https://doi.org/10.1038/s43586-024-00371-x

A data-centric perspective to fair machine learning for healthcare

Haoran Zhang, Walter Gerych & Marzyeh Ghassemi

Check for updates

Machine learning models are increasingly being deployed in real-world clinical settings and have shown promise in patient diagnosis, treatment and outcome tasks. However, such models have also been shown to exhibit biases towards specific demographic groups, leading to inequitable outcomes for under-represented or historically marginalized communities.

Many prior works propose algorithmic approaches to address these fairness issues. However, applying debiasing algorithms effectively requires first recognizing the underlying biases in the data (Fig. 1). Often, there is a disconnect between those developing the models, who may be skilled in algorithmic debiasing, and those curating the data, who possess the knowledge of where biases exist and how they manifest. This highlights the need for a data-centric approach to fair machine learning in healthcare. By focusing on identifying and addressing biases at their source, rather than relying solely on post-hoc algorithmic fixes, we can build machine learning models that are not only fair but also generalize across different healthcare environments and populations and over time.

Not learning bias is better than fixing it

Healthcare datasets contain historical biases that reflect long-standing disparities in medical treatment and diagnosis. For example, women are less likely to be correctly diagnosed with heart disease than men and are instead more frequently misdiagnosed with mental health conditions[1]. Machine learning models trained on such datasets subsequently learn these patterns, resulting in propagation of these biases. Large language models (LLMs) like GPT-4 have exhibited stereotypes related to race, ethnicity and sex in medical diagnoses[2]. These biases originate from the unfiltered internet data used for training, ranging from PubMed abstracts to Reddit posts.

Addressing these biases after they have been learned by a model, or during the learning process, is a challenging task that often results in partial or ineffective solutions. For example, debiasing methods that superficially reduce sex bias in word embeddings fail to address deeper biases in the embedding space that can resurface in downstream modelling[3]. One alternative approach is to eliminate these biases at the source – within the data itself – by removing or downweighting biased entries or by collecting new, unbiased data. One study predicting occupation from biographies demonstrated that removing low-quality, ambiguous and mislabelled samples improved downstream model fairness by reducing the occupations' associations with gender stereotypes[4].

Another data-centric intervention is generating synthetic unbiased data to counteract historical biases in under-representation[5]. Although this is not as effective as collecting real representative data, these models have improved both fairness and accuracy in medical image classification across sex and race, addressing short-term gaps in representation.

Beware of proxies in data and labels

Clinical machine learning often relies on predicting proxies of the target variable. For instance, healthcare cost may be used as a proxy for illness severity owing to the assumption that sicker individuals incur more costs. Proxies are difficult to avoid because outcomes like the severity of an illness are difficult to quantify, whereas healthcare cost is easily accessible. However, proxies can also be a source of bias. For example, using healthcare cost as a target was shown to underestimate the severity of illness for Black patients owing to systemic biases. Specifically, Black patients have reduced healthcare expenditures compared with white patients with the same level of need[6].

Proxies may also exist in the covariates, known as spurious correlations. For example, a model trained to infer mortality in patients with pneumonia predicted that individuals with asthma were less likely to die than people without[7]. This was true in the training data – patients with asthma received more care, resulting in fewer deaths. This is undesirable because the amount of medical attention a patient receives is an unobserved and invariant feature, while the decreased risk of patients with asthma is specific to the training data and unlikely to generalize. Crucially, the model having learned a proxy is not apparent until its predictions and associated training data are critically analysed.

There is a need for methods that mitigate the risks of biases that arise from proxies. Using interpretable models[7] or having human experts that specify causal relations and identify when model predictions break common sense[8] are steps towards this goal. For instance, a human-in-the-loop approach could identify that a model predicting a lower pneumonia risk score for patients with asthma is likely to be a spurious signal or, conversely, flag higher mortality risks for Black mothers as systemic bias that may be useful in decision-making. While causal models could ideally disentangle spurious correlations from true causal relationships, they are challenging in practice owing to observational limitations, unverifiable assumptions and the lack of controlled interventions in healthcare data.

