VIEWPOINT

Addressing Bias in Artificial Intelligence in Health Care

Ravi B. Parikh, MD, MPP
Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania.

Stephanie Teeple, BA
Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Amol S. Navathe, MD, PhD
Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania.

Author Audio Interview

Corresponding Author: Ravi B. Parikh, MD, MPP, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Blockley 1102, Philadelphia, PA 19104 (ravi.parikh@pennmedicine.upenn.edu).

Recent scrutiny of artificial intelligence (AI)–based facial recognition software has renewed concerns about the unintended effects of AI on social bias and inequity. Academic and government officials have raised concerns over racial and gender bias in several AI-based technologies, including internet search engines and algorithms to predict risk of criminal behavior. Companies like IBM and Microsoft have made public commitments to "de-bias" their technologies, whereas Amazon mounted a public campaign criticizing such research. As AI applications gain traction in medicine, clinicians and health system leaders have raised similar concerns over automating and propagating existing biases.[1]

But is AI the problem? Or can it be part of the solution? While potentially inadvertently contributing to bias, AI technologies, when used responsibly, may also help counteract the risk of bias in unique ways. Using AI to identify bias in health care may help identify interventions that could help correct biased clinician decision-making and possibly reduce health disparities.

Statistical and Social Bias in AI

Statistical bias refers to an algorithm that produces a result that differs from the true underlying estimate. Statistical bias is common in predictive algorithms for many reasons, including suboptimal sampling, measurement error in predictor variables, and heterogeneity of effects. For example, the Framingham Study risk factors have been used for decades to predict risk of cardiovascular disease. However, the original Framingham Study sampled from an overwhelmingly non-Hispanic white population. When applying the Framingham Risk Score to populations with similar clinical characteristics, the predicted risk of a cardiovascular event was 20% lower for black individuals compared with white individuals, indicating that the score may not adequately capture risk factors for some minority groups.[2]

Social bias in health care refers to inequity in care delivery that systematically leads to suboptimal outcomes for a particular group. Social bias could be caused by a statistically biased algorithm or by other human factors, including implicit or explicit bias. For example, clinicians may incorrectly discount the diagnosis of myocardial infarction in older women because these patients are more likely to present with atypical symptoms.[3] An AI algorithm that learns from historical electronic health record (EHR) data and existing practice patterns may not recommend testing for cardiac ischemia for an older woman, delaying potentially lifesaving treatment. Perhaps of more concern, clinicians may be more likely to believe AI that reinforces current practice, thus perpetuating implicit social biases.

Why Do AI Algorithms Automate and Perpetuate Bias?

Artificial intelligence and machine learning are limited by the quality of data on which they are trained. The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers. Generalizability and representativeness are also important considerations when interpreting randomized clinical trials.

However, the process by which the data are generated may be more important and particular to AI. If AI algorithms use data that are generated through a biased process, then the output may be similarly biased. This is a significant challenge when using clinical data sources like EHRs, insurance claims, or device readings because most of these data are generated as a consequence of human decisions. An algorithm to predict sepsis among patients in the emergency department, for example, may learn to use a test order for lactic acid to predict a poor outcome. However, the laboratory order may be more predictive of survival than the lactic acid value.[4] This is because a clinician is more likely to order the test for patients at risk of poor outcomes like death.

Artificial intelligence is also likely to incorrectly estimate risks for patients with missing data in the EHR. For example, among women with breast cancer, black women had a lower likelihood of being tested for high-risk germline mutations compared with white women, despite carrying a similar risk of such mutations.[5] Thus, an AI algorithm that depends on genetic test results is more likely to mischaracterize the risk of breast cancer for black patients than white patients.

While all predictive models may automate bias, AI may be unique in the extent to which bias is unrecognized (**Table**). Normally, clinicians have a pretest probability of an outcome and use the results of a diagnostic test to generate a posttest probability. However, clinicians may have a propensity to trust suggestions from AI decision support systems, which summarize large numbers of inputs into automated real-time predictions, while inadvertently discounting relevant information from nonautomated systems—so-called automation complacency.[6] For example, an AI-based early warning system can interpret changes in continuously monitored vital signs to alert an intensivist of a patient's impending clinical instability. However, AI-based decision support systems may produce a questionable or incorrect prediction. Hypothetically, an intensivist who is performing multiple concurrent tasks may inadvertently accept incorrect AI predictions unless there were obviously conflicting clinical information. This automation complacency could occur because AI predictions are framed around the outcome of interest and thus may be more salient to clinicians than an isolated test or laboratory result. Dedicated clinician training on interpreting AI outputs could ameliorate automation complacency.

Reducing Bias in AI

Although much of the discussion about AI and bias has focused on its potential for harm, strategies exist to mitigate such bias. When applied correctly, AI may be an effective tool to help counteract bias, an intractable problem in medicine.

