https://doi.org/10.1038/s42256-023-00651-3

Translating intersectionality to fair machine learning in health sciences

Elle Lett & William G. La Cava

Check for updates

Fairness approaches in machine learning should involve more than an assessment of performance metrics across groups. Shifting the focus away from model metrics, we reframe fairness through the lens of intersectionality, a Black feminist theoretical framework that contextualizes individuals in interacting systems of power and oppression.

There has been an explosion of research using machine learning (ML) to optimize health interventions. With this increase, concerns have risen that ML-based technologies may exacerbate health inequities[1]. In fair ML, investigators develop approaches that prevent models from disproportionately harming already oppressed and excluded populations. A fundamental challenge to the field is defining (un)fairness itself. In practice, fair ML focuses on eliminating differences in model performance across groups defined by a subset of demographic traits. However, we argue that this oversimplification has limited use in preventing ML models from becoming an adverse digital health determinant. Populations subjected to severe inequities in healthcare access, treatment and outcomes experience many intersecting systems of power and oppression. Furthermore, equilibrating model performance across groups does not guarantee equitable health outcomes when ML tools are deployed.

Intersectionality is particularly suited to address these challenges based on the two 'arms' of the framework: critical inquiry and critical praxis[2]. Critical inquiry relates to how we capture the effect of societal-level discrimination in modelling, and how, and for whom, (un)fairness is measured. Critical praxis requires expanding fairness beyond the narrow lens of model performance metrics, motivating us to identify more equitable approaches throughout the ML pipeline, including task definition, feature engineering, data processing, model training, validation, deployment and updating.

Translating core principles to fairness in ML

Collins and Bilge[2] articulate six core ideas for intersectionality (Table 1). For illustration, we consider the hypothetical task of predicting cardiovascular events among a cohort of US hospital patients inclusive of Black transgender women. The first two ideas – social inequality and intersecting power relations – are best understood together. In relation to our task, the social inequalities in access to routine, high-quality primary care and health insurance for Black transgender individuals are due, in part, to intersecting oppressive power systems such as racism[3] and transphobia[4]. In addition, understanding intersecting power relations requires a recognition of the multilevel nature of discrimination. On an interpersonal level, transgender individuals face discrimination and bias that results in avoidance, denial, or poorer quality healthcare. On a structural level, Black individuals are disproportionately segregated into 'food deserts' – geographical regions in which residents have limited access to affordable and nutritious food (such as fresh produce), with a related increased likelihood of adverse cardiovascular outcomes[5]. These inequalities and power relations directly map onto bias in ML as characteristics of the generating mechanism for training data. Decreased access to and frequency of healthcare leads to underrepresentation and increased missingness in training data[1]. Providers directly impact data quality when practicing biased care that varies treatment assignment or outcomes by social identities[6]. Together, these processes that generate social inequalities also coalesce to create data that biases models.

Social context relates to transportability of ML models. Power and oppression vary spatiotemporally. Anti-Black racism in the USA has unique manifestations, particularly in the form of racialized police violence[7]. For our hypothetical task, beyond mortality and injury effects of police violence, there are potential deleterious mental health[8] and gendered physical health effects on blood pressure and diabetes[9] that, if measured properly, may improve the accuracy of cardiovascular outcome predictions for Black trans women in the USA. However, that model would not transport to predictions for Black trans women in Brazil or the UK where the specific manifestations of anti-Black racism differ. Social context similarly varies on the subnational level, impeding transportability of models between regions within a country.

Relationality and complexity have broad implications for ML and fairness. The former emphasizes connectedness among social identities and systems, dissolving rigid boundaries between constructs such as race and class and highlighting how they are co-constituted: a racialized system is inherently classist and gendered. This concept is strongly related to intersecting power systems but also highlights the challenges of interpretability in ML; particularly for demographic and social inequality measures, it may be challenging to parse the individual contribution of a single feature to predictive accuracy.

