such as Agile. The article discusses how agile organizations wishing to align with CMMI V2.0 might begin that process. This article uncovers a set of issues that organizations should first consider and provides suggestions on how to leverage the resources that CMMI V2.0 provides to address these challenges.

In the second article, "Compositional Thinking in Cyberphysical Systems Theory," the authors focus on how to engineer safer and more secure cyberphysical systems. They argue that engineering safer and more secure cyberphysical systems requires system engineers to develop and maintain both static and dynamic model views. Their key point is that by verifying the composition of requirements, behavioral, and architectural models using category theory, this can assist in the modeling and analysis of safety-critical cyberphysical systems.

In the third article, "When Scientific Software Meets Software Engineering," the authors focus on scientific software development. They argue that the success of scientific software development depends on the specific computer languages employed. More specifically, the more general purpose a language is, the more flexibility it will provide; however, more rigorous engineering principles and validation and verification activities will be required. Their article aims to raise awareness among scientists, engineers, and language creators of their shared responsibility in developing more reliable scientific software.

In the final article, "Blockchain-Based Software Architecture Development for Service Requirements With Smart Contracts," the authors focus on how smart contracts can provide advanced and flexible development of distributed ledger applications. Their article is a survey on the progress of research into smart contracts, and they elaborate on the existing classification and compilation mechanisms of smart contract languages. They discuss how smart contracts form the foundation of blockchain 2.0. And their article illustrates one smart contract language and its compilation, contract deployment mechanism, and contract execution process.

In summary, I hope you enjoy this issue.

—Jeffrey Voas, Editor in Chief

Digital Object Identifier 10.1109/MC.2021.3107004
Date of current version: 17 November 2021

must be fought against. In the medical field, when research fails to be validated and is retracted, it may affect other connected research studies that were based on those retracted results. For example, *The Lancet* and *New England Journal of Medicine* recently retracted a study stating that hydroxychloroquine had no benefit to treating COVID-19 and in fact the retraction comments suggest that hydroxychloroquine could increase risk of death.[6]

By improving data validation techniques, we would hope that this would tease out poorer quality research and discourage unethical behavior. But what are potential solutions moving forward?

Research is occurring that is looking into the benefit of using blockchain to validate research and the argument is that blockchain should be able to increase transparency and increase visibility among multiple organizations.[7] Other research is focusing on artificial intelligence and machine learning techniques to validate research data.[8]

So, what else can the research community do to mitigate this concern?

- The research community should add incentives for researchers, such as allowing data to be an additional citable research product. (Note: IEEE's Dataport formulates data citations.)
- The community should require that research data is part of the publication submission process: collect, review, access, and archive the data artifacts.

DISCLAIMER

The authors are completely responsible for the content in this message. The opinions expressed are their own.

However, while all of this sounds great in theory, the extra time to do it slows down the process of information dissemination. So that is a tradeoff that cannot be quickly dismissed. And it probably depends on the criticality of the results being published; for example, medical results should be more critical than others.

In summary, the fabrication and falsification of research data are not new. But it is an increasingly challenging problem needing attention by the research community. If this is not acknowledged, a continuation of employing falsified and fabricated data for political advantage, graduation, employment promotion, and so on, will likely be a social and safety challenge going forward.

REFERENCES

1. S. Carpenter, "Harvard psychology researcher committed fraud, U.S. investigation concludes," *Science*, Sept. 2012. [Online]. Available: https://www.sciencemag.org/news/2012/09/harvard-psychology-researcher-committed-fraud-us-investigation-concludes. doi: 10.1126/article.26972.
2. D. Fanelli, "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data," *PloS One*, vol. 4, no. 5, p. e5738, 2009. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005738. doi: 10.1371/journal.pone.0005738.
3. H. Li, L. Zhu, M. Shen, F. Gao, X. Tao, and S. Liu, "Blockchain-based data preservation system for medical data," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–13, 2018. doi: 10.1007/s10916-018-0997-3.
4. T. Nugent, D. Upton, and M. Cimpoesu, "Improving data transparency in clinical trials using blockchain smart contracts," *F1000Research*, vol. 5, p. 2541, Oct. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5357027.1/#ref-1. doi: 10.12688/f1000research.9756.1.
5. D. Randall and C. Welser, "The irreproducibility crisis of modern science: Causes, consequences, and the road to reform," National Association of Scholars, New York, 2018. [Online]. Available: https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science/full-report
6. E. Edwards, "The Lancet retracts large study on hydroxychloroquine," NBC News, 2020. https://www.nbcnews.com/health/health-news/lancet-retracts-large-study-hydroxychloroquine-n1225091
7. A. Andrianov and B. Kaganov, "Blockchain in clinical trials-the ultimate data notary," *Appl. Clin. Trials*, vol. 27, no. 7/8, pp. 16–24, 2018.
8. "How AI is tackling fake academic research that is plaguing scientific community," *Analytics India Magazine*, 2018. https://analyticsindiamag.com/how-ai-is-tackling-fake-academic-research-that-is-plaguing-scientific-community/

JOANNA F. DEFRANCO is an associate professor of software engineering at The Pennsylvania State University, Malvern, Pennsylvania, 19355, USA. Contact her at jfd104@psu.edu.

JEFFREY VOAS, Gathersburg, Maryland, USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.