HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript; available in PMC 2023 May 15.

Published in final edited form as:

Nat Mach Intell. 2022 November ; 4(11): 909–911. doi:10.1038/s42256-022-00551-y.

Federated learning and Indigenous genomic data sovereignty

Nima Boscarino[1,✉], **Reed A. Cartwright**[2,3], **Keolu Fox**[4], **Krystal S. Tsosie**[3]

[1]Hugging Face, Brooklyn, New York, NY, USA.

[2]The Biodesign Institute, Arizona State University, Tempe, AZ, USA.

[3]School of Life Sciences, Arizona State University, Tempe, AZ, USA.

[4]Department of Anthropology and Global Health, University of California, San Diego, La Jolla, CA, USA.

Abstract

Indigenous peoples are under-represented in genomic datasets, which can lead to limited accuracy and utility of machine learning models in precision health. While open data sharing undermines rights of Indigenous communities to govern data decisions, federated learning may facilitate secure and community-consented data sharing.

Despite efforts to increase the inclusion of diverse populations in research studies, Indigenous peoples remain under-represented in genomics datasets[1–3]. Recent and past unethical research conduct involving the collection of biological and health data from Indigenous communities has strained relationships between Indigenous peoples and researchers[4,5], often resulting in tribal policies that restrict open data sharing. This poses a difficulty in fields such as bioinformatics in which the sharing of diverse sets of digital sequence information is vital for the development of robust machine learning models that inform genomic and clinical algorithms. The exclusion of Indigenous peoples from datasets limits the predictive accuracy of machine learning models, increases the potential for unintended biases and may contribute to misinformed data decisions affecting precision healthcare of Indigenous patients. Hence, there is a sustained need to facilitate secure and community-consented Indigenous data sharing that respects the right of Indigenous communities to self-govern data decisions concerning genomic information from their own peoples, known as Indigenous genomic data sovereignty[6].

Issues and concerns with sharing of Indigenous genomic data

Artificial intelligence (AI), machine learning (ML) and data science tools are continually shaping biomedical research, especially as related disciplines continue forward in the

✉ nima.boscarino@gmail.com .

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Machine Intelligence* thanks Jantina de Vries and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

digital data revolution. In 2021, the US National Institutes of Health (NIH) announced two initiatives that centre on AI and ML approaches, the Bridge2AI and AIM-AHEAD programs, which specifically address the lack of diversity in both research and data and are directed at addressing health inequities while preserving data privacy. The ethical components for these initiatives, however, are weighted heavily toward concepts such as open data access, which centre on researchers' data access and utility, but not communities' interests. Although access to genomic data may facilitate research that addresses the genomic divide, open data access conflicts with Indigenous data sovereignty (IDS), creating a pressing and timely dilemma.

The 'open data' movement supports the open sharing of data as a means to 'democratize' data access for researchers, but Indigenous peoples and communities question whether this transfers too much data decision-making authority outside of Indigenous data governance structures. Unlike for other subsets of individuals who may be included in genomic research, the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) recognizes and respects the self-determination and self-governance of Indigenous peoples[7]. This includes the right to "maintain, control, protect and develop […] their technologies […], including human and genetic resources," which is exemplified by calls from global Indigenous peoples for the cessation of large-scale diversity genome projects in their communities[8]. Despite efforts over the past 20 years to increase the inclusion of under-represented populations in datasets, as of 2019 Indigenous peoples still constituted less than 1% of genomic datasets[3], while contemporaneous sources estimated that Indigenous peoples accounted for 7% of the global population (https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2019/08/sp-2019-08-09-IndigenousDay-en-1.pdf).

Cultural knowledge is in many cases considered sacred, and there are sustained concerns against infringements. Intentions of genetically modifying the *kalo* (taro) plant in Hawaii, for instance, have been opposed because the plant is "culturally recognized by Native Hawaiians as being an ancestor"[5]. These issues are further compounded for genomic data. The protection of those data is especially relevant to Indigenous communities as a result of experiences under colonialism, working with researchers from primarily white institutions. Medical research carries a legacy of structural racism that has affected minority communities worldwide and has resulted in mistrust and skepticism[9]. For Indigenous groups, this has continued to be fuelled by disrespect, unapproved secondary research and "broken promises"[10]. Misuse of Indigenous genomic data undermines sovereignty, with the inclusion of group identifiers such as tribal community names in genomic data enabling commercial use by being "misrecognized, commodified, and sold as ancestry tests"[11]. Preventative measures such as removing identifying markers from genomic data are not reliable options because anonymized data can still be exploited to identify individuals and relatives[12]. Given this context, open-access data policies are fundamentally not suitable for Indigenous genomic data and may act as "just another form of colonial dispossession"[13]. Ultimately, Indigenous peoples should have "control over and benefit from information generated from their communities" and decide for themselves whether to make their data openly accessible.

