All Together Now: Data Work to Advance Privacy, Science, and Health in the Age of Synthetic Data

Lindsay Fernández-Rhodes

College of Health and Human Development, Department of Biobehavioral Health; Social Science Research Institute; Population Research Institute; Clinical and Translational Science Institute; 219 Biobehavioral Health Building, 296 Henderson Drive, Pennsylvania State University, University Park, PA 16802, USA
Email: fernandez-rhodes@psu.edu

Jennifer K. Wagner

School of Engineering Design and Innovation; Department of Anthropology; Department of Biomedical Engineering; Institute for Computational and Data Sciences; Huck Institutes of the Life Sciences; Rock Ethics Institute; Pennsylvania State University, University Park, PA 16802 USA and Penn State Law, University Park, PA 16802 USA
Email: jkw131@psu.edu

There is a disconnect between data practices in biomedicine and public understanding of those data practices, and this disconnect is expanding rapidly every day (with the emergence of synthetic data and digital twins and more widely adopted Artificial Intelligence (AI)/Machine Learning tools). Transparency alone is insufficient to bridge this gap. Concurrently, there is an increasingly complex landscape of laws, regulations, and institutional/ programmatic policies to navigate when engaged in biocomputing and digital health research, which makes it increasingly difficult for those wanting to "get it right" or "do the right thing." Mandatory data protection obligations vary widely, sometimes focused on the type of data (and nuanced definition and scope parameters), the actor/entity involved, or the residency of the data subjects. Additional challenges come from attempts to celebrate biocomputing discoveries and digital health innovations, which frequently transform fair and accurate communications into exaggerated hype (e.g., to secure financial investment in future projects or lead to more favorable tenure and promotion decisions). Trust in scientists and scientific expertise can be quickly eroded if, for example, synthetic data is perceived by the public as "fake data" or if digital twins are perceived as "imaginary" patients. Researchers appear increasingly aware of the scientific and moral imperative to strengthen their work and facilitate its sustainability through increased diversity and community engagement. Moreover, there is a growing appreciation for the "data work" necessary to have scientific data become meaningful, actionable information, knowledge, and wisdom–not only for scientists but also for the individuals from whom those data were derived or to whom those data relate. Equity in the process of biocomputing and equity in the distribution of benefits and burdens of biocomputing both demand ongoing development, implementation, and refinement of embedded Ethical, Legal and Social Implications (ELSI) research practices. This workshop is intended to nurture interdisciplinary discussion of these issues and to highlight the skills and competencies all too often considered "soft skills" peripheral to other skills prioritized in traditional training and professional development programs. Data scientists attending this workshop will become better equipped to embed ELSI practices into their research.

Keywords: bioethics, data privacy, data work, health research, synthetic data

1. Introduction

The breadth of this workshop is deliberate, intended to bring together scholars from diverse areas of expertise and to promote interdisciplinary understandings foundational to the development and use of digital twins for biomedical research.[1] This effort responds to growing recognition of need for an interdisciplinary workforce prepared to seize opportunities and overcome challenges for digital twins, exemplified by the recent recommendations of the National Academies of Sciences, Engineering, and Medicine.[2,3]

The title of this workshop—*All Together Now: Data Work to Advance Privacy, Science, and Health in the Age of Synthetic Data*—itself is significant, with multiple levels of meaning to shine light on areas in which biocomputing can be enhanced. The reference to "all together now" refers not only to the importance of interdisciplinary, multidisciplinary, transdisciplinary collaboration but also to participatory, community-engaged research and collaborative governance. "Data work" draws attention to recent anthropological scholarship[4,5] as well as the novel biocomputing approaches that are now possible (e.g., synthetic data and digital twins, see, e.g., Foraker et al.[6], Moore et al.[7]). The explicit mention of "privacy, science, and health" is intended to draw attention to three distinct but interconnected international human rights (Articles 12, 27, and 25(1) of the Universal Declaration of Human Rights, respectively) that underlie ongoing debates about AI governance around the world and influence Fair Information Practice Principles (FIPPs), FAIR guiding principles for scientific data,[8] CARE principles[9] and more. The scope of this workshop is further intended to help attendees situate their biocomputing research more deliberately within the revised National Institute on Minority Health and Health Disparities (NIMHD) research framework for digital health equity.[10]

In this workshop, we will begin by highlighting scholars who routinely utilize synthetic data, digital twining, 'fake' data, simulations, or other obfuscation of data to ensure data privacy. They will present on how this limits the utility of data and/or their explainability. To address the tension these data present for public engagement, we will have community, implementation scientists, and communication scholars present on best practices on how these new data technologies can (and should) be incorporated into community engagement activities, to ensure that all populations have access to these new scientific approaches and insights. Workshop attendees will learn approaches to manage the gap between a) public expectations of science or assumptions of how biocomputing is performed, and b) the reality of the modern healthcare system, methodologic innovations within the biomedical research data ecosystem.

