EDITORIAL

Automation Bias and Assistive AI
Risk of Harm From AI-Driven Clinical Decision Support

Rohan Khera, MD, MS; Melissa A. Simon, MD, MPH; Joseph S. Ross, MD, MHS

Related article page 2275

At the point of care, artificial intelligence (AI) algorithms have been developed to augment diagnostic decisions and suggest appropriate care pathways,[1] by leveraging complex information in a patient's electronic health record, such as imaging, documentation, and diagnostic testing. With an increasing number of technologies integrated into the diagnosis, management, and even treatment of patients,[2] the promise of AI to enhance accuracy, reduce errors, reduce clinician burnout, and improve clinical workflows may appear imminent.[3]

Most AI algorithms are designed to be assistive technologies—augmenting, not replacing, clinicians' decision-making.[4] AI models are imperfect and lack the broader clinical context that may be relevant for patient care. The expectation is that the diagnostic performance of clinicians supported by AI will exceed those of clinicians without such support.

In practice, however, clinicians are challenged by how to best interpret the information they receive from AI tools. Novel AI technologies are "black boxes" and clinicians may be unsure of whether or when to make a decision that runs counter to a recommendation based on the AI algorithm providing assistance. To address this, model developers have begun adding a layer of explainability so that clinicians can better interpret the model predictions and understand when models are relying on heuristics rather than clinically relevant data elements.[5] These heuristics can bias AI model predictions and may be the result of development in selective, nonrepresentative populations,[6] inadequate adherence to development best practices, and limited validation. The US Food and Drug Administration (FDA) has called for explainability of model outputs in its draft guidance addressing AI technologies for clinical decision support.[7]

In this issue of *JAMA*, Jabbour et al[8] evaluate the impact of explainability for AI model output on clinician diagnosis and further examine how systematically biased models may affect patient care supported by AI-based assistive diagnostic aids. In total, 457 clinicians from 14 US states were asked to respond to a respiratory distress–related clinical vignette and rate the probability of 3 possible diagnoses: chronic obstructive pulmonary disease, pneumonia, or heart failure. Each clinician's diagnostic performance was assessed over 9 vignettes: 2 baseline vignettes without AI support, 6 for which clinicians received AI support for interpretation of the chest radiograph, either with or without explainability metrics accompanying this interpretation, and 1 final vignette with support from an "expert clinician" consult. For the 2 baseline vignettes without AI support, diagnostic accuracy was 73%, whereas diagnostic accuracy for the final vignette was 81%, establishing the lower and upper bounds of average diagnostic accuracy for these vignettes.

In addition to testing the impact of explainability, this study also examined the impact of systematic bias. Among the 6 AI-supported vignettes, 3 reported predictive outputs from a standard model with a known accuracy of 75%, whereas the other 3 reported predictive outputs from a biased model that systematically ascribed a higher diagnostic probability for pneumonia based on advanced age and for heart failure based on high body mass index.

The results from Jabbour et al suggest that a more careful approach to evaluating AI tools is warranted before their rapid adoption, even when AI is used as assistive technology. For vignettes with AI support using the standard model, clinicians' diagnostic accuracy increased only modestly, from 73% without AI support to 76%. For vignettes with AI support using the standard model paired with explainability heatmaps highlighting the predictive areas on chest radiographs, diagnostic accuracy improved slightly more to 78%. However, for vignettes with AI support using the systematically biased model, clinicians' diagnostic accuracy dropped substantially to 62%. This large drop in performance was not remedied by explainability heatmaps that demonstrated inappropriate clinical sources for the predictions (ie, information from bones and soft tissues on the radiographic image, instead of lungs or heart). Even with this layer of explainability, clinician diagnostics accuracy only improved slightly (64%) and remained much lower than accuracy without any AI support.

These findings are concerning. Although the study highlights the potential value of explainability metrics to accompany assistive AI-based diagnostic tools, it also clearly illustrates the major challenge of clinicians' relying on assistive technologies, often referred to as *automation bias*.[9] Even in controlled settings, without the usual pressures on time, clinicians favored automated decision-making systems, relying on the AI-based tool, despite the presence of contradictory or clinically nonsensical information. If a model performs well for certain patients or in certain care scenarios, such automation bias may result in patient benefit in those settings. However, in other settings where the model is inaccurate—either systematically biased or due to imperfect performance—patients may be harmed as clinicians defer to the AI model over their own judgment. Worryingly, errors resulting from automation bias are likely to be further compounded by the usual time pressures faced by many clinicians.

For AI-based assistive technology developers, the study of Jabbour and colleagues illustrates the harm that may result from health system adoption of biased models. Having a "clinician-in-the-loop" overseeing the AI does not overcome the challenges of AI systems failing to provide accurate information, regardless of whether the source of the predictions is highlighted for clinicians. Therefore, the bar for clinical decision support using novel technology needs to reflect the challenges likely to be encountered in clinical practice.[10]

The study demonstrates that offering explainability metrics for predictions to clinicians, expecting they will then weigh that information before making a decision, may be ineffective. The limited value of explainability metrics in this study may reflect both the nature of explainability strategies used (ie, heatmaps) and that clinicians do not have the requisite training in evaluating these measures. As AI-based assistive technology is embedded in care systems, clinicians will need training on the interpretation of technology outputs, how to evaluate the quality of the provided information using available measures of explainability, and how to infer the common sources of bias, including derivation of data from nonrepresentative populations.

