Machine learning

https://doi.org/10.1038/s41591-024-02892-0

Using unlabeled data to enhance fairness of medical AI

Rajiv Movva, Pang Wei Koh & Emma Pierson

AI models for tasks such as pathology and dermatology struggle to generalize to new patient groups or hospitals that they were not trained on; learning more robust features from unlabeled data could prevent overfitting to the training distribution and thereby increase fairness.

Despite the number of studies that report expert-level performance of clinical machine learning algorithms[1], clinicians, computer scientists and regulators remain cautious. At the heart of their uncertainty lies the question of generalization: an algorithm with near-perfect performance on its training data may not perform as well when tested on new data with different properties, an issue known as 'distribution shift'. Distribution shifts can lead to health inequities: a dermatology model trained primarily on lighter-skinned white patients may perform less well on race groups with different skin pigmentation[2]. Because it is challenging to obtain large, labeled medical imaging datasets, models are often trained on relatively small and specific labeled datasets, and consequently fail to generalize[3].

A promising recent paradigm for combatting distribution shifts involves training models on unlabeled datasets to learn generalizable representations of data[4]. Unlabeled datasets are often much larger and more diverse than labeled datasets, exposing models to more variation — which helps them to generalize. In this issue of *Nature Medicine*, articles by Vaidya et al.[5] and Ktena et al.[6] present two approaches that exemplify this paradigm: Vaidya et al.[5] apply a model trained on a large, unlabeled dataset to extract more generalizable image features, and Ktena et al.[6] train a model on unlabeled data to generate synthetic images used to augment the labeled dataset. Both approaches yield improvements in accuracy, and improve generalization to new hospitals and minority patient groups who are less represented in training sets.

Vaidya et al.[5] train a model to predict cancer subtype from whole-slide microscopy images, a task which is crucial to guide treatment and predict prognosis. To predict cancer subtype, their model uses an encoder to extract features from the images, then makes a prediction from the extracted features. Their key finding is that the use of a naive encoder that is trained on non-medical images results in substantial disparities across patient race groups in terms of accuracy. However, the authors improve performance and reduce disparities by replacing the non-medical encoder with an encoder trained on a far larger dataset of unlabeled pathology images spanning diverse tissue types, cell morphologies, staining conditions, image resolutions, and tasks[7]. Because it is trained on a diverse array of relevant images, the pathology encoder learns a richer feature set that is more robust to distribution shifts than the non-medical image encoder.

Ktena et al.[6] take a different approach to improve generalization and fairness. They focus on predicting metastatic breast cancer (from histopathology images), chest conditions (from X-rays) and skin carcinomas (from photos). Their key innovation is to train their model not just on their original training set, but on 'synthetic' images that are created by training an image generation model on unlabeled data, thus augmenting their original dataset. Where possible, they sample more synthetic images from groups that were underrepresented in the original dataset, increasing demographic diversity. This approach works well: their model performs better on images from different contexts, such as hospitals that were not seen in training or patient groups with different demographic makeup. Overall, their method mitigates accuracy disparities across sex and race groups.

Vaidya et al.[5] and Ktena et al.[6] provide two illustrations of a critical point for medical machine learning: we can improve generalization by training models not just from small, labeled, task-specific datasets, but also from unlabeled datasets that can be much larger and more diverse. Vaidya et al.[5] use a pathology-specific encoder that aggregates data from many different settings, some of which may look more like the condition in which we want to deploy the model; Ktena et al.[6] generate synthetic images that look like the new distribution. Both approaches incorporate information from unlabeled data to improve generalization. More broadly, unlabeled data is only one of many outside sources of information that we can incorporate to improve generalization of models trained on small, labeled datasets. For example, other work makes use of outside information such as the overall prevalence of a disease[8] or prior knowledge about the distribution shift[9]. Incorporating outside information may be particularly useful in low-resource settings, in which it can be difficult to obtain enough clinician-labeled data from a new hospital or patient group.

The studies also have implications for reducing bias in clinical algorithms. Many generic methods for bias reduction have been

developed, but these methods can harm accuracy in some medical settings[10]. Vaidya et al.[5] and Ktena et al.[6] suggest that incorporating outside information (in this case from unlabeled data) can improve accuracy and reduce bias simultaneously, by uplifting performance for all groups — and might represent a promising approach even when generic bias methods do not perform well. Substantiating this, Vaidya et al.[5] test several existing methods for bias reduction and find that they have minimal effects on fairness, compared with using the pathology-specific encoder.

These studies[5,6] also reiterate the importance of evaluating algorithmic performance on subgroups of the population, because evaluating only on the entire population can conceal subgroup disparities. Such evaluations remain challenging, owing in part to lack of available data. Therefore, a high priority must be the continued collection of large, diverse samples, including from underserved populations, which must be made widely available to researchers[11,12]. The approaches developed by Vaidya et al.[5] and Ktena et al.[6] focus primarily on widely known benchmarks and well-resourced hospitals; future work should assess whether the methods explored in these papers will prove useful for other deployment settings, such as hospitals outside of the USA. Datasets should be annotated with relevant clinical and demographic variables, mitigating the limitations of some existing datasets (for example, a lack of granular race data or social determinants of health data[13]). Reviewers and regulators should encourage subgroup evaluations. Such evaluations will shed light on algorithms that perform inequitably and fail to generalize, encouraging the development of methods such as those pursued by Vaidya et al.[5] and Ktena et al.[6], and aiding the development of robust and equitable algorithms.

Rajiv Movva[1], Pang Wei Koh[2] & Emma Pierson[1] ✉
[1]Department of Computer Science, Cornell Tech, New York, NY, USA.
[2]Department of Computer Science, University of Washington, Seattle, WA, USA.
✉e-mail: emma.pierson@cornell.edu

Published online: 19 April 2024

References

1. Rajpurkar, P. et al. *Nat. Med.* **28**, 31–38 (2022).
2. Kinyanjui, N. M. et al. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (eds. Martel, A. L. et al.) Vol. 12266, 320–329 (2020).
3. Quiñonero-Candela, J. et al. (eds.) *Dataset Shift in Machine Learning* (MIT Press, 2022).
4. Sagawa S. et al. In *Int. Conference on Learning Representations* https://go.nature.com/3wRZsW6 (2022).
5. Vaidya, A. et al. *Nat. Med.* https://doi.org/10.1038/s41591-024-02885-z (2024).
6. Ktena, I. et al. *Nat. Med.* https://doi.org/10.1038/s41591-024-02838-6 (2024).
7. Chen, R. J. et al. Preprint at https://doi.org/10.48550/arXiv.2308.15474 (2023).
8. Balachandar, S., Garg, N. & Pierson, E. In *Int. Conference on Learning Representations* https://go.nature.com/4a7Pqyg (2024).
9. Gao, I. et al. In *Proc. 40th International Conference on Machine Learning* 202, 10800–1083, (PMLR, 2023).
10. Pfohl, S. R. et al. *J. Biomed. Inform.* **113**, 103621 (2021).
11. The All of Us Research Program Investigators. *N. Engl. J. Med.* **381**, 668–676 (2019).
12. Mullainathan, S. & Obermeyer, Z. *Nat. Med.* **28**, 897–899 (2022).
13. Movva, R. et al. In *Proc. 8th Machine Learning for Healthcare Conference* **219**, 443–472 (PMLR, 2023).

Competing interests

The authors declare no competing interests.