The Pitfalls of Black Box AI in Medicine

(An editorial)

Yasmine Madan
Honours Bachelor of Health Sciences, Year II, McMaster University
madany@mcmaster.ca

The Pitfalls of Black Box AI in Medicine

Abstract:

Artificial intelligence (AI) is augmenting human capacities and is slowly infiltrating every industry—even medicine. The literature shows various possible applications of AI that aid in clinical decision making. Yet AI is not being widely adopted into clinical practice. The purpose of this editorial is to explain how black box AI models contribute to this lack of adoption, and to advocate for the implementation of explainable AI.

Editorial:

Over the past decade, artificial intelligence (AI) has become a common subject of healthcare research. AI shows great promise for clinical decision support, predicting health outcomes, and detecting abnormalities in imaging.[1] A search in the PubMed database demonstrates a rapidly increasing trend in the number of publications involving artificial intelligence in medicine between 2010 and 2021 (Figure 1). This accelerating research output demonstrates the heightened interest and the variety of possible applications of AI. Despite this, there are currently only a limited number of examples of AI being successfully used in clinical practice.[2] How is it that we have been creating, testing, and demonstrating the usefulness of AI models for the past decade, yet there isn't widespread adoption of AI in clinical practice?

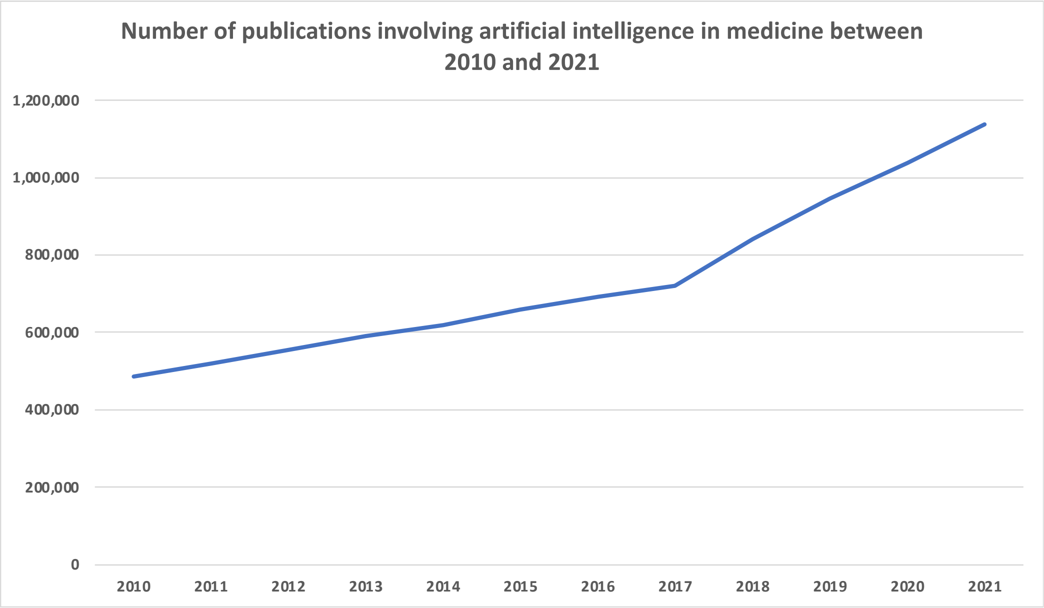

Figure 1: Absolute number of publications mentioning medicine and AI in their title or abstract between 2010 and 2021. The data were gathered from the PubMed database, with search terms: [medicine or healthcare or clinical] AND [artificial intelligence or deep learning or machine learning].

Clinicians have expressed valid concerns regarding the lack of transparency of AI algorithms. For instance, a recent study evaluated clinicians' trust of an AI clinical decision support tool for the diagnosis of COVID-19 based on chest CT images.[3] Clinicians ranked reliability and trust of the AI system on a five-point Likert scale (five corresponding to the most positive response). The mean rank was only 2.15 for reliability and 2.12 for trust.