The study also forecasts broader challenges with emerging AI technology in more complex clinical scenarios. This study focuses only on clinical diagnosis using a radiographic image, which clinicians are routinely trained to read and interpret for respiratory distress. However, as care algorithms evolve to infer information not directly apparent to clinicians and rely on the ability of AI to identify complex hidden signatures,[11,12] there is a critical need for methodological innovation and clinician training that goes beyond explainability and provides true interpretability—where clinicians can infer the factors that resulted in the information.[13] In these settings, heatmaps are woefully inadequate because they are designed to simply make visible model heuristics (ie, shortcuts in making the predictions and whether predictions were based on information relevant to the clinical condition).[14]

The work of Jabbour and colleagues should inform the FDA as it evaluates and authorizes the use of a growing number of AI-based diagnostic tools. The current regulatory evaluation by the FDA is focused on model performance characteristics and its stability across different development and validation populations. Although this approach may ensure consistency of the model's performance across diverse populations, it does not fully capture the potential downstream negative consequences resulting from algorithmic assisted care. As clinicians react to information available from algorithms to decide on care pathways, all biases that are inherent in the model are further propagated and amplified.[15] To safeguard patients from unintended harm, evaluations of how clinicians interact with the model output, which downstream care decisions hinge on the algorithm and biases that arise with intended and unintended uses, are needed.

Clinical decision support tools based on imperfect AI assistive technologies have the potential to result in patient harm because clinicians may trust the output of AI tools over their own judgment. The bar for AI developers and regulatory agencies to put a product into clinical use must, therefore, be high. The task of interpreting the outputs of AI models cannot be off-loaded to clinicians, especially during a deluge of AI-driven tools that lack adequate controls, and better strategies are needed to go beyond explainability and to enable true interpretability. The future of AI-supported care is rapidly approaching, but the primary goal of implementing these tools—to improve patient care—must not be forgotten in the excitement over the technology.

ARTICLE INFORMATION

Author Affiliations: Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut (Khera); Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut (Khera); Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut (Khera, Ross); Associate Editor, *JAMA* (Khera, Simon); Department of Obstetrics and Gynecology, Northwestern Medicine Feinberg School of Medicine, Chicago, Illinois (Simon); Section of General Internal Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut (Ross); Deputy Editor, *JAMA* (Ross).

Corresponding Author: Rohan Khera, MD, MS, Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, 195 Church St, Sixth Floor, New Haven, CT 06510 (rohan.khera@yale.edu).

Conflict of Interest Disclosures: Dr Khera reported receiving grants from the National Heart, Lung, and Blood Institute (NHLBI) and the Doris Duke Charitable Foundation, and receiving grants to his institution from Bristol Myers Squibb and Novo Nordisk, being a cofounder of Evidence2Health and Ensight-AI outside the submitted work; holding a provisional patent (No. 63/428,569), related to the article. Dr Ross reported receiving grants from the FDA, Johnson & Johnson, the Medical Devices Innovation Consortium, the Agency for Healthcare Research and Quality, the NHLBI, and Arnold Ventures outside the submitted work; serving as an expert witness at the request of Relator's attorneys, the Greene Law Firm, in a qui tam suit alleging violations of the False Claims Act and Anti-Kickback Statute against Biogen Inc that was settled September 2022. No other disclosures were reported.

REFERENCES

1. Chen JH, Dhaliwal G, Yang D. Decoding artificial intelligence to achieve diagnostic excellence: learning from experts, examples, and experience. *JAMA*. 2022;328(8):709-710. doi:10.1001/jama.2022.13735

2. Wu K, Wu E, Theodorou B, et al. Characterizing the clinical adoption of medical AI through US insurance claims. *NEJM AI*. Published online November 9, 2023. doi:10.1056/AIoa2300030

3. Khera R, Butte AJ, Berkwits M, et al. AI in medicine—*JAMA's* focus on clinical outcomes, patient-centered care, quality, and equity. *JAMA*. 2023;330(9):818-820. doi:10.1001/jama.2023.15481

4. Bitterman DS, Aerts HJWL, Mak RH. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health*. 2020;2(9):e447-e449. doi:10.1016/S2589-7500(20)30187-4

5. Reddy S. Explainability and artificial intelligence in medicine. *Lancet Digit Health*. 2022;4(4):e214-e215. doi:10.1016/S2589-7500(22)00029-2

6. Shachar C, Gerke S. Prevention of bias and discrimination in clinical practice algorithms. *JAMA*. 2023;329(4):283-284. doi:10.1001/jama.2022.23867

7. Clinical decision support software: guidance for industry and Food and Drug Administration staff. US Food and Drug Administration. Published September 2022. Accessed November 15, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

8. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *JAMA*. Published December 19, 2023. doi:10.1001/jama.2023.22295

9. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect

mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121-127. doi:10.1136/amiajnl-2011-000089

10. Dorr DA, Adams L, Embí P. Harnessing the promise of artificial intelligence responsibly. *JAMA*. 2023;329(16):1347-1348. doi:10.1001/jama.2023.2771

11. Sangha V, Nargesi AA, Dhingra LS, et al. Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation*. 2023;148(9):765-777. doi:10.1161/CIRCULATIONAHA.122.062646

12. Holste G, Oikonomou EK, Mortazavi BJ, et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur Heart J*. 2023;44(43):4592-4604. doi:10.1093/eurheartj/ehad456

13. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. *Cardiovasc Diabetol*. 2023;22(1):259. doi:10.1186/s12933-023-01985-3

14. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9

15. Vaid A, Sawant A, Suarez-Farinas M, et al. Implications of the use of artificial intelligence predictive models in health care settings: a simulation study. *Ann Intern Med*. 2023;176(10):1358-1369. doi:10.7326/M23-0949