npj | digital medicine

www.nature.com/npjdigitalmed

EDITORIAL OPEN

Check for updates

Bias in AI-based models for medical applications: challenges and mitigation strategies

Artificial intelligence systems are increasingly being applied to healthcare. In surgery, AI applications hold promise as tools to predict surgical outcomes, assess technical skills, or guide surgeons intraoperatively via computer vision. On the other hand, AI systems can also suffer from bias, compounding existing inequities in socioeconomic status, race, ethnicity, religion, gender, disability, or sexual orientation. Bias particularly impacts disadvantaged populations, which can be subject to algorithmic predictions that are less accurate or underestimate the need for care. Thus, strategies for detecting and mitigating bias are pivotal for creating AI technology that is generalizable and fair. Here, we discuss a recent study that developed a new strategy to mitigate bias in surgical AI systems.

npj Digital Medicine (2023)6:113 ; https://doi.org/10.1038/s41746-023-00858-z

BIAS IN MEDICAL AI ALGORITHMS

Artificial intelligence (AI) technology is increasingly applied to healthcare, from AI-augmented clinical research to algorithms for image analysis or disease prediction. Specifically, within the field of surgery, AI applications hold promise as tools to predict surgical outcomes[1], aid surgeons via computer vision for intraoperative surgical navigation[2], and even as algorithms to assess technical skills and surgical performance[1,3–5].

Kiyasseh et al.[4] highlight this potential application in their work deploying surgical AI systems (SAIS) on videos of robotic surgeries from three hospitals. They used SAIS to assess the skill level of surgeons completing multiple different surgical activities, including needle handling and needle driving. In applying this AI model, Kiyasseh et al.[4] found that it could reliably assess surgical performance but exhibited bias. The SAIS model showed an underskilling or overskilling bias at different rates across surgeon sub-cohort. Underskilling was the AI model downgrading surgical performance erroneously, predicting a particular skill to be lower quality than it actually was. Overskilling was the reverse—the AI model upgraded surgical performance erroneously, predicting a specific skill to be of higher quality than it was. Underskilling and overskilling were measured based on the AI-based predictions' negative and positive predictive values negative, respectively.

STRATEGIES TO MITIGATE BIAS

The issue of bias being exhibited, perpetuated, or even amplified by AI algorithms is an increasing concern within healthcare. Bias is usually defined as a difference in performance between subgroups for a predictive task[6,7]. For example, an AI algorithm used for predicting future risk of breast cancer may suffer from a performance gap wherein black patients are more likely to be assigned as "low risk" incorrectly. Further, an algorithm trained on hospital data from German patients might not perform well in the USA, as patient population, treatment strategies or medications might differ. Similar cases have already been seen in healthcare systems[8]. There could be many different reasons for this performance gap. Bias can be generated across AI model development steps, including data collection/preparation, model development, model evaluation, and deployment in clinical settings[9]. With this particular example, the algorithm may have been trained on data predominantly from white patients, or health records from Black patients may be less accessible. Additionally, there are likely underlying social inequalities in healthcare access and expenditures that impact how a model might be trained to predict risk[6,10]. Regardless of the cause, the impact of an algorithm disproportionately assigning false negatives would include fewer follow-up scans, and potentially more undiagnosed/untreated cancer cases, worsening health inequity for an already disadvantaged population. Thus, strategies to detect and mitigate bias will be pivotal to improving healthcare outcomes. Bias mitigation strategies may involve interventions such as pre-processing data through sampling before a model is built, in-processing by implementing mathematical approaches to incentivize a model to learn balanced predictions, and post-processing[11]. Further, as experts can be aware of biases specific to datasets, "keeping the human in the loop" can be another important strategy to mitigate bias.

With their SAIS model, Kiyasseh et al.[4] developed a strategy called TWIX to mitigate bias. TWIX is an add-on application that taught the SAIS model to add a prediction of the importance of video clips that was used to assess surgical skill. They hypothesized that the SAIS model's bias might be due to the system latching onto unreliable video frames for assessment. TWIX requiring model predictions of video clip importance served a similar role to human assessors explaining the rationale for assessments. Kiyasseh et al.[4] found that TWIX mitigated SAIS model bias, improving model performance both for the disadvantaged surgeon sub-cohorts and for surgical skill assessments overall. This accomplishment is beneficial not only for this particular use case but also implies that this type of bias mitigation strategy could be used to continue to improve AI applications in the future.

A LOOK INTO THE FUTURE—CHALLENGES WITH CONTINUOUSLY LEARNING AI MODELS

Bias within AI algorithms must continue to be studied and mitigated as AI technology develops. Looking into the future, one question that will most definitely arise is what level of bias is acceptable for an AI algorithm[4]. This is analogous to the question of what accuracy threshold is acceptable for a particular AI system[4]. Previous groups suggested that any performance discrepancy is indicative of algorithmic bias, but expecting completely bias-free systems before implementation

is unrealistic[12]. Performance discrepancy may also differ based on the data and population an AI algorithm is trained on and then subsequently applied to. Currently, there is significant heterogeneity in terms of the datasets AI algorithms are trained with within algorithm types themselves[13,14]. The question of whether AI algorithms may need to be more generalizable, trained on larger and more diverse datasets to be applied to broader populations, or more localized and applied narrowly remains to be addressed. In any case, AI models will have to be explainable[15] with transparent methodologies so that these questions can be studied and debated in the coming years.

