npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01399-9

Ethical data acquisition for LLMs and AI algorithms in healthcare

Marta Williams, Wasie Karim, Justin Gelman & Marium Raza

Artificial intelligence (AI) algorithms will become increasingly integrated into our healthcare systems in the coming decades. These algorithms require large volumes of data for development and fine-tuning. Patient data is typically acquired for AI algorithms through an opt-out system in the United States, while others support an opt-in model. We argue that ethical principles around autonomy, patient ownership of data, and privacy should be prioritized in the data acquisition paradigm.

Artificial intelligence (AI) innovation has permeated nearly every industry, including healthcare. AI refers to computer technology that reasons and cognitively functions in a similar manner to humans. Recent advancements in AI include large language models (LLMs) - AI tools trained on large amounts of data to simulate human conversation. LLMs like ChatGPT and Bard attract great excitement, with human-like responses to queries across diverse knowledge areas. LLMs can excel at specific, task-oriented problems either through 'fine-tuning' on specific, relevant datasets or by leveraging techniques such as in-context learning or Chain-of-Thought (CoT) prompting. The effectiveness of these techniques can vary depending on the model architecture, training data, and whether the model is open-source or proprietary[1]. Within healthcare, LLMs are already being used to assist with administrative tasks such as clinical note-writing and patient portal communications[2,3]. These applications highlight the potential of LLMs to transform healthcare delivery by improving efficiency and patient engagement.

Moreover, AI has the potential to significantly improve patient outcomes as AI tools continue to be developed and integrated into clinical practice. For example, recent trial data suggest AI may improve care for patients suspected to be having a myocardial infarction[4]. Furthermore, AI-assisted imaging technology is already in use to aid physicians in the real-time identification of cancerous polyps during colonoscopies. Without a doubt, the further integration of AI into medical practice is not only inevitable but also poised to revolutionize healthcare if done appropriately.

However, it must be acknowledged that AI models are developed within the confines of existing structural inequities, and without deliberate effort, are at risk of perpetuating them[5,6]. While historical cases like Dr. Sims's experiments on enslaved women and the Tuskegee syphilis study illustrate a long-standing precedent of medical exploitation, modern examples demonstrate that these issues have the potential to persist in contemporary AI-driven healthcare[7]. For instance, Obermeyer et al. found racial bias in a widely used commercial health algorithm, where Black patients assigned the same risk level as White patients were sicker[8]. Similarly, convolutional neural networks for skin lesion classification show significantly reduced diagnostic accuracy for Black patients, as they are often trained on datasets where only 5% to 10% of the images come from Black individuals[9]. These examples underscore the capacity for AI to exacerbate existing disparities if not developed with equity in mind and reflect the ongoing need to address bias in healthcare algorithms today[10].

The challenge for future development lies in acquiring data that is representative of diverse patient populations, without impeding on patient rights or worsening existing population health disparities. Currently, data for algorithm development is acquired in two ways: opt-in and opt-out. Opt-in involves patients explicitly providing informed consent to include their health data in an AI training dataset. Opt-out, the current default in the United States[11], involves automatically including patient data in AI training datasets unless patients specifically choose otherwise. We aim to define the benefits and pitfalls of each model, and argue that ethics should be prioritized over financial incentives for future LLM development.

Comparing opt-in and opt-out

The opt-out model for collecting patient health data has advantages (Table 1), nearly guaranteeing sample sizes representing the full spectrum of diversity and thereby ensuring more accurate AI models and output. This method is easily scalable, with higher consent rates than seen with opt-in models[3], and provides a wealth of data with minimal expense or paperwork. The disadvantages of the opt-out method, however, are significant. Bypassing an explicit informed consent process limits patient autonomy; patients may be unaware that their data is utilized, or that they have the option to limit that use. If patients are not explicitly asked about the use of their data, they are also unlikely to be compensated for profits garnered by models trained using their data.

The opt-in model for data collection has significant advantages in terms of informed consent and patient autonomy (Table 1). The current default model for most of Europe based on the European Union General Data Protection Regulation[12], opting-in requires that patients be informed about, and provide consent for, the use of their data in AI development. The opt-in model improves patient trust and prioritizes transparency between researchers, healthcare providers, and patients. The disadvantages for opt-in data collection models are lower consent rates and consent bias[11]. Opt-in