Complexity emphasizes the intrinsic challenges of applying intersectionality, including selecting among the various definitions of fairness. For example, statistical parity, in which the prediction rate for an outcome must be equivalent, may be inappropriate when baseline class membership varies substantially by group, such as in our hypothetical task with cardiovascular disease. Equalizing false positive and/or false negative rates may be more appropriate. However, these definitions have theoretical trade-offs, both with overall accuracy[10] and between definitions[11], so selection must be tailored to the research question. Notably, recent empirical work has shown that large fairness gains can be made with negligible accuracy losses across diverse data and health policy applications, reinforcing the case for building fairness-aware models[12]. Complexity also suggests that some scenarios are inappropriate for ML tools; the real-world context of discrimination may preclude building an ML model that is sufficiently equitable to avoid

Table 1 | Intersectionality core ideas for ML researchers

Intersectionality core idea	Implications for ML and fairness
Social inequalities	**Data generating mechanism:** Training data exhibits health inequities due to social inequalities (such as wealth, education and housing stability) that are driven by interconnected socio-structural systems of power and oppression.
Intersecting power relations and relationality	
Social context	**Generalizability:** Models built on a biased sample of participants subject to only a subset of the social contexts of the target population (for example, predominantly white, cisgender samples) will not generalize to the entire population **Transportability:** Models built in one social context, such as predictions for Black individuals in the southeastern USA, may not transport to another, such as Black individuals in the Pacific Northwest.
Relationality	**Interpretability:** Systems of discrimination and oppression are inter-related and co-constituted such that it may be difficult to parse the individual contributions to predictive accuracy of corresponding features.
Complexity	**Measuring (un)fairness:** Selecting the appropriate fairness definitions in the model fitting step must be tailored to the specific prediction task, social context and data. **Discretion:** Some use cases may not be appropriate for ML if data cannot sufficiently represent marginalized groups or tools cannot be fairly deployed.
Social justice	**Community participation:** Incorporate and centre individuals from marginalized backgrounds throughout the ML pipeline **Impact:** Use post-deployment studies to determine whether the benefits of ML tools are experienced equitably across groups and if corresponding health inequities are being decreased.

causing harm to populations already made vulnerable by intersecting power relations.

The last core idea, social justice, is straightforward: the goal of fair ML should be equitable health impacts. Ideally, rather than eliminating differential model bias, healthcare ML should reduce health inequities and, for our hypothetical task, reduce the excess burden of cardiovascular disease on Black trans women.

Community participation

Intersectionality centres oppressed and excluded communities as the 'source' of knowledge on how systems of discrimination impact their lives and their health. The current status quo of researchers defining prediction tasks without community input systematically excludes the perspectives of marginalized groups. Consistent with the social justice tenet, intersectional fairness requires that we use community-based participatory research (CBPR) frameworks and enable non-academics to help define the prediction task and oversee the development and implementation pipeline[13]. CBPR approaches must include adequate compensation for labour provided by community research partners to ensure that the process is equitable and non-extractive[14].

Training dataset construction

Poor representation of marginalized communities in training data is a primary source of model bias[1]. Most healthcare-related ML tools are built on data from academic health systems, which often serve populations that differ from community hospitals. Deploying models trained on data that excludes marginalized groups can amplify existing health inequities. Therefore, we need to re-imagine dataset construction to prospectively address representation deficits. Academic centres can pool data from nearby community hospitals with similar social contexts to increase the sample size of intersectional marginalized groups. Importantly, there is a potential trade-off with overall prediction accuracy as pooled data sources become more dissimilar, but this may be tempered by improvements in group-specific prediction accuracy, particularly among populations that often carry the highest disease burden. Defining which populations to enrich for in training data should be based on the specific disease context, prediction task, and intervention. For example, a model trained for predicting triple-negative breast cancer treatment response should enrich for Black women with the disease, as they are subject to a disproportionate incidence and mortality burden[15].

Data pre-processing

Pre-processing features related to social contexts is an exercise in political power. The common practice of collapsing underrepresented groups decides who 'counts' and to whom a model must be fair. For Indigenous populations in the USA, the collapse of Native Americans into a heterogenous 'Other Race' category, or their exclusion from analysis, has contributed to their erasure from public health statistics and the scientific record[16]. Regarding ML fairness, such practices obscure model biases that impact minoritized communities. These practices are enforced under the guise of statistical sample size limitations and become default without interrogation. We advocate for disaggregation and transparent reporting of how demographic data are treated in ML models with emphasis on potential biases introduced by pre-processing. Disaggregation must be tailored, emphasizing groups who are marginalized within the specific context of the prediction task and implementation environment while balancing privacy concerns to prevent introducing new harms.

Feature engineering

Most ML fairness focuses on social identities (such as race and gender) and algorithms that satisfy group fairness constraints, imposing (near) equality of some metric across groups defined by shared demographic traits. This approach flattens the multilevel interfaces of power and privilege (such as racism and sexism) into individual characteristics. However, social identities function as imperfect proxies for social context, limiting the predictive power of models built exclusively on these features.