2. Workshop Topics and Presenters

2.1. *Opportunities and Challenges of Synthetic Data, Digital Twins, and Data Governance*

2.1.1. *Expanding Information Accessibility through Synthetic Data*

Presented by: Randi Foraker, PhD, MA, FAHA, FAMIA, FACMI (University of Missouri)

Synthetic healthcare data allow informaticians, data scientists, and clinicians to unlock siloed data and provide access to clinical researchers and consumers (e.g., students, citizen scientists) for improving the health of patients. The core benefit of synthetic data in medicine is that they can address obstacles to rapid research, methods development, and data sharing by representing the trends and relationships in the data without exposing the individual patients, and data — and therefore knowledge — can be shared while protecting individuals' privacy. This talk will explore the current and future state of synthetic data, highlighting its ability to support data sharing, to address privacy and confidentiality, and to advance national and international initiatives. The presenter will share their own work with synthetic data, which spans statistical validation (comparing results of analyses between real and computationally derived data); national and international research partnerships; and leveraging synthetic data for informatics, biostatistics, and data science education.

2.1.2. *The Role of Synthetic Data in Patient Privacy, Healthcare, and Biomedical Research*

Presented by: Jason Moore, PhD, FACMI, FIAHSI, FASA (Cedars-Sinai Medical Center)

Paramount to healthcare and biomedical research is the protection of patient privacy and the security of their data. Synthetic data may address these concerns by providing artificial data points that preserve the correlation structure and patterns of the original patient data. The presenter will review artificial intelligence methods the generation of synthetic data and their use in clinical and biomedical research. These will include deep learning and large language model approaches. They will highlight several use cases from the literature and will discuss the use of synthetic data for creating digital twins that might improve the prediction of clinical outcomes. Limitations and challenges of these methods will be discussed.

2.1.3. *The Long View on Emerging Data Science Technologies*

Presented by: Anjali Deshmukh, MD, JD (Georgia State University College of Law)

This talk will examine emerging data technologies in children, focusing on privacy, longitudinal impacts, and FDA regulation. Real world data of children's health outcomes are difficult to obtain and analyze, and technologies including digital twins have the potential to solve the current limitations of claims data. Yet, the potential benefits must be considered against the risk. Current data policy choices will impact children's privacy rights and drug safety over the long-term. Therefore, understanding current FDA regulations and creating regulatory flexibilities to optimize outcomes for children over their lives is important.

2.1.4. *Technical Approaches to Balance Patient Privacy and Shared Analytic Utility*

Presented by: John Wilbanks (Astera Institute)

National biobanks such as the UK Biobank, the All of Us Research Program, and similar emergent state-level efforts around the world hold the promise of driving novel research on large, diverse participant cohorts. Many such biobanks simultaneously center ideals of aggressive participant empowerment and Open/FAIR science, which can create tensions between protecting an individual patient's privacy while seeking thousands of researchers to generate analysis and insights. Some privacy-enhancing technology approaches can enable multi-party computation, or create synthetic data sets, which can introduce other tensions between data availability and data trustability. This talk will explore how intentional choices in cloud architecture can address these tensions, with specific examples drawn from the All of Us Researcher Workbench and the Broad Institute's Data Science Platform.

2.2. *Data Work from the Perspective of Scholars in Community Engagement, Ethics, and Science Communication*

2.2.1. *Biorepositories and Group Harm: A Choice Architecture for Researchers*

Presented by: Meg Doerr, MS LGC (Sage Bionetworks)

This talk will help workshop participants (1) distinguish between individual and group harm from research; (2) appreciate why group harm should be a primary consideration of AI researchers and those that enable AI research including data access committees, ethics boards, and funders; (3) renew their understanding of current regulations on individual and group harm in research; and (4) learn about new tools (created in a project funded by the Robert Wood Johnson Foundation) to aid researchers, data access committees, ethics boards, and funders in enabling responsible AI-driven research.

