Making decisions

Bias in artificial intelligence and data-driven diagnostic tools

CPD

Yves Saint James Aquino

Background
Although numerous studies have shown the potential of artificial intelligence (AI) systems in drastically improving clinical practice, there are concerns that these AI systems could replicate existing biases.

Objective
This paper provides a brief overview of 'algorithmic bias', which refers to the tendency of some AI systems to perform poorly for disadvantaged or marginalised groups.

Discussion
AI relies on data generated, collected, recorded and labelled by humans. If AI systems remain unchecked, whatever biases that exist in the real world that are embedded in data will be incorporated into the AI algorithms. Algorithmic bias can be considered as an extension, if not a new manifestation, of existing social biases, understood as negative attitudes towards or the discriminatory treatment of some groups. In medicine, algorithmic bias can compromise patient safety and risks perpetuating disparities in care and outcome. Thus, clinicians should consider the risk of bias when deploying AI-enabled tools in their practice.

ARTIFICIAL INTELLIGENCE (AI) has the potential to revolutionise clinical medicine, particularly in image-based diagnosis and screening.[1] Although AI applications are generally limited to augmenting clinician skills or helping with certain clinical tasks, full automation may be possible in the near future.[2] A more sophisticated set of AI approaches, called deep learning, has found success in detection tasks (eg determining the presence or absence of signs of a disease) and classification tasks (eg classifying cancer type or stage).[3] Lin et al[4] offer a summary of the ways in which AI will transform primary care, which include, among others, AI-enabled tools for automated symptom checking, risk-adjusted panelling and resourcing based on the complexity of the medical condition, as well as the automated generation of clinical notes based on clinician–patient conversations.

However, there are concerns that the benefits of AI in improving clinical practice may be hampered by the risk of AI replicating biases. A growing number of cases across industries show that some AI systems reproduce problematic social beliefs and practices that lead to unequal or discriminatory treatment of individuals or groups.[5] This phenomenon is referred to as 'algorithmic bias', which occurs when the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal impacts.[6]

Aim

This paper aims to provide a brief overview of the concept of algorithmic bias and how it may play out in the context of clinical decision making in general practice.

Understanding bias in AI

The first step in understanding algorithmic bias is dispelling the myth that technologies, especially AI enabled and data driven, are free of human values. AI relies on data generated, collected, recorded and labelled by humans. If AI systems remain unchecked, whatever biases exist in the real world that are embedded in the data will be incorporated into the AI algorithms.[5]

Cycle of biases

Algorithmic bias can be considered as an extension, if not a new manifestation, of a pernicious cycle of biases that include social, technological and clinician biases (Figure 1). Social bias refers to behaviours, beliefs or practices that treat individuals or groups unequally in an unjust manner. Social biases can manifest as prejudice, understood as negative attitudes or false generalisations felt or expressed towards an individual or group based on characteristics, such as race or ethnicity, sex or gender and socioeconomic status.[7] Social biases can also manifest as discrimination, which refers to practices that deny individuals or groups equality of treatment, usually involving actions that directly harm or disadvantage the target individual or group.[7]

Clinician bias refers to a set of cognitive tendencies of clinicians to make decisions based on incomplete information or subjective factors, or out of force of habit.[8] Yuen et al[8] describe common cognitive biases, including 'availability bias' and 'confirmation bias'. The former refers to making decisions based on what is immediately familiar to the clinician, whereas the latter refers to assigning unjustified preference to findings only because they confirm a diagnosis. Such cognitive biases could amplify health inequities resulting from broad social prejudices against marginalised groups. As with any member of society, clinicians are susceptible to culturally pervasive prejudicial beliefs that can manifest as implicit bias, or the unconscious association of negative attributes to an individual or a group.[9] In Australia, implicit bias may explain Aboriginal and Torres Strait Islander people's (hereafter respectfully referred to as Aboriginal people) experience of health inequities, particularly for chronic conditions. Evidence shows that implicit bias among practitioners tends to underestimate Aboriginal people's experience of pain, leading to less comprehensive assessment and subsequently delays in treatment because of mismanagement.[10]

Medical technologies are said to be biased when errors or outputs systematically lead to unequal performance among groups.[11] Studies have shown that some physical attributes or mechanisms of medical devices are biased against certain demographics.[11] For example, pulse oximeters that use light to measure blood oxygenation have been shown to be less accurate for people with darker skin tones.[12] The authors of a retrospective cohort study that examined pulse oximetry sensitivity among Black, Hispanic, Asian and White patients based on data collected from 324 centres in the US argue that 'systematic underdiagnosis of hypoxemia in Black patients is likely attributed to technical design issues, but the decision to tolerate the miscalibration for Black patients has been collective despite the available evidence'.[13] These findings show that unequal performance of medical devices, such as pulse oximeters, risks exacerbating health inequities that negatively impact the already marginalised groups.

