Generic Contrast Agents
Our portfolio is growing to serve you better. Now you have a *choice*.
FRESENIUS KABI
VIEW CATALOG

AJNR

Understanding Bias in Artificial Intelligence: A Practice Perspective

Melissa A. Davis, Ona Wu, Ichiro Ikuta, John E. Jordan, Michele H. Johnson and Edward Quigley

AJNR Am J Neuroradiol 2024, 45 (4) 371-373
doi: https://doi.org/10.3174/ajnr.A8070
http://www.ajnr.org/content/45/4/371

This information is current as of May 21, 2025.

Understanding Bias in Artificial Intelligence: A Practice Perspective

Melissa A. Davis, Ona Wu, Ichiro Ikuta, John E. Jordan, Michele H. Johnson, and Edward Quigley

ABSTRACT

SUMMARY: In the fall of 2021, several experts in this space delivered a Webinar hosted by the American Society of Neuroradiology (ASNR) Diversity and Inclusion Committee, focused on expanding the understanding of bias in artificial intelligence, with a health equity lens, and provided key concepts for neuroradiologists to approach the evaluation of these tools. In this perspective, we distill key parts of this discussion, including understanding why this topic is important to neuroradiologists and lending insight on how neuroradiologists can develop a framework to assess health equity–related bias in artificial intelligence tools. In addition, we provide examples of clinical workflow implementation of these tools so that we can begin to see how artificial intelligence tools will impact discourse on equitable radiologic care. As continuous learners, we must be engaged in new and rapidly evolving technologies that emerge in our field. The Diversity and Inclusion Committee of the ASNR has addressed this subject matter through its programming content revolving around health equity in neuroradiologic advances.

ABBREVIATIONS: AI = artificial intelligence; ASNR = American Society of Neuroradiology; TAT = turnaround time

Many artificial intelligence (AI) tools currently in clinical practice involve neuroimaging, including tools for detection, acquisition, and segmentation. It is important for neuroradiologists to evaluate these tools for clinical efficacy and safety, including how the use of these tools will impact patient care and workflow. There is ample literature to help neuroradiologists understand the basic principles and technology of AI and how to approach the evaluation and validation of AI tools. Although the original literature focused on the scientific development process continues to evolve, there is increasing interest in the potential biases of these types of learning applications.[1]

Health disparities in neurologic diseases are well-characterized and cross many sociodemographic variables, including race, socioeconomic status, and insurance status. The effects on population health are highlighted through the study of social determinants of health, which can serve as key drivers of health disparities.[1] Such disparities can have negative compounding effects on the health care continuum and, ultimately, patient outcomes. Understanding AI through the lens of health equity is necessary to recognize bias that might be introduced in AI algorithms and to mitigate biases that can occur.

In the fall of 2021, several experts in this space delivered a Webinar hosted by the American Society of Neuroradiology (ASNR) Diversity and Inclusion Committee, focused on expanding the understanding of bias in AI and provided key concepts for neuroradiologists to approach evaluation of these tools. In this perspective, we distill key insights from the dynamic discussion that ensued. The source Webinar is available as enduring content on the ASNR Education Connection (https://www.pathlms.com/asnr/courses/56243/video_presentations/268414#).

Why Is It Important for Neuroradiologists to Care about Bias in AI?

As long as AI algorithms are relegated to the role of cognitive assistant, algorithmic bias might not be a pressing issue for neuroradiologists. Once AI models are used to predict outcomes, manage care, or order workflow, potential bias in the collection or labeled training data needs to be considered. One can imagine that under-representation of populations in the training data can lead to inaccurate predictions of outcomes. One example is an application using algorithms to segment out the "core" infarct on CT perfusion imaging.[2] If the core infarct volume is larger than a certain threshold, some studies have suggested that there will be no benefit to the patient from endovascular treatment and therefore the patient should be excluded.[3] However, subsequent

Received August 29, 2023; accepted after revision October 17.

From Yale University (M.A.D., M.H.J.), New Haven, Connecticut; Massachusetts General Hospital (O.W.), Charlestown, Massachusetts; Mayo Clinic Arizona, Department of Radiology (I.I.), Phoenix, Arizona; Stanford University School of Medicine (J.E.J.), Stanford, California; and University of Utah (E.Q.), Salt Lake City, Utah.