Ultimately, a diverse team must carefully consider each step of the model training and deployment pipeline, including what features to collect, the conditions of collection and integrating additional data sources to form a robust view of the system. The All of Us dataset, which emphasizes data collection from under-represented groups, exemplifies these robust data collection practices. Efforts to standardize datasets with datasheets that describe the data collection context, cohort selection, intended use and known biases can empower this decision-making and the creation of more robust models.

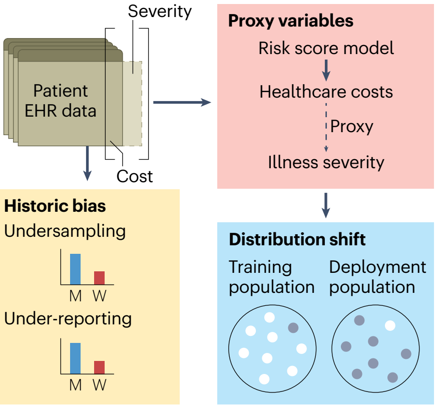

Fig. 1 | Common sources of bias in healthcare datasets that are best resolved with a data-centric approach. These include historic bias, the use of proxy variables and distribution shifts based on the training and deployment populations. EHR, electronic health record; M, men; W, women.

Probing data under distribution shift

Even for models that are fair and performant in distribution on the training domain, distribution shifts – changes in the data distribution between the training and deployment environments –remain a barrier to maintaining fairness during deployment. In healthcare, distribution shifts occur frequently owing to changes in patient characteristics, evolving medical practices, deployments in new hospitals or the introduction of new diseases such as COVID-19. Such distribution shifts can cause the model fairness to behave unpredictably[9]. Maintaining fairness under distribution shift is crucial for model reliability in deployment.

The first step to diagnose loss of fairness is to probe the shift in data distribution. Prior works have suggested deducing the structure of these shifts by conditional independence[10]. When the cause of the failure of fairness transfer is known, mitigation strategies can be effectively applied to target such shifts. This data-centric approach enables a more in-depth understanding of fairness issues and facilitates more robust solutions that can adapt to changing healthcare landscapes.

Given the potential for distribution shifts and the safety-critical nature of healthcare models, we advocate for restrictions on how, where and when deployed machine learning models can be used to make predictions in healthcare settings. By limiting the manner, place and time of model use, we can ensure that fairness is maintained across different contexts and populations. Data-centric approaches are critical to concretely define the boundaries of these restrictions, by characterizing the distributions within the training data and considering the magnitude and impact of potential shifts.

Haoran Zhang, Walter Gerych & Marzyeh Ghassemi ✉
Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, Cambridge, MA, USA.
✉e-mail: mghassem@mit.edu

Published online: 27 November 2024

References

1. Maserejian, N. N. et al. Disparities in physicians' interpretations of heart disease symptoms by patient gender: results of a video vignette factorial experiment. *J. Womens Health* **18**, 1661–1667 (2009).
2. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
3. Gonen, H. & Goldberg, Y. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J. et al.) 609–614 (Association for Computational Linguistics, 2019).
4. Schröder, S. et al. Measuring fairness with biased data: a case study on the effects of unsupervised data in fairness evaluation. in *Advances in Computational Intelligence. IWANN 2023* vol. 14134 (eds Rojas, I. et al.) 134–145 (Springer, 2023).
5. Ktena, I. et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30**, 1166–1173 (2024).
6. Obermeyer, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
7. Caruana, R. et al. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (ACM, 2015).
8. Srivastava, M., Hashimoto, T. & Liang, P. Robustness to spurious correlations via human annotations. in *Proceedings of the 37th International Conference on Machine Learning* Vol. 119, 9109–9119 (PMLR, 2020).
9. Yang, Y. et al. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
10. Schrouff, J. et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. in *Advances in Neural Information Processing Systems* Vol. 35, 19304–19318 (NeurIPS, 2022).

Acknowledgements

This work was supported in part by a National Science Foundation (NSF) 22-586 Faculty Early Career Development Award (no. 2339381), a Gordon & Betty Moore Foundation award and a Google Research Scholar award.

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.