Another issue for the future is whether AI algorithms will be able to be changed/edited, just as Kiyasseh et al.[4] added TWIX to their existing SAIS algorithm. An AI algorithm can either be locked—once the algorithm is trained, the model provides the same result when the same input is applied—or adaptive[16]. In this case, the AI model could be updated continuously as it learns from new data over time rather than becoming outdated within a few years. However, continuous learning also possesses the risk of increasing or adding new bias if the new data are biased[17]. Thus, methodologies for regular bias detection and continual bias mitigation will be key to AI implementation.

From a regulatory standpoint, new initiatives also aim to tackle the issue of biased data in AI systems. The STANDING Together initiative (standards for data diversity, inclusivity, and generalizability), launched in September 2022, aims to develop recommendations for the composition (who is represented) and reporting (how they are represented) of datasets underpinning medical AI systems[18]. Further, the FDA has recognized challenges due to bias in AI and ML algorithms and released an action plan ("Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan") in January 2021[9,19], emphasizing the importance of identifying and mitigating bias in AI systems[9]. As part of the FDA Action Plan, the FDA intends to support the piloting of real-world performance monitoring[19], allowing for the detection of bias after deployment. Further, to meet regulatory challenges that come with continuously adopting AI models, the FDA recently released a draft guidance to develop a less burdensome regulatory approach supporting the iterative improvement of, e.g., AI models while continuing to assure their safety and effectiveness[20]. These types of regulatory steps should be encouraged, as they will become increasingly necessary to ensure the minimization of bias without the blockade of AI innovation.

CONCLUSION

The integration of AI into medical technology and healthcare systems is only going to increase in the coming years. Key to AI model integration and usability will be bias mitigation. Kiyasseh et al. describe an innovative approach to bias mitigation with their TWIX system. As technology continues to develop, the push toward bias mitigation occurs at all levels—from model development and over training to deployment and implementation. This effort will require checks and balances from innovators, healthcare institutions, and regulatory entities.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 27 April 2023; Accepted: 6 June 2023;
Published online: 14 June 2023

Mirja Mittermaier[1,2]✉, *Marium M. Raza*[3] *and Joseph C. Kvedar*[3]
[1]*Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Infectious Diseases, Respiratory Medicine and Critical Care, Berlin, Germany.* [2]*Berlin Institute of Health at Charité—Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.* [3]*Harvard Medical School, Boston, MA, USA.* ✉*email: Mirja.mittermaier@charite.de*

REFERENCES

1. Ma, R. et al. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digital Med.* **5**, 187 (2022).
2. Chadebecq, F., Vasconcelos, F., Mazomenos, E. & Stoyanov, D. Computer vision in the surgical operating room. *Visc. Med.* **36**, 456–462 (2020).
3. Kiyasseh, D. et al. A multi-institutional study using artificial intelligence to provide reliable and fair feedback to surgeons. *Commun. Med.* **3**, 42 (2023).
4. Kiyasseh, D. et al. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *NPJ Digital Med.* **6**, 54 (2023).
5. Kiyasseh, D. et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng*. https://doi.org/10.1038/s41551-023-01010-8 (2023).
6. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
7. Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y. & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Med.* **6**, 55 (2023).
8. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
9. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Commun. Med.* **1**, 25 (2021).
10. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* **9**, 010318 (2019).
11. Xu, J. et al. Algorithmic fairness in computational medicine. *EBioMedicine* **84**, 104250 (2022).
12. Townson, S. *Manage AI Bias Instead of Trying to Eliminate It.* https://sloanreview.mit.edu/article/manage-ai-bias-instead-of-trying-to-eliminate-it/2023 (MIT Sloan Management Review, 2023).
13. Gubatan, J. et al. Artificial intelligence applications in inflammatory bowel disease: emerging technologies and future directions. *World J. Gastroenterol.* **27**, 1920–1935 (2021).
14. Moglia, A., Georgiou, K., Georgiou, E., Satava, R. M. & Cuschieri, A. A systematic review on artificial intelligence in robot-assisted surgery. *Int. J. Surg.* **95**, 106151 (2021).
15. Theunissen, M. & Browning, J. Putting explainable AI in context: institutional explanations for medical AI. *Ethics Inf. Technol.* **24**, 23 (2022).
16. Benjamens, S., Dhunnoo, P. & Mesko, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digital Med.* **3**, 118 (2020).
17. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **27**, 2020–2023 (2020).
18. Ganapathi, S. et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat. Med.* **28**, 2232–2233 (2022).
19. FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. www.fda.gov/media/145022/download (2021).
20. FDA. *Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial (2023).

ACKNOWLEDGEMENTS

M.M. is a fellow of the BIH—Charité Digital Clinician Scientist Program funded by the Charité—Universitätsmedizin Berlin, the Berlin Institute of Health at Charité, and the German Research Foundation (DFG).

AUTHOR CONTRIBUTIONS

M.M. wrote the first draft. M.M.R. contributed to the first draft and provided critical revisions. J.C.K. provided critical revisions. All authors critically reviewed and revised the manuscript and approved the final manuscript.

COMPETING INTERESTS

J.C.K. is the Editor-in-Chief of *npj Digital Medicine*. M.M. and M.M.R. declare no competing interests.