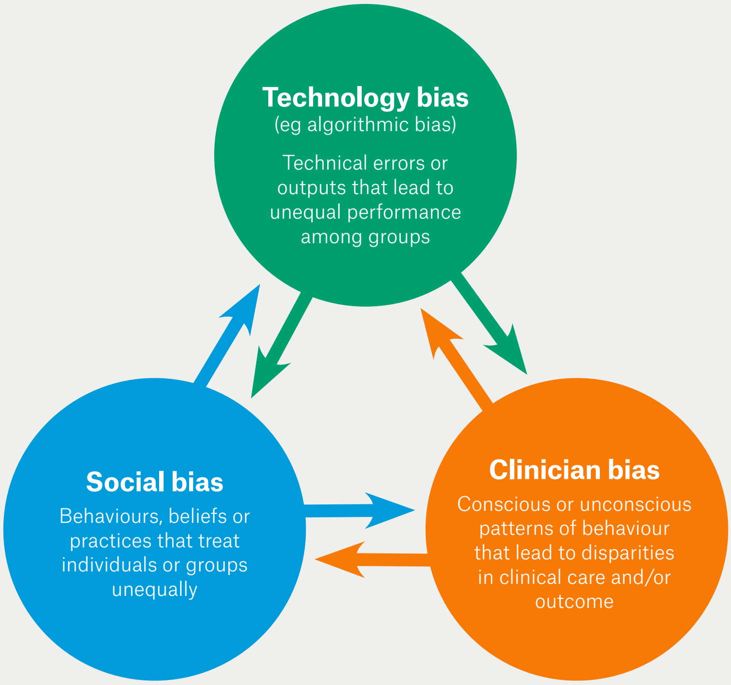
Technology bias
(eg algorithmic bias)
Technical errors or outputs that lead to unequal performance among groups
Social bias
Behaviours, beliefs or practices that treat individuals or groups unequally
Clinician bias
Conscious or unconscious patterns of behaviour that lead to disparities in clinical care and/or outcome

Figure 1. Cycle of biases.

From old to new

AI algorithms risk perpetuating existing social, technological and clinician biases by the continued use of datasets that do not represent real-world populations. Marginalised groups based on race, sex and sexuality have a long history of being absent or misrepresented in datasets,[14] which are typically coming from electronic health records or social surveys. AI algorithms based on non-representative datasets may perform accurate prediction, classification or pattern recognition specific to the majority groups they are trained with, but tend to have performance issues in recognising patterns outside the majority groups.[14] Evidence of algorithmic bias has been demonstrated in algorithms used to identify dermatological lesions that are often trained with images of lesions from White patients.[15] When tested on patients with a darker skin colour, the accuracy of the algorithms is 50% lower than what the developers claimed.[15] Other examples demonstrate sex-based disparities, such as the algorithms used for predicting cardiac events that are trained predominantly using datasets from male patients.[16]

In addition to replicating real-world bias through non-representative datasets, AI algorithms that are initially established as 'fair' may develop biases. These so-called latent biases, or biases 'waiting to happen', can occur in a number of ways.[17] One way biased performance occurs is when an AI algorithm is trained using datasets in one location (eg a local hospital in a high-income city). The algorithm could be proven to perform fairly in that location, but may turn out to be biased when transferred to another hospital, city or country when the algorithm interacts with local data. Another way for biased performance to occur is the introduction by AI algorithms of 'categorically new biases', biases that do not just mimic social biases, but 'perniciously reconfigure our social categories in ways that are not necessarily transparent to us, the users'.[18]