Please address correspondence to Melissa A. Davis, MD, MBA, Yale University, 330 Cedar Street, New Haven, CT 06510; e-mail: Melissa.a.davis@yale.edu

http://dx.doi.org/10.3174/ajnr.A8070

studies have shown that core infarcts predicted by acute CTP fail to manifest on follow-up noncontrast CT after successful recanalization, especially when patients are treated early.[2] In this situation, incorrect prediction of the algorithm of a large-volume infarction might have precluded a beneficial treatment if treatment was guided purely by a computer algorithm.

There is increasing interest in using generative algorithms for many clinical neuroimaging applications, ranging from increasing resolution on low-resolution data to being able to convert one technique to another, such as a CT-to-MR imaging conversion. The synthetics are very effective and can even recreate susceptibility artifacts. The accuracy of such transformations depends highly on the training data. If there is bias in the data, inaccurate transformations can arise. This issue is highlighted by the Face-Depixelizer example (Tg-bomze/Face-Depixelizer; https://github.com/tg-bomze/Face-Depixelizer). Like methods that convert low-resolution images to high-resolution images, the Face-Depixelizer can take a compressed low-resolution image and generate an image with quality similar to that of the original image. This technique holds great promise for applications, ranging from data storage to streaming. However, it was quickly discovered that the Depixelizer failed for nonwhite faces, which were incorrectly transformed into faces with white features.[4]

Tools to Approach Understanding Bias

By integrating fairness into the machine learning lifecycle, we can mitigate unfairness. The machine learning lifecycle can be simplified to the model-development phase, deployment phase, and feedback reverting to the development phase to refine the model performance. By defining fairness requirements and by involving diverse stakeholders at the model-development stage, potential bias can be mitigated. Sources of bias include definition of the task, data set construction, and cost-function for the training algorithm. Poor task and cost-function definitions can lead inadvertently to racially biased machine learning algorithms. Imbalances in the training data can lead to underdiagnoses in the under-represented data set. For models that focus on positive or negative predictive values, the prevalence of the disease in the training and validation data should match real-world clinical distribution. A detailed discussion on the consequences of data mishandling is found in a review by Rouzrokh et al.[5]

One method to potentially mitigate problems using machine learning algorithms for clinical decision-making is explainability. For example, to explain the classification of hemorrhage subtypes,[6] attention maps were used to highlight which features had the greatest weights in the decision of the algorithm. The authors showed that regions corresponding to SAH and intraventricular hemorrhage were appropriately highlighted. In contrast, there is the risk of data leakage in which features indirectly associated with disease prevalence are incorrectly interpreted by the model to be the disease itself. This was demonstrated in a pneumonia classification program that focused on metallic tokens as the most relevant feature in patients with pneumonia instead of the lungs because the tokens could be used to identify which hospital provided the data, ie, the hospital more likely to have pneumonia cases.[7] Although the algorithm nominally had high accuracy with its training and testing data, the algorithm did not actually learn the true, relevant features in the lungs and would likely fail when deployed at other hospital centers.

We must always consider such effects in daily practice and the influence of these technologies on our decision-making, impacting patient care.[8] Consider the following examples.

Examination Triage

AI algorithms may aid in the detection of emergent findings such as acute intracranial hemorrhage and cervical spine fractures. O'Neill et al[9] showed that simply marking an examination as having an emergent finding does not affect report turnaround time (TAT), but grouping all marked examinations at the top of the reading list does result in improved TAT. However, we might take a step back and look at the bias of an AI system and how that might propagate unfairness. As an example, acute intracranial hemorrhage detection was initially thought to have high sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. Yet, subsequent clinical implementations of the same algorithm at other institutions showed a positive predictive value of 81.3%, with decreased performance for older patients.[10] If an older patient with an acute intracranial hemorrhage and a false-negative AI result waited longer for his or her scan to be read due to a triage system set up around AI results, that scenario would not be equitable.