Federated learning as an innovative strategy for secure data sharing

Federated learning is a technique for training machine learning models using datasets distributed across multiple devices or data centres[14]. Instead of all data being required to pass through a central server, as in traditional machine learning training, a partially trained model is distributed from a central server to independent nodes, which train the model on the data that they have access to. The independent nodes then return the trained model to the central server. This is in contrast to open-access methods, as the independent nodes can contribute to training models without ever exposing their data. Individuals or institutes with large datasets that need to remain private, due to legal or business limitations, can use federated learning to collaboratively train models on their combined datasets. Federated learning is increasingly gaining traction in healthcare informatics[15,16] – particularly due to regulations such as the US Health Insurance Portability and Accountability Act (HIPAA) and the EU General Data Protection Regulation (GDPR) – and could address Indigenous privacy concerns in genomics research. Specifically, federated learning as a tool can empower Indigenous communities to engage with research that sets out to bridge the "genomic divide"[2], while retaining community ownership, control, access and possession of their data[17].

Because genomic sequences can be used for machine learning in bioinformatics research, there is an incentive to continue exploring privacy-preserving approaches. Federated learning is showing promising potential for genome-wide association studies[16,18] and may provide an additional layer of privacy, assurance and control by limiting the data that are exposed to external researchers, possibly by selectively restricting data access to parts of genomes, metadata or whole sequences.

Furthering privacy-preserving machine learning techniques is an active field of research, with federated learning being at the forefront. Techniques such as differential privacy, secure multi-party computation and homomorphic encryption are being leveraged to minimize data leakage[19], while advances in network topologies continue to optimize training performance. Not all improvements in these domains are directly relevant to genomics, however. Differential privacy, for example, may introduce too much noise into the data to be useful for genomics, although there are efforts to resolve such issues[20].

Federated learning can support community-driven data decision-making

Federated learning is well placed to be compatible with existing principles and guidelines used in Indigenous genomics research projects, such as the OCAP (ownership, control, access, and possession)[17] and CARE (collective benefit, authority to control, responsibility, ethics)[21] frameworks. Establishing the necessary infrastructure for federated learning would maintain ownership and control of data under Indigenous Data Sovereignty while empowering communities to "assess benefits, harms, and potential future uses based on community values and ethics"[21] to block unauthorized secondary research.

There is an increased awareness that ethical and equitable biological and genomic research should centre principles of community-based participation. Engaging community decision-

making in the research process should not end at data collection, however. Many data-governance frameworks highlight the FAIR (findability, accessibility, interoperability and reusability) principles[22] to facilitate the sharing of data collected from under-represented communities, but it has been argued that these principles grant too much power to researchers and circumvent Indigenous data governance and authority[10]. Supplemental data governance frameworks such as OCAP and CARE highlight Indigenous community data decision-making in their principles but still rely on researchers to implement these in their research.

Among the ways that Indigenous peoples around the world have been exercising their right to data sovereignty and community-driven data decision-making is by establishing biobanks[2] and creating Indigenous Background Variant Databases[1]. These initiatives are ongoing, and there are opportunities to expand their scope with federated learning systems to allow the communities that run the projects to choose how their data can continue to be used.

Next steps and issues to consider

For researchers and communities considering taking a federated-learning-based approach to managing genomic data, there will be a number of details to consider. First, researchers should focus on the importance of engaging with Indigenous communities as partners in research, as opposed to simply subjects of research. This requires ensuring that beyond treating data appropriately, researchers also handle collected DNA and samples with respect. Additionally, although a research team may make a case for federated-learning-based data governance, the decision to move forward with the implementation must ultimately lie in the community's hands.

Simply offering federated learning to communities does not absolve researchers of unethical behaviour or usage of data. If the implementation of such systems is initiated by researchers, care must be taken to comprehensively inform study participants about the effects of the research being carried out, potential future usage of their data and the workings of federated learning systems. Additionally, communities should have the ultimate say in which research and analysis they may be included in, and consent should be revocable at any time.

There are also questions around the granularity of the data that would be made available. For example, releasing community-level identifiers poses risks, and some communities may choose to refrain from including such fields in their data even at the expense of affecting the data's research value. Conversely, federated learning could be used to protect community-level identifiers in aggregate studies until a community elects to release them.

Any tool or practice that is used in projects involving Indigenous data should uphold and satisfy the principles set forth by Indigenous communities themselves. Currently, the literature on federated learning does not consider Indigenous perspectives, but there is an opportunity to approach data privacy, artificial intelligence and technology as a whole with Indigenous epistemologies[23]. Of major importance is the ownership and administration of the infrastructure for federated learning systems. Currently, there is a reliance on researchers and external parties to host data because of a technological gap between communities and

research institutes. Any Indigenous community deciding to utilize federated learning would need to invest in the required infrastructure up front and develop a long-term maintenance plan, as the organization that owns the data also has to own and maintain the infrastructure that allows other people to analyse them. This can be paired with mandating capacity building as part of research projects, which is a common practice in research involving Indigenous communities.

