Ethics of AI in global health research

Cape Town, 29&30 November 2022

Global Forum on Bioethics in Research

Governance paper
A silent trial is critical to accountable and justice-promoting implementation of artificial intelligence in healthcare

Melissa D McCradden[1,2,3]
[1]The Hospital for Sick Children – Department of Bioethics
[2]Peter Gilgan Centre for Research & Learning – Genetics & Genome Biology Research Program
[3]Dalla Lana School of Public Health

Context

Artificial intelligence (AI) – a subset of machine learning (ML) – refers to a computational system that can run inference on novel cases based on a mapping of inputs (data) to outputs (labels). The ability to predict accurately the current or future state of a patient's health is a valuable property, yet to do so does not guarantee the patient will benefit. In a recent systematic review of clinical AI tools, only two fifths of those with good technical performance also were associated with a concomitant improvement to patient outcomes[1]. These trials are essential for two reasons: 1) to reliably establish knowledge about the causal effect of an AI tool on a given patient outcome; and 2) because AI can introduce novel and unanticipated errors, including automation bias[2] and algorithmic injustice[3]. Ensuring that AI adoption actually benefits patients reflects a commitment to evidence-based practice, responsible health data stewardship, organizational accountability, and efficient use of hospital resources. Additionally, the manner of technological integration and perception of value alignment can contribute positively or negatively to healthcare worker satisfaction with the tool, as we have seen with the introduction of the electronic health record. For all these reasons, the demonstration that an AI tool benefits patients is crucial to ethical integration for both patients and healthcare workers.

AI tools are most reliable when the population and environment on which they were trained is highly similar to that in which they are integrated. As such, generalizability of an AI system's performance is a matter of controversy[4]. Attempts to utilize systems like Watson for Oncology have proven to be highly inappropriate when using them internationally. Assuming, rather than testing, that a model's performance will be retained in a new environment is a recipe for failure. Low- and Middle-Income Countries (LMICs) may reasonably wish to adopt AI technologies, some of which may be developed and trialed externally. In some cases, the population distribution will differ substantially from the training population, which poses additional concerns to the fairness properties of the system.

How should international healthcare organizations go about ascertaining that performance will be retained in their local population and promote rather than compromise fairness to patients? This governance paper proposes the silent trial as a critical process to protect patients' interests and serve justice.

<u>The silent trial</u>

In model development, an algorithm is initially developed and validated using historical data from a curated dataset as an initial proof-of-concept that some output (label) can be effectively mapped given particular inputs (data). While a valuable component of responsible AI evaluation[5], such validation is not sufficient to guarantee similar performance in the clinical environment. As such, the silent trial (aka shadow trial or silent mode) refers to the deployment of a model in the anticipated clinical environment, where the model is running inference on active cases and making predictions – however, these predictions are seen only by a research team and do not influence patient care. The predictions are recorded and compared to the true clinical outcomes or human-

defined labels that they are predicting (e.g., a radiologist's confirmation of a given diagnosis). The silent trial thus enables two goals: 1) demonstrating the ecological validity of the model and 2) offers the kind of information that can establish clinical equipoise, which is the justificatory basis on which interventional trials are considered ethically permissible[6].

Silent trial and justice

In the context of LMICs, requiring that a silent trial be conducted onsite prior to committing to adopting a particular AI model can be a valuable process for protecting patients' interests. While compliance with standardized reporting guidelines and transparent reporting of clinical trials can help institutions identify whether a model could have utility for them[7], good models can still fail to generalize to new settings. The silent trial verifies whether a model will perform adequately on the local patient population prior to actually using the model to inform the care of patients. Institutions can thus demonstrable that they are accountable to their patients by verifying the AI's appropriateness in an empirical sense.

As is well-recognized by now, AI systems reflect the biases embedded in societal injustices, spitting out patterns of inequity, inaccessibility, and prejudice when not adequately overseen. Several case examples in the literature and public discourse describe the use of AI that resulted in systemic disadvantages to racialized, marginalized, and/or oppressed groups. These consequences are further problematic in their lack of recourse – affected persons typically have no avenue to dispute the output, and users do not have the tools to identify where the AI went wrong.

As such, increasing attention is being paid to the importance of algorithmic audits as a mechanism for the more robust characterization of algorithmic performance (i.e., with respect to clinical and demographic sub-groups)[7]. As fairness is a core concern of AI, the principle of distributive justice suggests two key steps must be taken: first, establishing the distributive benefits and burdens of a given system, and second, identifying opportunities for correction, revision, or redress. For AI, step 1 would look like an algorithmic audit at the silent trial stage to characterize model performance across the locally relevant population and subgroups of patients. Step 2 would involve reflection and engagement regarding the suitable options to address potential discrepancies in that distribution. Relational ethics highlights the importance of engaging those most affected by AI under-performance or discrimination, respecting their lived experience as valuable knowledge, and collaboratively identifying solutions that serve justice[3].