Harvard Medical School, Boston, MA, USA. e-mail: martawilliams@hms.harvard.edu

Table 1 | Advantages and Disadvantages of Opt-In and Opt-Out Models

Relevant factors	Opt-in consent	Opt-out consent
Autonomy	Respects patient autonomy/choice	Offers limited autonomy
Informed Consent	Involves explicit informed consent	May bypass the informed consent process
Participation Rates	Limited participation	Greater participation
Administrative Efficiency	May require investment in recruitment, advertisement, messaging	Easily scalable, ease of collection
Selection Bias	Biases model to skew male, higher education level, and higher SES	Larger, more diverse samples that may yield more accurate AI models
Trust Building	Enhanced transparency improves trust between researchers, healthcare providers, and patients	Patients, especially vulnerable patient populations, may not be aware of their right to opt out if nobody explicitly tells them
Equity and Fairness/Benefit Sharing	Patients can be informed and possibly compensated for their data contributions	Patients might not be compensated for profit gained from an algorithm trained on their data

procedures tend to be biased towards the inclusion of patients who are male, more highly educated, and of higher socioeconomic status[11], thus AI models trained using that data are likely to be similarly skewed. Opt-in procedures are also more labor-intensive and expensive because of the time, money, and paperwork required to inform patients and document their consent.

Call to action

Given the rapid integration of AI into healthcare, it is imperative for the healthcare and AI communities to prioritize patient needs as the central focus while advancing the implementation of this technology.

Ideally, the opt-in model would address concerns of patient autonomy by requiring informed consent before collecting patient data. Yet, this model risks perpetuating existing biases by failing to recruit a representative patient population as so many other well-intentioned healthcare efforts currently do. We considered how an opt-in model could more successfully recruit patients across the socioeconomic and educational spectrum. Opportunities could involve direct compensation for data collection or discounted healthcare services. Yet even when compensation is offered, studies often struggle to recruit underrepresented populations due to historical injustices and structural health inequities, including financial and transportation barriers to study participation[13,14]. For example, between 2015 and 2019, 78% of FDA clinical trial participants were non-Hispanic whites, despite them comprising only 61% of the population[13]. While a 2023 study found that a $100 incentive was more effective at increasing participation among white and affluent people than among those from low-income or non-white households, a larger $500 incentive closed the participation gap among different racial, ethnic, and socioeconomic groups, indicating that sufficient financial incentives may persuade underrepresented patients to opt in[15]. However, financial compensation to this degree would be unsustainable considering the prodigious amount of patient data required for a well-functioning LLM.

Given the financial incentives and ease of data collection offered by opt-out models, it is likely that opt-out models will predominate despite their potential ethical disadvantages. Advancements in privacy-preserving methodologies, such as differential privacy and federated learning, may provide a path forward for the opt-out model while upholding ethical principles[16]. Differential privacy introduces carefully calibrated noise into datasets, ensuring that individual data points remain unidentifiable while preserving the overall utility of the data. Federated learning enables decentralized AI training by keeping patient data within local systems and sharing only aggregated updates, reducing risks associated with centralized repositories and enhancing data security. By integrating these technologies into data collection frameworks, it will be possible to maintain patient privacy and autonomy while still acquiring the diverse, representative datasets necessary for unbiased AI model development. Alongside technological developments, it is imperative that opt-out models be made with transparency in mind to appeal to patients concerned about privacy; these actions will not only address immediate ethical concerns but also foster trust and inclusivity as AI technologies become more commonplace in healthcare.

The following are three specific actions we recommend:

1. Patients must be provided with clear and concise terms and conditions that are less than one page and written in patient-centered language rather than legal jargon.

   Unlike those of smartphones and social media platforms that are so long that most people skip past them, consent in an equitable opt-out system must be accessible to patients of all health literacy levels. Healthcare providers and administrators should create clear, concise terms and conditions, as well as educational materials that help patients understand the impact of their data on predictive analytics. These materials can be provided to patients at the outset explaining the use of their data and their right to opt-out, rather than their data “silently” being used behind a wall of lengthy terms and conditions. Additionally, securing consent at regular intervals through online portals or in-person reminders at clinic visits can ensure that patients remain aware of their rights and can exercise autonomy within an opt-out system.

2. Patients must be the ultimate owners of their data, and infrastructure must be built to both protect patient data and allow patients to readily extricate their information from databases by request.

   Whether opt-in or opt-out models are pursued, concerns regarding data ownership still stand. When eliciting concerns regarding AI in healthcare, patients consistently voice fears about rising costs to incorporate this novel technology[17]. Given the inevitable profit motive underlying LLM implementation in healthcare, it is hard to imagine a sustainable system in which patients are asked to provide their data for free to develop models for which they are subsequently charged. This dilemma will require extensive conversation on patient data ownership, compensation, and reimbursement for the use of AI technology. The commodification of patient data, without adequate safeguards and fair compensation, risks perpetuating the legacy of exploitation exemplified by Henrietta Lacks, whose cells were used to generate enormous profits without any compensation to her or her family[7]. If healthcare and technology companies do not draw firm boundaries on patients’ right to own their data, we may see a continuation of healthcare exploiting our most vulnerable communities.

