Fairness metrics for health AI: we have a long way to go

Amarachi B. Mbakwe,[a] Ismini Lourentzou,[a] Leo Anthony Celi,[b,c,d,*] and Joy T. Wu[e]

[a]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
[b]Institute of Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA
[d]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[e]Department of Radiology, Stanford Medicine, Stanford University, Stanford, CA 94305, USA

eBioMedicine
2023;90: 104525
Published Online 14 March 2023
https://doi.org/10.1016/j.ebiom.2023.104525

The use of Artificial Intelligence (AI) is on track to revolutionize healthcare, with performance in medical tasks such as clinical diagnosis often being comparable to expert-level accuracy, at least in the laboratory. AI can play a significant role in healthcare, enabling clinicians to make more accurate and timely diagnoses and devise effective treatment plans. However, the amplification of pre-existing healthcare inequity with the use of AI models is a legitimate concern. Recent works have shown that medical imaging AI models can easily encode and learn patient-sensitive characteristics[1] and cause significant performance disparities between patient subgroups.[2] Therefore, it is encouraging to see more attempts, such as from Glocker et al.,[3] that evaluate methods for assessing how sensitive patient information, such as ethnicity and sex, are encoded and possibly used in model predictions. Unfortunately, for many diagnostic and prognostic clinical applications, the "ground truth" used for fairness assessment metrics may already be embedded with biases and laced with suboptimal outcomes that are not explained by clinical features. As such, the medical AI community needs to go beyond solely evaluating the clinical readiness of AI models with metrics that are predicated on potentially biased and constantly shifting clinical ground truth.

A critical step towards addressing AI model bias and subgroup disparities is the establishment of common principles, guidelines, and standards that model developers adhere to. These standards would need to emphasize the importance of fairness and transparency in AI systems' design and deployment. Proper documentation of model performance across patient subgroups is a minimum requirement. Depending on the clinical use case, models should be designed and evaluated with additional impact metrics that consider existing health inequities and possible harm for disadvantaged subgroups. Recent work by MEDFAIR,[4] a benchmark for building and evaluating fair medical imaging models, is a contribution towards this. An ideal guideline would need to cover requirements for appropriate debiasing techniques and evaluation metrics for different sources of bias. These include, but are not limited to, bias arising from dataset composition, model feature encoding, the use of learned demographic features (also known as shortcut features), and bias in ground truth labels.

Datasets can encode bias, such as from underrepresentation of already disadvantaged subgroups. Clinician bias can also be reflected in data and learned by AI. In medical images, bias may even be introduced from access to different quality scanners. These biases in the data should be documented, e.g., by using "datasheets for datasets".[5] Federated learning methods can also aid in training/tuning model(s) on more varied databases from different parts of the world and/or from underrepresented subgroups. Moreover, dataset bias mitigating strategies may be helpful, including dataset preprocessing, e.g., reweighing unintended features so that they are statistically independent of the target/outcome label.[6] However, it is unclear how well these methods work for medical images.

Model feature encoding is another source of bias. AI models can identify race and sex from medical images across modalities and use these characteristics to detect diseases, even when such characteristics are not associated with the diagnosis.[1] Even after removing sensitive information from datasets, which may not even be possible for medical images, models can still encode and use other correlated features for prediction. The "fairness through awareness" framework[7] shows why we cannot assume sensitive information has been expunged from a dataset. The framework also offers a metric-based approach for ensuring that a model's labeling of similar individuals is indeed similar.

Furthermore, models can inherit disparities from medical data through learning to depend on correlations between unrelated input features (e.g., nonbinary gender, immigration status), and the predicted outcomes. Glocker et al. highlighted difficulties in detecting what information is used in model predictions,[3] despite trying a range of methods from transfer learning, multitask learning, and unsupervised exploration of feature representations. Besides these methods, algorithmic transparency, explainability and interpretability,[8,9] focus instead on understanding how encoded input features are used for model prediction. Without an in-depth understanding of what features AI models use in making predictions, the promise of AI may not be realized.

DOI of original article: https://doi.org/10.1016/j.ebiom.2023.104467
*Corresponding author.
E-mail address: lceli@mit.edu (L.A. Celi).

Few have explored metrics that quantify the effect of training on potentially biased ground truth labels. The closest in the fairness literature involves social welfare functions[6] that aim to capture the underlying social phenomena and inequities when the model learns from data. More work is needed to develop metrics that are not completely reliant on ground truth labels for assessing readiness of medical imaging AI tools. Short of such metrics, intra- and post-processing de-biasing techniques may help reduce subgroup performance disparity. An example was employed in recent work on neural network pruning and fine-tuning for chest X-ray classifiers.[10]

AI in healthcare is intended to improve access to quality healthcare, especially for those who are marginalized. It is worrisome to find evidence across many works that these models utilize non-clinical demographic attributes and are likely to propagate existing disparities. Current attempts to understand how imaging models encode and use non-clinical demographic information for prediction are encouraging, but are still limited. More interdisciplinary communication and collaboration between AI researchers, healthcare providers, social scientists, and the public would be needed to advance fairness, transparency, and accountability of medical imaging models.

Contributors

All the authors participated in the outline development, writing and editing of the manuscript.

Declaration of interests

L.A.C. received support for attending meetings and/or for travel by Massachusetts Institute of Technology, and cloud credits from Amazon, Google, and Oracle. The other authors have no conflicts of interest to declare.

References

1 Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modeling study. *Lancet Digit Health*. 2022;4(6):e406–e414.
2 Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176–2182.
3 Glocker B, Jones C, Bernhardt M, Winzeck S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *eBioMedicine*. 2023;89:104467. https://doi.org/10.1016/j.ebiom.2023.104467.
4 Zong Y, Yang Y, Hospedales T. MEDFAIR: benchmarking fairness for medical imaging. In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. 2023.
5 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM*. 2021;64(12):86–92.
6 Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst*. 2012;33(1):1–33.
7 Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 2012:214–226.
8 Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput Biol Med*. 2022;140:105111.
9 Jungmann F, Ziegelmayer S, Lohoefer FK, et al. Algorithmic transparency and interpretability measures improve radiologists' performance in BI-RADS 4 classification. *Eur Radiol*. 2023;33(3):1844–1851.
10 Marcinkevics R, Ozkan E, Vogt JE. Debiasing deep chest x-ray classifiers using intra-and post-processing methods. In: *Machine Learning for Healthcare Conference*. PMLR; 2022:504–536.