A more sophisticated set of approaches to AI called deep learning (eg artificial neural networks) have shown promise in extracting even more complex patterns of information from large datasets by using multiple processing layers.[14] Some of the areas in which deep learning applications have shown some success are for the classification of melanoma,[19] prediction of cardiovascular events[20] and COVID-19 diagnosis.[21] An increasing number of reports raise concerns that deep learning-based algorithms could amplify health disparities due to biases embedded in the training data. For example, a study by Larrazabal et al[22] examined a model based on deep neural networks for computer-aided diagnosis of thoracic diseases using X-ray images. The findings of that study showed a consistent decrease in performance when using male patients for training and female patients for testing (and vice versa).[22]

Countering algorithmic bias

Regulatory bodies, including the US Food and Drug Administration[23] and Australia's Therapeutic Goods Administration,[24] are in the process of establishing frameworks that specifically address the challenges raised by AI systems. Within the AI research and development community, experts are developing debiasing methods, which involve the measurement of biases followed by bias removal through a 'neutralise and equalise approach'.[25] Currently, however, the role of clinicians in minimising the risk (or impact, if unavoidable) of algorithmic bias remains an open question. One possible intervention in clinical practice is ensuring that clinicians continue to use their skills and judgment, including critical thinking and empathy, when using AI-enabled systems.[26] Clinicians should continue to challenge the myth that AI systems are completely objective and bias free.

Currently, there are very few procedural mechanisms consisting of checklists or specific steps to counter algorithmic bias. One of these is the Algorithmic Bias Playbook developed by Obermeyer et al at the Center for Applied AI at the University of Chicago Booth.[27] The playbook was intended for health managers (eg chief technical and chief medical officers), policymakers and regulators and includes a four-step process to guide the bias-auditing processes for health institutions.[27] At present, however, it remains challenging to develop procedural tools for individual general practitioners with no background in data science or AI.

Conclusion

AI systems have shown great promise in improving clinical practice, particularly in clinical tasks that involve identifying patterns in large, heterogeneous datasets to classify a diagnosis or predict outcomes. However, there are concerns that AI systems can also exacerbate existing problems that lead to health disparities. There is growing evidence of algorithmic bias, whereby some AI systems perform poorly for already disadvantaged social groups. Algorithmic bias contributes to a persistent cycle that consists of social bias, technological bias and clinician bias. As with other types of biases in medicine, algorithmic bias has practical implications for general practice: it can compromise patient safety, lead to over- or underdiagnosis, delays in treatment and mismanagement. Thus, clinicians should consider the risk of bias when using or deploying AI-enabled tools in their practice.

Key points

- AI systems have the potential to greatly improve clinical practice, but they are not free of errors and biases because they are built on datasets that are generated, collected, recorded and labelled by humans.
- Algorithmic bias refers to the tendency of some AI systems to perform poorly for disadvantaged or marginalised groups.
- One cause of algorithmic bias is the use of datasets that are not representative of real-world populations.
- Clinicians should be aware of the risk of algorithmic bias and seek information about datasets and evidence of performance when deploying AI-enabled tools in their practice.

Author

Yves Saint James Aquino MD, MRes, MSc, PhD, Research Fellow, Australian Centre for Health Engagement, Evidence and Values, School of Health and Society, University of Wollongong, Wollongong, NSW

Competing interests: YSJA is a committee member of Bellberry Limited, a national, private, not-for-profit Australian company that provides services as a Human Research Ethics Committee.

Funding: The author is part of The Algorithm Will See You Now project funded by the National Health and Medical Research Council (Ideas Grant Scheme 1181960). This project examines the ethical, legal and social implications of using AI for diagnosis and screening.

Provenance and peer review: Commissioned, externally peer reviewed.