3. Government and healthcare organizations must immediately invest in creating and enforcing regulatory standards to ensure patient safety and trust.

   Placing the burden of ethical practice entirely on individual organizations and patients is insufficient. Many patients lack the data and technology literacy necessary to make truly informed decisions about contributing their data[18,19]. Patients cannot and should not be expected to be capable of interrogating the safety, transparency, and reversibility of their data contributions to LLMs. Therefore, responsibility must also fall on other partners—such as funders, healthcare systems, healthcare administrators, data use committees, and others involved in the AI healthcare enterprise—to contribute to the ethical integration of AI. Adopting frameworks like differential privacy and federated learning will ensure that patient data is utilized ethically and inclusively, minimizing risks while maintaining the utility of datasets. It is also imperative to uphold the trust that underpins an opt-out system by establishing a third-party regulatory board, supported by the government, to develop and enforce transparency and safety standards while ensuring ongoing compliance through rigorous oversight.

With significant existing patient skepticism surrounding AI, we must anticipate and respond proactively to concerns to ensure LLMs are representative, transparent, and respectful of patient autonomy. LLMs have already begun to revolutionize science, medicine, and the speed at which we can advance in any given field. We have already seen the immense benefit AI can provide patients, and the AI community must ensure these benefits are available to *all* patients by combating the reinforcement of existing biases in emerging technologies.

Received: 3 June 2024; Accepted: 16 December 2024;
Published online: 24 December 2024

References

1. Naveed, H. et al. A Comprehensive Overview of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2307.06435 (2024).
2. Hudelson, C. et al. Selection and implementation of virtual scribe solutions to reduce documentation burden: a mixed methods pilot. *AMIA Summits Transl. Sci. Proc.* **2024**, 230–238 (2024).
3. Garcia, P. et al. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Netw. Open* **7**, e243201 (2024).
4. Lin, C. et al. Artificial intelligence–powered rapid identification of ST-Elevation Myocardial Infarction via Electrocardiogram (ARISE) — A pragmatic randomized controlled trial. *NEJM AI* **1**, AIoa2400190 (2024).
5. Jindal, A. Misguided artificial intelligence: how racial bias is built into clinical models. *J. Brown Hosp. Med.* **2**, 1–6 (2022).
6. Agarwal, R. et al. Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy Technol.* **12**, 100702 (2023).
7. Baptiste, D. et al. Henrietta Lacks and America's dark history of research involving African Americans. *Nurs. Open* **9**, 2236–2238 (2022).
8. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
9. Kamulegeya, L. et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. *Afr. Health Sci.* **23**, 753–763 (2023).
10. Raza, M. M., Venkatesh, K. P. & Kvedar, J. C. Promoting racial equity in digital health: applying a cross-disciplinary equity framework. *Npj Digit. Med.* **6**, 1–3 (2023).
11. de Man, Y. et al. Opt-in and opt-out consent procedures for the reuse of routinely recorded health data in scientific research and their consequences for consent rate and consent bias: systematic review. *J. Med. Internet Res.* **25**, e42131 (2023).
12. What is GDPR, the EU's new data protection law? *GDPR.eu* https://gdpr.eu/what-is-gdpr/ (2018).
13. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Women in Science, Engineering, and Medicine; Committee on Improving the Representation of Women and Underrepresented Minorities in Clinical Trials and Research. *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. (National Academies Press (US), Washington (DC), 2022).
14. Fairley, R. et al. Increasing clinical trial participation of black women diagnosed with breast cancer. *J. Racial Ethn. Health Disparities* **11**, 1701–1717 (2024).
15. Dutz, D. et al. Representation and hesitancy in population health research: evidence from a COVID-19 Antibody Study. Working Paper at https://doi.org/10.3386/w30880 (2023).
16. Abadi, M. et al. Deep learning with differential privacy. in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* 308–318 (Association for Computing Machinery, New York, NY, USA, 2016). https://doi.org/10.1145/2976749.2978318.
17. Richardson, J. P. et al. Patient apprehensions about the use of artificial intelligence in healthcare. *Npj Digit. Med.* **4**, 1–6 (2021).
18. Nguyen, A., Mosadeghi, S. & Almario, C. V. Persistent digital divide in access to and use of the Internet as a resource for health information: Results from a California population-based study. *Int. J. Med. Inf.* **103**, 49–54 (2017).
19. Cohort bias in predictive risk assessments of future criminal justice system involvement | PNAS. https://www.pnas.org/doi/10.1073/pnas.2301990120

Author contributions

MW, WK, JG, and MR all contributed equally to this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Marta Williams.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.