partners with researchers to grant them access to large and diverse datasets and compute resources that can enable them to conduct novel and impactful research in various subfields of healthcare, such as epidemiology, genomics, and clinical trials. We're pleased to see these projects adopting best practices like on data governance and data documentation (Gebru et al. 2021).

We are excited about the potential of these projects to improve the state of the art and the state of the practice of LLMs in the healthcare domain. However, we also recognize that there are still many open and important questions about the impacts and implications of using these systems in real-world settings, such as their effects on the quality and equity of healthcare delivery, their risks and challenges for privacy and security, and their ethical and legal ramifications. We encourage both projects to also provide datasets that can help researchers and practitioners to address these questions, to better understand the realized impacts when models are deployed in the real world.

References

Brundage, Miles, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike, and Pamela Mishkin. (2022). "Lessons Learned on Language Model Safety and Misuse." OpenAI. https://openai.com/blog/language-model-safety-and-misuse/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. "Improving Alignment of Dialogue Agents via Targeted Human Judgements." arXiv preprint. arXiv: 2209.14375.

OpenAI. 2018. "OpenAI Charter." https://openai.com/charter/

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv preprint. arXiv: 2203.02155.

Comment Judy Gichoya

Introduction

In their chapter, Mullainathan and Obermeyer utilize a clinical use case for predicting sudden cardiac death from electrocardiograms (ECGs) to

Judy Gichoya is an associate professor in the Department of Radiology and Imaging Sciences at Emory University School of Medicine.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges/comment-health-data-platforms-gichoya.

show the role of health data platforms in research and product development. Their clinical case selection is interesting—because sudden cardiac death (SCD) occurs in less than 1 percent of the population, and most deaths occur in low-risk patients, which makes mitigation difficult.

One View Is No View

Current acquisition for health data occurs primarily in healthcare delivery institutions, which often represents a snapshot of patients suffering from an ailment. These data tend to be blind to wellness—yet increasingly these "well" data are continuously collected by commercial entities. For example, in the case of SCD, I noted that smart watches like the Apple Watch collect oxygenation levels, pulse rate, and limited form of ECGs in many patients at various states including young, old, and healthy (Perez et al. 2019; Seshadri et al. 2020; Marcus 2020). A different example is the smart pumps for breastfeeding, which collect and track various breast milk amounts and variations with time of data. Obermeyer agrees that commercial vendors have data that are missing from data platforms, but further notes that such data are limited in their usefulness because they are not linked to patient outcomes. In the case of ECG detection of atrial fibrillation or abnormal rhythms, the commercial vendor is usually unaware of the patient outcome including interventions that occur in the hospital. While this may be the case today, it should not surprise the community when the status quo changes and the commercial vendors purchase clinical data to link and enrich their datasets. Incentives to stimulate public–private partnerships and also include citizen science (who may have ability to download and share their own data) could encourage collaborative work to enrich available health datasets.

Justifying the Need for Health Data Platforms

Obermeyer describes two platforms—the Nightingale open science platform (Mullainathan and Obermeyer 2022) and Dandelion—a commercial data platform for AI development. Dandelion currently supports five large health systems that provide all the raw clinical data and in turn receive clean and structured data. Dandelion provides these data to various stakeholders with a strict focus on products that benefit patients and also shares revenue back to the participating health system. In this paper we observe two variations of health data platforms—one that is problem/dataset driven (Nightingale), while the second one is process driven with a larger amount of data provided through a single contract.

Health data platforms overcome the barriers for data access by shortening the duration to access new data (which usually takes years and is rarely successful); democratizing data access beyond researchers and institutions with more resources; providing a unified data management process including data use agreements; supporting technological advances including cloud integration; and serving as a safe harbor for datasets to be continually improved to prevent dataset expiry. Despite the huge promise of data platforms, their business models rely on locking participants into the platform with no interoperability. This is challenging as these platforms are relatively new in their development and adoption, and not one platform fits multiple end user needs. Moreover, these platforms are owned by startups with high failure rates, which poses a theoretical risk that a selected platform may not be in use in the future.