AI image recognition algorithms are created by machine learning systems that detect patterns from large databases of information. For example, a system that is given thousands of CTs and the associated diagnoses, can recognize patterns linking the two. When this "trained" system is then given a CT that it has never "seen" before, it can predict the diagnosis based on that pattern. Although these algorithms can make correct predictions with a high degree of accuracy, they are not perfect, and sometimes make errors. However unlike humans, who can provide a rationale, when AI systems make errors, it is difficult to understand why the error occurred.[4] Because the algorithm develops and matures independently as it encounters more data ("learning"), ultimately even the programmer who created the original algorithm may not be able to understand why it is now reaching a certain conclusion. For the clinicians who would use these systems, this presents a dilemma, as they cannot interrogate the *reason* the AI system is predicting a certain diagnosis or recommending a certain treatment. The term used to describe such AI models is "black boxes" – wherein only an input and an output can be seen, while the inner-workings remain hidden inside a black box.

Clinical decision making requires transparency in the decision-making process, whereby black box algorithms present a limitation. This challenge raises complex questions such as: Can we assure that a black box algorithm is valid and free of bias without knowing its inner workings? How can we trust a recommendation from an AI model when we are oblivious to how it was reached?

Black box algorithms risk hiding errors and biases, which threatens the trustworthiness of the output for end users. For instance, one AI algorithm that was used to triage patients learned to assign patients with asthma at an inappropriately low risk of death by pneumonia. It was later discovered that this was because the algorithm was trained on a biased dataset consisting of patients with asthma who had already received an active physician intervention (a population which does not reflect the typical asthma population that the algorithm would be used for).[5] In a second example, an AI algorithm incorrectly predicted that Black patients required less healthcare resources than White patients, due to source data reflecting historical systemic discrimination restricting access to healthcare to Black communities.[6] These instances suggest that black box algorithms are only as useful as the data that was used to "create" them, and this represents an important vulnerability regarding their use medicine.

Trust is the foundation of medical practice. Thus, for AI models to have successful uptake, they must gain the trust of both clinicians and their patients. The known biases and errors that may arise from black box algorithms erode this trust, and if clinicians were to use black box models, they would have to blindly trust the algorithm (and thus the validity of the dataset that it was created from). This "blind trust" is an emerging ethical challenge in medical decision-making.[7] The black box dilemma becomes particularly concerning when an AI model makes a prediction that the clinician disagrees with. In this case, the clinician does not understand why the algorithm came to its decision, which hinders their ability to justify their own decisions and further contributes to a lack of trust. For patients, they may be left with conflicting recommendations between an unexplained AI model and their doctor, in some cases jeopardizing trust in their provider and/or the medical system.

We must find ways to overcome these barriers and harness the power of AI to assist healthcare providers to improve care. It is important to acknowledge that AI isn't perfect, but we must also recognize that humans make mistakes too. In fact, a survey of 726 pediatricians showed that almost half (45%) reported misdiagnosis at least once or twice per month.[8] Here, AI used alongside clinical judgement might alert and prompt providers regarding possible missed (and particularly obscure) diagnoses, acting as an instantaneous "second opinion" that can be taken with a grain of salt.

Overcoming the barrier of lack of trust may be achievable by creating explainable AI models: systems that can make predictions, but can also explain to clinicians why and how they came to these conclusions. We must open the black box by working towards a future that prioritizes the integration of explainable AI in clinical decision making. With the right algorithms by our side, we can dare to grow.

References

1. Vinny PW, Vishnu VY, Padma Srivastava MV. Artificial intelligence shaping the future of neurology practice. Medical Journal Armed Forces India. 2021;77(3):276–82.
2. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with Artificial Intelligence. BMC Medicine. 2019;17(1).
3. Goel K, Sindhgatta R, Kalra S, Goel R, Mutreja P. The effect of machine learning explanations on user trust for automated diagnosis of COVID-19. Computers in Biology and Medicine. 2022;146:105587.
4. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for Artificial Intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making. 2020;20(1).
5. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for Healthcare. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
7. Berg JW, Appelbaum PS, Lidz CW, Parker LS. The concept and ethical justification of informed consent. Informed Consent. 2001;14-38
8. Singh H, Thomas EJ, Wilson L, et al. Errors of diagnosis in pediatric practice: a multisite survey. Pediatrics 2010;126:70-9.