Correspondence to:
yaquino@uow.edu.au

References

1. Aquino YSJ, Rogers WA, Braunack-Mayer A, et al. Utopia versus dystopia: Professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform 2023;169:104903. doi: 10.1016/j.ijmedinf.2022.104903.
2. Richards B, Sage Jacobson S, Aquino YSJ. Regulation of AI in health care: A cautionary tale considering horses and zebras. J Law Med 2021;28(3):645–54.
3 Nichols JA, Chan HWH, Baker MAB. Machine learning: Applications of artificial intelligence to imaging and diagnosis. Biophys Rev 2019;11(1):111–18. doi: 10.1007/s12551-018-0449-9.
4. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med 2019;34(8):1626–30. doi: 10.1007/s11606-019-05035-1.
5. Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems – an introductory survey. WIREs Data Mining Knowl Discov 2020;10(3):e1356.
6. Kordzadeh N, Ghasemaghaei M. Algorithmic bias: Review, synthesis, and future research directions. Eur J Inf Syst 2022;31(3):388–409.
7. Dovidio JF, Hewstone M, Glick P, Esses VM. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. In: Dovidio JF, Hewstone M, Glick P, Esses VM, editors. The SAGE Handbook of Prejudice, Stereotyping and Discrimination. Thousand Oaks: SAGE Publications, 2010; p. 3–28.
8. Yuen T, Derenge D, Kalman N. Cognitive bias: Its influence on clinical diagnosis. J Fam Pract 2018;67(6):366;368;370;372.
9. McDowell MJ, Goldhammer H, Potter JE, Keuroghlian AS. Strategies to mitigate clinician implicit bias against sexual and gender minority patients. Psychosomatics 2020;61(6):655–61. doi: 10.1016/j.psym.2020.04.021.

10. O'Brien P, Bunzli S, Lin I, et al. Addressing surgical inequity for Aboriginal and Torres Strait Islander people in Australia's universal health care system: A call to action. ANZ J Surg 2021;91(3):238–44.
11. Kadambi A. Achieving fairness in medical devices. Science 2021;372(6537):30–31. doi: 10.1126/science.abe9195.
12. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020;383(25):2477–78. doi: 10.1056/NEJMc2029240.
13. Valbuena VSM, Barbaro RP, Claar D, et al. Racial bias in pulse oximetry measurement among patients about to undergo extracorporeal membrane oxygenation in 2019-2020: A retrospective cohort study. Chest 2022;161(4):971–78. doi: 10.1016/j.chest.2021.09.025.
14. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns (N Y) 2021;2(10):100347. doi: 10.1016/j.patter.2021.100347.
15. Kamulegeya LH, Okello M, Bwanika JM, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. BioRxiv 2019. doi: 10.1101/826057.
16. van Smeden M, Heinze G, Van Calster B, et al. Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. Eur Heart J 2022;43(31):2921–30. doi: 10.1093/eurheartj/ehac238.
17. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc 2020;27(12):2020–23. doi: 10.1093/jamia/ocaa094.
18. Waller RR, Waller RL. Assembled bias: Beyond transparent algorithmic bias. Minds and Machines 2022;32(3):533–62. doi: 10.1007/s11023-022-09605-x.
19. Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019;119:11–17. doi: 10.1016/j.ejca.2019.05.023.
20. Cheung CY, Xu D, Cheng CY, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. Nat Biomed Eng 2021;5(6):498–508. doi: 10.1038/s41551-020-00626-4.
21. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for COVID-19. J Big Data 2021;8(1):18. doi: 10.1186/s40537-020-00392-9.
22. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci USA 2020;117(23):12592–94. doi: 10.1073/pnas.1919012117.
23. Ebrahimian S, Kalra MK, Agarwal S, et al. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. Acad Radiol 2022;29(4):559–66. doi: 10.1016/j.acra.2021.09.002.
24. Goergen SK, Frazer HM, Reddy S. Quality use of artificial intelligence in medical imaging: What do radiologists need to know? J Med Imaging Radiat Oncol 2022;66(2):225–32. doi: 10.1111/1754-9485.13379.
25. Schlender T, Spanakis G. 'Thy algorithm shalt not bear false witness': An evaluation of multiclass debiasing methods on word embeddings. In: Baratchi M, Cao L, Kosters WA, Lijffijt J, van Rijn JN, Takes FW, editors. Artificial Intelligence and Machine Learning. BNAIC/Benelearn 2020. Communications in Computer and Information Science, Vol. 1398. Cham: Springer, 2021. doi: 10.1007/978-3-030-76640-5_9.
26 Royce CS, Hayes MM, Schwartzstein RM. Teaching critical thinking: A case for instruction in cognitive biases to reduce diagnostic errors and improve patient safety. Acad Med 2019;94(2):187–94. doi: 10.1097/ACM.0000000000002518.
27. Obermeyer Z, Nissan R, Stern M, et al. Algorithmic bias playbook. Center for Applied AI at Chicago Booth, 2021. Available at www.ftc.gov/system/files/documents/public_events/1582978/algorithmic-bias-playbook.pdf [Accessed 20 March 2023].