[Commentary] Artificial Intelligence, or Artifact Intelligence? Most AI Is Not Ready for Practice

Peter Muennig[1], Chris Pirazzi

1 Columbia University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This commentary emphasizes the limitations of artificial intelligence (AI) in medical diagnosis, highlighting instances of erroneous diagnoses due to inadequate training data and a lack of explainability. It stresses the need for rigorous data curation and the cost implications of developing safe medical AI systems. The authors caution against overenthusiastic adoption of AI in medicine and call for awareness of these limitations among medical professionals and stakeholders, underscoring the importance of a cautious and resource-intensive approach to AI in healthcare.

Peter Muennig, MD, MPH*

Professor, Health Policy and Management

Mailman School of Public Health

Columbia University

Chris Pirazzi

Computer scientist

***Corresponding Author:**

Peter Muennig, MD, MPH
722 W. 168th Street, 4th Floor
New York, NY 10032
Phone: +1 347-533-3415
Email: pm124@columbia.edu

Keywords: Artificial Intelligence (AI), Medical diagnosis, Data quality, Explainable AI (XAI), Healthcare technology.

Today's advanced artificial intelligence (AI) models, such as ChatGPT, can outscore most physicians on the MCAT. However, most AI models were not carefully built with billions in funding, so we must take great care to understand the limitations of AI as a tool for safe medical diagnosis. The major AI models we are familiar with were built with hundreds of thousands of human hours performing safety and accuracy checks. The same is not true of machine learning used in most medical applications branded as "AI." [1][2][3][4][5] Rather, they were sometimes built with inadequate training and on datasets with limitations, in some cases producing inaccurate diagnoses.

Over the last decade, numerous peer-reviewed medical journals reported that AI products outperformed humans at tasks such as reading medical images. [1][2][3][4][5] However, more recent attempts to deploy these systems in practice reveal shocking deficiencies that arose because we missed one basic fact: despite the name, AI models are not "intelligent." They are simply computer code designed to find correlations between inputs and outcomes, not unlike other correlational analyses used for hypothesis generation in the social sciences.

No dermatologist would mistake a benign mole for cancer just because it is marked with a pen, but at least one AI model will reliably do so. [1] No radiologist would base the diagnosis of pneumonia on whether the radiograph was taken by a portable emergency room scanner, but at least one AI product will. [2] In one study, an AI system was making diagnoses based primarily on the type and location of a hospital rather than actual clinical content, [3] identifying hospitals using scanner-specific image noise that humans can neither see nor remove. This system diagnosed pneumonia in the acromion process.[2] We are not arguing that all AI products sold to providers are dangerous, but rather that some of the products may not have been vetted "in the wild," where the predictions matter.

Another issue is that training data used often comes from affluent white males, leading to misdiagnoses in females and patients of color. [2][4] Yet, the greatest clinical utility of AI in its current state is to assist primary care providers in low-income settings or developing countries. Certainly, providers serving whites in rural America are also in need of help, but more diverse datasets are needed to improve the generalizability of AI systems going on the market.

How could products go to market that do not make accurate diagnoses? These AI applications did exactly what we designed them to do: find correlations. Correlations are subject to confounding by extraneous variables (hospital ID, machine type). Additional examples of spurious correlations can be found on Lurktech.[6]

To build a safe and effective AI diagnosis system, medical professionals need to weed out poor-quality, confound-laden data that software companies sometimes use to train these systems. [3] and then demand an AI system that can reveal exactly what image features it is using to make each diagnosis (so-called “Explainable AI”) and repeatedly check that the AI’s decisions are medically meaningful at every stage of product development.

How much does safe medical AI development cost? We can get a hint from large language models such as ChatGPT, which took years to develop with multi-billion-dollar investments and with feedback between large teams of human expert reviewers and model developers to handle biases and program in safety measures – and they still sometimes fabricate information. The massive monetary investment required to produce safe non-human medical assistants is too often overlooked by overly enthusiastic developers and hospital administrators eager to reduce costs.

As hundreds of articles on AI are published each month (the New England Journal of Medicine even released a journal dedicated to AI advances), editors, reviewers, and readers need to be aware of these pitfalls. These models are not vetted in the same way as those that can ace the MCAT. It is important to remember why medicine was built on randomized trials and not on associational studies: correlation does not equal causation.

References

1. [a, b, c] *Winkler, J. K., Fink, C., Toberer, F., et al. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA Dermatology, 155(10), 1135-1141. https://doi.org/10.1001/jamadermatol.2019.1735*
2. [a, b, c, d, e] *Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Medicine, 15(11), e1002683. https://doi.org/10.1371/journal.pmed.1002683*
3. [a, b, c, d] *Compton, R., Zhang, L., Puli, A., & Ranganath, R. (2023). When More is Less: Incorporating Additional Datasets Can Hurt Performance By Introducing Spurious Correlations. arXiv preprint. https://doi.org/10.48550/arXiv.2308.04431*
4. [a, b, c] *Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., et al. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27, 2176-2182. https://doi.org/10.1038/s41591-021-01595-0*
5. [a, b] *Roberts, M., Driggs, D., Thorpe, M., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3, 199-217. https://doi.org/10.1038/s42256-021-00307-0*
6. [^] *Pirazzi, C. Medical AI is going just great. Available online at: https://lurkertech.com/medai/. Accessed 11/30/2023.*