Table. Artificial Intelligence Bias in Health Care

Example of Bias	Type of Bias	Potential Reasons for Bias	Methods to Address Bias
Low sensitivity of Framingham Risk Score in minority subgroups	Statistical	Algorithm training sample differs significantly from the population of interest	Oversample minority subgroups in training sample; tailor predictions or scores for specific subgroups
Delayed diagnosis of lung cancer in patients with low socioeconomic status or who lack transportation access to clinic	Social	Underlying disparities in diagnosis	Create flags for model uncertainty in predictions for certain high-risk subgroups
Missing data in electronic health record-based data sets due to lack of patient follow-up	Statistical and social	Missing data	Base predictions on "upstream" data at presentation of illness, not on subsequent follow-up data

First, AI decision support tools could be used to identify real-time bias in physician decision-making. Many nonmedical factors affect physician decision-making; situations with high cognitive load, such as decision-making at the end of a clinic day, are particularly prone to bias. If rational AI predictions and clinician decision-making differ in these situations, clinicians could be alerted in real time about decisions that are at risk of bias. For example, an AI algorithm may flag a possibly questionable opioid prescription at the end of a primary care clinician's day, providing a needed check on this decision. There are fledgling examples of using AI to identify disparities. When applied to unstructured data from psychiatry notes, AI algorithms demonstrated greater documentation of anxiety and chronic pain topics for white patients and psychosis topics for black, Hispanic, and Asian patients. Alerting clinicians to these disparities in documentation in real time could improve care of patients by making implicit biases in their practice more salient.[7]

Second, because most AI bias is related to the data-generating process, the primary solution may be to preferentially use unbiased data sources. Uniform collection of large amounts of data on all patients is now possible because of more routine use of noninvasive monitoring. Examples of relatively unbiased, uniform data sources include recorded vital sign data during surgical operations or triage data collected from the first hour after emergency department presentation, "upstream" of clinician judgments. Randomized trial data also could be used preferentially instead of observational data to support AI development, although it would be important to access which patients had been enrolled in the clinical trials. In many regards, the potential bias in AI is similar to concerns raised in clinical trials, in that participants are often nonrepresentative of the general population.

Other steps could help facilitate addressing bias in health care AI. For instance, existing standards, including the PROBAST tool to assess risk of bias in prediction models, can aid algorithm developers in selecting representative training sets and appropriate predictor variables.[8] In addition, algorithm predictions and subsequent actions could be tracked continuously to help ensure that outputs are not reinforcing existing social biases. Algorithm developers also could use certain sensitivity checks, including creating simulated data sets with high numbers of omitted variables and conducting counterfactual simulations, to determine how robust predictions are to omitted variable bias. For data sets that are necessarily collected after clinician decisions, algorithm developers could seek to oversample underrepresented populations to mitigate statistical bias.

Conclusions

Artificial intelligence is making its way into clinical practice. Because of its reliance on historical data, which are based on biased data generation or clinical practices, AI can create or perpetuate biases that may worsen patient outcomes. However, by strategically deploying AI and carefully selecting underlying data, algorithm developers can mitigate AI bias. Addressing bias could allow AI to reach its fullest potential by helping to improve diagnosis and prediction while protecting patients.

ARTICLE INFORMATION

Published Online: November 22, 2019. doi:10.1001/jama.2019.18058

Conflict of Interest Disclosures: Dr Parikh reported receipt of personal fees from GNS Healthcare. Dr Navathe reported receipt of grants from the Hawaii Medical Service Association, the Anthem Public Policy Institute, the Commonwealth Fund, Oscar Health, Cigna Corporation, and the Donaghue Foundation and personal fees from Navvis Healthcare, Agathos Inc, University Health System (Singapore), Elsevier Press, Navahealth, the Cleveland Clinic, the Medicare Payment Advisory Commission, and Embedded Healthcare; he reported being an uncompensated board member for Integrated Services Inc. No other disclosures were reported.

Funding/Support: This work was supported in part by the Penn Center for Precision Medicine (Dr Parikh) and the Pennsylvania Universal Research Enhancement (CURE) Program and Robert Wood Johnson Foundation (Dr Navathe).

Role of the Funders/Sponsors: The funders had no role in the preparation, review, or approval of the manuscript or decision to submit the manuscript for publication.

REFERENCES

1. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544-1547. doi:10.1001/jamainternmed.2018.3763

2. Gijsberts CM, Groenewegen KA, Hoefer IE, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One*. 2015;10(7):e0132321. doi:10.1371/journal.pone.0132321

3. Canto JG, Goldberg RJ, Hand MM, et al. Symptom presentation of women with acute coronary syndromes: myth vs reality. *Arch Intern Med*. 2007;167(22):2405-2413. doi:10.1001/archinte.167.22.2405

4. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361:k1479. doi:10.1136/bmj.k1479

5. McCarthy AM, Bristol M, Domchek SM, et al. Health care segregation, physician recommendation, and racial disparities in *BRCA1/2* testing among women with breast cancer. *J Clin Oncol*. 2016;34(22):2610-2618. doi:10.1200/JCO.2015.66.0019

6. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. 2010;52(3):381-410. doi:10.1177/0018720810376055

7. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167-E179. doi:10.1001/amajethics.2019.167

8. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376