Examination Scheduling

Maximizing examinations performed in a day results in maximal profits for a radiology practice. Identifying delays, potential "no-shows," and potential times for additional examinations could optimize scheduling and potentially improve patient satisfaction. Pianykh et al[11] showed that while several AI models have been used to predict examination times/delays, these models all decline in quality with time. Various factors could affect the model such as seasonal/migrant workers, a sudden influx of refugees, major industry closure with layoffs and loss of health care benefits, or a global pandemic. These factors could be addressed with the implementation of continuous-learning AI that repeatedly uses updated data input from hospital information systems to dynamically address fluctuations in population health and identify barriers to health care that can be addressed, such as lack of dependable transportation, lack of health insurance, lack of childcare, or an inability to pay.

DISCUSSION

In this perspective, we have introduced resources to aid the neuroradiologist in learning and contemplating the intersection of AI and health equity. By leveraging examples of clinical workflow implementation of these tools, we can begin to see how AI tools will impact discourse on equitable radiologic care. As end users of these tools, we are responsible for understanding potential pitfalls and implicit biases that may affect our ability as physicians and neuroradiologists to deliver equitable high-quality care to our patients.

As continuous learners, we must be engaged especially as new and rapidly evolving technologies emerge in our field. AI is the

newest of these advances, and there is an urgent need to remain focused on health equity within radiology as we begin to leverage and automate these rapidly evolving tools. The Diversity and Inclusion Committee of the ASNR has taken on this task in collaboration with the Computer Science and Informatics Committee and the Artificial Intelligence Committee. Through this programming content, learners can access in-depth discussions regarding health equity in neuroradiologic advances.

CONCLUSIONS

Since 2020, the ASNR Diversity and Inclusion Committee has hosted Webinars spanning medical-social objectives to core science objectives. These types of Webinars allow neuroradiologists to engage in digestible content like bias in AI. The Webinar discussed in this article focused on the intersection of health equity and bias, with the goal of introducing imaging experts to these concepts in meaningful ways.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Davis MA, Lim N, Jordan J, et al. **Imaging artificial intelligence: a framework for radiologists to address health equity, from the AJR Special Series on DEI.** *AJR Am J Roentgenol* 2023;221:302–08 CrossRef Medline
2. Martins N, Aires A, Mendez B, et al. **Ghost infarct core and admission computed tomography perfusion: redefining the role of neuroimaging in acute ischemic stroke.** *Interv Neurol* 2018;7:513–21 CrossRef Medline
3. Albers GW, Marks MP, Kemp S, et al; DEFUSE 3 Investigators. **Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging.** *N Engl J Med* 2018;378:708–18 CrossRef Medline
4. Menon S, Damian A, Hu S, et al. **Pulse: self-supervised photo upsampling via latent space exploration of generative models.** *arXiv* 2020 CrossRef https://arxiv.org/abs/2003.03808. Accessed June 23, 2023
5. Rouzrokh P, Khosravi B, Faghani S, et al. **Mitigating bias in radiology machine learning, 1: data handling.** *Radiol Artif Intell* 2022;4: e210290 CrossRef Medline
6. Lee H, Yune S, Mansouri M, et al. **An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets.** *Nat Biomed Eng* 2019;3:173–82 CrossRef Medline
7. Zech JR, Badgeley MA, Liu M, et al. **Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study.** *PLoS Med* 2018;15:e1002683 CrossRef Medline
8. Filippi CG, Stein JM, Wang Z, et al. **Ethical considerations and fairness in the use of artificial intelligence for neuroradiology.** *AJNR Am J Neuroradiol* 2023;44:1242–48 CrossRef Medline
9. O'Neill TJ, Xi Y, Stehel E, et al. **Reprioritization of the reading worklist using artificial intelligence has a beneficial effect on the turnaround time for interpretation of head CT with intracranial hemorrhage**. *Radiol Artif Intell* 2021;3:e200024 CrossRef Medline
10. Voter AF, Meram E, Garrett JW, et al. **Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage.** *J Am Coll Radiol* 2021;18:1143–52 CrossRef Medline
11. Pianykh OS, Langs G, Dewey M, et al. **Continuous learning AI in radiology: implementation principles and early applications.** *Radiology* 2020;297:6–14 CrossRef Medline