In organizations with no management structure for interacting and developing these relationships, I anticipate new job titles and managerial units will develop to provide funding and governance and serve as a liaison to ensure compliance and shared benefits flow back to organizations. Cost estimation will remain difficult due to lack of transparency for individual platform components to allow comparison between onsite and cloud servers. As noted by most participants, high infrastructure cost is a barrier to use of healthcare data platforms. A one-year review of our program shows an estimated cost of $100,000 for cloud computing compared to $60,000 for onsite GPU servers. Today, institutions lack capacity in the office of technology transfer to deal with intellectual property and ownership of innovations developed from the shared data.

Incentives Matter

In addition to data sharing incentivized for social good, funders like the National Institutes of Health are mandating data sharing for all funded research. In countries where there is a single payer/universal health care like in Canada, mandates to share all data are easy to enforce compared to multipayer systems. For such cases, health data platforms can be easily harmonized and data easily linked to other data sources like death registries. In a competitive marketplace, commercial partners like insurance providers have enough capital to purchase other data sources and link to the datasets, and these are limited in access. It is important to note that provision of incentives does not equal availability of valuable data, and the process of making data machine learning ready cannot be understated.

Justifying the Status Quo

In 2017, *The Economist* described data as the new oil based on how much data was generated and the combined market forces of the big technology companies in the United States. Coupled with numerous startups working in this space and large venture capital investments for health AI, it should not be a surprise that organizations realized the value of their data, and were reluctant to share it. As the hype has settled and reality set in about the expensive nature of data curation and lack of infrastructure to process and share data easily (as most healthcare organizations do not use cloud solutions and maintain onsite data infrastructure), organizations no longer prioritize data sharing. Legal consequences of data breaches and penalties in the background of ongoing debate that data can never be fully anonymized curb the enthusiasm of organizations to share data. Lastly, a conflict in values—where healthcare organizations are seen to serve the public good versus "evil profit companies" who want to capitalize health data—presents ongoing discourse. Public perception on how data are shared will get worse as more lawsuits and class actions arise when data are not used for the purposes they were intended for, as is the case of the National Health Service in the UK and DeepMind (BBC News 2021).

Concluding Thoughts

To harness the potential of AI for improving healthcare outcomes and reduce costs, data must be democratized and made accessible to researchers and industry. Health data platforms lower this barrier through streamlined data access (agreements and contracting) and improved data quality through curation efforts that provide machine-learning-ready datasets. As organizations decide on which health data platform to adopt, it is important for them to understand the sustainability of the selected platform including the contractual agreements that limit data migration and platform interoperability. To effectively use health data platforms, organizations must develop business units with competency in compliance, data science, finances, intellectual property, and legal expertise of data use agreements. At the society level, incentives must be aligned to promote data sharing and public private partnerships that provide a view on the health side of the patients who interact with healthcare systems.

References

BBC News. 2021. "DeepMind Faces Legal Action over NHS Data Use." October 1, 2021.

The Economist. 2017. "The World's Most Valuable Resource Is No Longer Oil, But Data." May 6, 2017.

Marcus, G. M. 2020. "The Apple Watch Can Detect Atrial Fibrillation: So What Now?" *Nature Reviews Cardiology* 17: 135–36.

Mullainathan, S., and Z. Obermeyer. 2022. "Solving Medicine's Data Bottleneck: Nightingale Open Science." *Nature Medicine* 28: 897–99.

Perez, M. V., K. W. Mahaffey, H. Hedlin, J. S. Rumsfeld, A. Garcia, T. Ferris, V. Balasubramanian, A. M. Russo, A. Rajmane, L. Cheung, et al. 2019. "Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation." *New England Journal of Medicine* 381: 1909–17.

Seshadri, D. R., B. Bittel, D. Browsky, P. Houghtaling, C. Drummond, M. Y. Desi, and A. M, Gillinov. 2020. "Accuracy of Apple Watch for Detection of Atrial Fibrillation." *Circulation* 141 (8): 702–3.