As previously noted[8], LMICs are not identical in population, character, or context. The silent trial provides additional opportunities to truly integrate (rather than 'deploy') a model by engaging those using the model in its implementation. One can imagine that human factors evaluation and stakeholder engagement will be increasingly important as the world outside of healthcare continues to be marred by ethical controversies involving AI. The silent trial provides a technical foundation to verify performance and an evaluative process whereby stakeholders are actively engaged in shaping AI implementation. The process thereby operationalizes relational ethics values that emphasize human interconnectedness and moral obligations to others[9] – a much-needed juxtaposition to the algorithmic coldness of AI.

Case example

Our institution is currently trialing a classification model to identify obstructive hydronephrosis in infants[10]. Despite the initially strong model performance (AUROC 90%) when moved into the silent trial stage we observed a decrease in performance (to an AUROC of 50%). Upon investigation, it was revealed that the patients in the silent trial were significantly younger and more likely to have right-sided obstruction of the kidney. Controlling for these differences improved the model performance. The team next addressed differences in image processing to improve the model performance to an AUROC of 85%. The model was additionally evaluated for performance across sex, laterality of hydronephrosis, ultrasound machine, and the patient's home postal code. The team found a retained performance of over 90% sensitivity across all variables. By conducting a silent trial, the team was able to establish the generalizability of the model to the live clinical setting and assure themselves of its performance across relevant patient subgroups. Notably, we were

unable to assess race or ethnicity as such data is not yet routinely collected in Canada; despite the use of postal code as a proxy, we acknowledge the limitations of this work.

The need for the silent trial is evident when considering the known limitations of model generalizability. For example, Straw and Wu[11] assessed the performance of four classifier models that predict the presence of liver disease from a commonly used dataset. They observed a higher false negative among women generally across all studies, indicating that more women than men could have a missed diagnosis using the model. A caveat, however, is that although they addressed the issue of class imbalance, there is limited information into the effects of potential confounding variables (a noted limitation in many ML-based works exploring bias[12]).

An additional opportunity embedded in the silent trial paradigm is the chance to seek user feedback and engagement around model integration choices[10,13]. This can be done through quantitative or qualitative activities. When faced with model performance discrepancies, there is a need to make choices about how the effects will be mitigated. The silent trial provides the empirical characterization needed to ground discussions about the most appropriate actions to take for model integration to promote equity, transparency, and ethical decision-making.

Conclusion and recommendation

This paper advocates for widespread adoption of a silent trial prior to the integration of any AI tool. This step enables the operationalization of distributive justice and provides a reliable empirical foundation for ensuring AI tools will benefit patients across different contexts. Algorithmic auditing can be a key piece of the silent trial to identify failure modes, disproportionate error rates and other performance metrics, inform postdeployment safety monitoring, identify the need for postdeployment recalibration, and mitigate risks to groups[8]. Audits thereby are a way to operationalize distributive justice as well as healthcare's commitments to evidence-based integration and good clinical decision-making. A common goal across all contexts is the use of AI to alleviate health burdens and augment the humanistic aspect of medicine. Adopting processes that enable accountable decisions and provide an empirical foundation for informed decisions on fairness and justice is one step toward this end.

References

1. Zhou, Q., Chen, Z.H., Cao, Y.H. and Peng, S., 2021. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *NPJ digital medicine*, *4*(1), pp.1-12.
2. Tschandl, P., C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. 2020. Human-computer collaboration for skin cancer recognition. *Nature Medicine* 26(8): 1229–1234.
3. Birhane, A., 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2), p.100205.
4. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), e489-e492.
5. Wiens, J., S. Saria, M. Sendak, M. Ghassemi, V. X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, et al. 2019. Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine 25(9):1337–40.
6. McCradden, M.D., Anderson, J.A., A. Stephenson, E., Drysdale, E., Erdman, L., Goldenberg, A. and Zlotnik Shaul, R., 2022. A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics*, *22*(5), pp.8-22.
7. Liu, X., Glocker, B., McCradden, M.M., Ghassemi, M., Denniston, A.K. and Oakden-Rayner, L., 2022. The medical algorithmic audit. *The Lancet Digital Health*.
8. McCradden, M.D. and Chad, L., 2021. Screening for facial differences worldwide: equity and ethics. *The Lancet Digital Health*, *3*(10), pp.e615-e616.
9. Ewuoso, C., 2021. An African Relational Approach to Healthcare and Big Data Challenges. *Science and Engineering Ethics*, *27*(3), pp.1-18.

10. Kwong, J. C., Erdman, L., Khondker, A., Skreta, M., Goldenberg, A., McCradden, M. D., ... & Rickard, M. (2022). The silent trial-the bridge between bench-to-bedside clinical AI applications. *Frontiers in digital health*, 163.
11. Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ health & care informatics*, *29*(1).
12. Bernhardt, M., Jones, C., & Glocker, B. (2022). Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, *28*(6), 1157-1158.
13. Tonekaboni, S., Morgenshtern, G., Assadi, A., Pokhrel, A., Huang, X., Jayarajan, A., ... & Goldenberg, A. (2022, April). How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at ICU. In *Conference on Health, Inference, and Learning* (pp. 169-182). PMLR.

This paper was prepared for GFBR 2022. Further details on the meeting are available at www.gfbr.global.