2.2.2. *Nothing About Us, Without Us Leading*

Presented by: Maile Tauali‘i, PhD MPH (Hawaii Permanente Medical Group)

Indigenous Peoples are often the target of research and not the owners of the research process. We are also domestically dependent under nations and are often subjected to rules and decisions made about us and not with us. So, when we speak about “own-voice” research, we are speaking in opposition to colonial settler science which subjects us to decisions made without us. We want our voices heard. Learning objectives: 1) Participants will be able to identify 3 laws that uphold Indigenous ownership of data 2) Participants will have 3 strategies to respectfully engage with Indigenous Peoples 3) Participants will learn 3 examples where Indigenous Peoples rights were violated by scientists.

2.2.3. *Bounded Justice, the Performance of Trust, and Anti-Racism in Biocomputing*

Presented by: Melissa C. Creary, PhD, MPH (University of Michigan)

Drawing upon her prior published works on bounded justice, the public performativity of trust[11] and the application of anti-racism in informatics,[12] the speaker will discuss best practices for fostering community engagement while simultaneous embracing new data practices for biocomputing.

2.2.4. *Communication Data, Communicating Science*

Presented by: Jasmine McNealy, PhD, JD (University of Florida)

The hallmark of any interaction between scientists and the public is communication. Communication is important for developing and sustaining relationships, for building trust, and enhancing partnerships. Effective communication is important for interactions with marginalized and/or vulnerable communities, especially those whose distrust of biomedical researchers is born of past missteps and harmful programs. Therefore, scientists should be able to communicate with both media and publics beyond those connected to academia and scholarly research. This is particularly important for helping the public to understand novel data practices.

2.2.5. *Soulful Innovation: A New Framework to Create Responsible Technologies of the Future*

Presented by: Samira Kiani, MD (University of Pittsburgh)

The speaker proposes a new framework for innovation by revisiting our relationship with ourselves, our relationship with the impact we create, the spaces in which innovation happens and our collective. Through this framework– called “soulful innovation”--we ask how we can move away from the culture of “be first” and “star is born” to a culture that celebrates our “collectiveness” and puts inner human values at the core of innovation.

3. Conclusion

The workshop will conclude with a discussion panel facilitated by the organizers involving all of the workshop presenters to what they would like to see from data scientists who use synthetic data in the near future and to address questions and comments from workshop attendees.

By attending in this workshop, participants will gain 1) expertise in understanding how new data technologies that use obfuscation are being implemented in the biomedical sciences, 2) awareness of the potential opportunities and concerns related to these practices with respect to participant and community engagement, and 3) familiarity with the best practices for fostering community engagement and science communication while simultaneously embracing these new data practices.

4. Acknowledgments

We would like to the acknowledge the Penn State Social Science Research Institute, Population Research Institute, and the Department of Biobehavioral Health for their support of LFR.

References

1.National Academies of Sciences, Engineering, and Medicine. Opportunities and Challenges for Digital Twins in Biomedical Research: Proceedings of a Workshop–in Brief. *The National Academies Press* (2023) doi:https://doi.org/10.17226/26922.
2.National Academies of Sciences, Engineering, and Medicine. *Briefing Slides: Foundational Research Gaps and Future Directions for Digital Twins*. https://nap.nationalacademies.org/resource/26894/briefing-slides-digital-twins.pdf (2023).
3.National Academies of Sciences, Engineering, and Medicine. *Foundational Research Gaps and Future Directions for Digital Twins*. https://doi.org/10.17226/26894 (2024).
4.Fiske, A., Prainsack, B. & Buyx, A. Data Work: Meaning-Making in the Era of Data-Rich Medicine. *J Med Internet Res* **21**, e11672 (2019).
5.Fiske, A., Degelsegger-Márquez, A., Marsteurer, B. & Prainsack, B. Value-creation in the health data domain: a typology of what health data help us do. *Biosocieties* 1–25 (2022) doi:10.1057/s41292-022-00276-6.
6.Foraker, R., Mann, D. L. & Payne, P. R. O. Are Synthetic Data Derivatives the Future of Translational Medicine? *JACC Basic Transl Sci* **3**, 716–718 (2018).
7.Moore, J. H. *et al.* SynTwin: A graph-based approach for predicting clinical outcomes using digital twins derived from synthetic patients. *Pac Symp Biocomput* **29**, 96–107 (2024).
8.Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
9.Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* **8**, 108 (2021).
10.Richardson, S., Lawrence, K., Schoenthaler, A. M. & Mann, D. A framework for digital health equity. *NPJ Digit Med* **5**, 119 (2022).
11.Creary, M. & Gerido, L. H. The Public Performativity of Trust. *Hastings Cent Rep* **53 Suppl 2**, S76–S85 (2023).
12.Platt, J. *et al.* Applying anti-racist approaches to informatics: a new lens on traditional frames. *J Am Med Inform Assoc* **30**, 1747–1753 (2023).