It is possible that Indigenous-led consortia could offer centralized solutions for Indigenous communities looking to pool resources and utilize federated learning for medical or commercial purposes, and such plans could be used as justification to apply for instrumentation funding. These platforms, and the policies governing them, could be developed by Indigenous-led technology firms, research institutes and think tanks in collaboration with the communities and researchers that will use them. Ideally, such technologies would be developed from the ground up and would be rooted in Indigenous worldviews; however, the costs of the engineering and product development process could potentially be prohibitive. As such, it may be appropriate and necessary to incorporate both open-source and white-label components to accommodate the available budgets.

Although most conversations related to IDS tend to localize Indigenous peoples of North America, it is important to realize that Indigenous sovereign authority is intrinsic and is not colonially defined. Therefore, we should consider all global Indigenous peoples to have sovereign domain over their peoples' data. However, this is complicated by unequal legal protections and differences in research investments that make certain goals unattainable for Indigenous peoples in certain parts of the world. Some countries, for example, might not possess the resources for precision medicine or to even pursue this type of genomics research. With this in mind, a question is raised regarding whether federated learning could present a way to think about Indigenous peoples' data beyond genomics. Exploring the applications of federated learning for data in these settings could require consultative and development processes involving the leadership and collaboration of Indigenous peoples' from across the world, and might be carried out via research frameworks that emphasize intersectionality and Indigenous self-determination[24].

Acknowledgements

N.B. thanks J. T. Topham and G. Pistilli for their advice and feedback on the manuscript. K.S.T. and K.F. are grateful to the ENRICH (Equity for Indigenous Research and Innovation Coordinating Hub) program at New York University and University of Waikato, funded in part through a gift from the Minderoo Foundation, for their support for machine-focused solutions that uplift Indigenous data sovereignties.

References

1. Caron NR et al. Front. Public Health 8, 111 (2020). [PubMed: 32391301]
2. Caron NR, Boswell BT, Deineko V & Hunt MA JCO Glob. Oncol. 6, 120–123 (2020). [PubMed: 32031443]
3. Mills MC & Rahal C Commun. Biol. 2, 1–11 (2019). [PubMed: 30740537]
4. Reardon J & TallBear K Curr. Anthropol. 53, S233–S245 (2012).
5. Taniguchi NK, Taualii M & Maddock J Int. Indig. Policy J 3, (2012).
6. Tsosie KS, Yracheta JM, Kolopenuk JA & Geary J Am. J. Bioeth. 21, 72–75 (2021).

7. United Nations General Assembly. United Nations Declaration on the Rights of Indigenous Peoples, A/RES/61/295 (2008).
8. Harry D, Howard S & Shelton BL Indigenous People, Genes and Genetics: What Indigenous People Should Know About Biocolonialism (Indigenous Peoples Council on Biocolonialism, 2000).
9. Geneviève LD, Martani A, Shaw D, Elger BS & Wangmo T BMC Med. Ethics 21, 17 (2020). [PubMed: 32075640]
10. Tsosie KS, Fox K & Yracheta JM Nature 591, 529–529 (2021).
11. Fox KN Engl. J. Med. 383, 411–413 (2020).
12. Bonomi L, Huang Y & Ohno-Machado L Nat. Genet. 52, 646–654 (2020). [PubMed: 32601475]
13. Tsosie KS, Yracheta JM, Kolopenuk J & Smith RWA Am. J. Phys. Anthropol. 174, 183–186 (2021). [PubMed: 33244756]
14. Kairouz P et al. Found. Trends Mach. Learn. 14, 1–210 (2021).
15. Kaissis GA, Makowski MR, Rückert D & Braren RF Nat. Mach. Intell. 2, 305–311 (2020).
16. Wu X et al. Brief. Bioinform. 22, bbaa090 (2021). [PubMed: 32591779]
17. Schnarch BJ Aborig. Health. 1, 80–95 (2004).
18. Nasirigerdeh R et al. Genome Biol. 23, 32 (2022). [PubMed: 35073941]
19. Zhou J et al. Preprint at https://arxiv.org/abs/2104.10501 (2021).
20. He Z, Li Y, Li J, Li K, Cai Q & Liang Y Tsinghua Sci. Technol. 23, 389–395 (2018).
21. Carroll SR et al. Data Sci. J. 19, 43 (2020).
22. Wilkinson MD et al. Sci. Data 3, 160018 (2016). [PubMed: 26978244]
23. Lewis JE, Arista N, Pechawis A & Kite SJ Des. Sci. 10.21428/bfafd97b (2018).
24. Levac L et al. Learning Across Indigenous and Western Knowledge Systems and Intersectionality: Reconciling Social Science Research Approaches (Univ. Guelph, 2018).