In public health and sociology there is extensive literature on measuring racism as a multidimensional system and process[17], with extensions to sexism[18] and other forms of discrimination. These approaches conceptualize discrimination as latent constructs estimated by linking several data sources on social inequalities (such as economic resources, housing access, carceral data) and/or laws and policies at various levels of geographic granularity. Recent work has begun to illustrate how measures of social determinants of health can improve predictive accuracy of ML models leaving room for continued expansion of similar approaches[19]. Also worth noting are recent causal approaches that conceptualize fairness as multi-level with macro-level causes impacting model performance for individuals based on protected attributes[20]. These approaches are unified in that they attempt to capture the complexity of how socio-structural systems interact with individuals to produce health and contribute to model (un)fairness.

Model training

Group fairness definitions and algorithms are commonly used to optimize ML models. These approaches have three common limitations: (1) single-axis definitions of fairness; (2) dichotomization of privilege; and (3) group size dependence. The first limitation is most common: constraining fairness based on groups defined by a single protected attribute only accommodates a single axis of discrimination. Even among group fairness definitions that are multi-axis, there is a theory–practice gap due to model fitting software that only allow one attribute, regardless of the definition[21].

Dichotomization of privilege is another oversimplification of discrimination. Within a protected attribute, the severity of discrimination may vary between classes. Therefore, intersectionality requires fairness definitions that accommodate heterogeneity in violations along protected attributes. For example, in the USA, anti-Black and anti-Indigenous racism is uniquely pervasive and manifests across police brutality[7], chronic illnesses, and politics[3] in ways that are not as severe for other ethnoracial groups. Some approaches would collapse all minoritized ethnoracial groups into a single 'unprivileged' group[21]. As a result, fairness violations among these groups are treated equivalently, regardless of different experiences of discrimination. This dichotomization of privilege violates principles of intersectionality and fails to optimize accuracy for populations that are most vulnerable to harm.

Some fairness definitions consider several protected attributes simultaneously, in principle accounting for multiple axes of power and moving toward intersectional fairness. However, all incorporate a group size dependence that deprioritizes intersectional groups who are underrepresented in the training data. There are three common remedies: (1) including a population frequency weight in the fairness measure[22]; (2) imposing a threshold that excludes small groups from the fairness measure/algorithm[23]; and (3) specifying a Bayesian prior that smooths fairness estimates for small groups[24].

These approaches control overfitting by improving the stability of fairness metric estimates. Without these constraints, estimates among groups with small sample sizes are less likely to generalize to future data. This highlights a tension between the theory of intersectionality and the pragmatic considerations of statistical computation. Intersectionality centres and even prioritizes the many marginalized individuals who exist at the convergence of intersecting power systems. By contrast, for statistical necessity, these approaches de-emphasize or even exclude those very groups.

Validation, deployment and updating

As with training data curation, validation datasets should enrich for populations most at risk for harm. Specifically, we advocate for purposeful recruitment, data collection and pooling to increase the representation of marginalized groups in validation datasets. In addition, investigators should report performance metrics for each intersectional position, so that end-users know for whom it is most valid. For example, for a hypothetical model to identify patients who will not maintain antiretroviral therapy for HIV in the USA, the validation data might purposively sample Black women, who represented the greatest proportion of new HIV cases among women in 2018[25]. Oversampling may inflate positive predictive value for these groups, which underscores the need for intersectional position-specific reporting of validation metrics. Recent work has also shown that using group-specific thresholds to equilibrate recall across groups can produce fair positive predictive value rates[12].

Post-deployment studies are necessary to determine the effects of ML models. Clinical decision-making is multifactorial and integrates perspectives from patients, providers, administrators and payers, such that even 'statistically' fair models can widen health inequities. Therefore, health systems should conduct audits to ensure that the benefits from ML technologies are distributed equitably and, if not, collaborate with implementation scientists to identify system failures that drive inequities. Ideally, integration of new ML technology should be governed by community advisory boards of potential patients likely to be impacted. Impact evaluations should be continuous to account for model drift. Stakeholders should collaborate to pre-specify criteria for updating models or retiring them for severe fairness violations. These practices will ensure that ML does not worsen health inequities and may reduce them.

Conclusion

Fair ML has disproportionately focused on statistical definitions, fitting algorithms and metrics, without situating the field in the context of an unjust society in which model outputs have consequences that can compound health inequities. We adapt intersectionality to fair ML through its two arms: (1) inquiry, emphasizing how we quantify and correct for algorithmic injustice in models; and (2) praxis, identifying processes that promote justice in the generation and implementation of new technologies throughout the ML pipeline. We hope intersectional ML fairness can extend fair ML from balancing predictive accuracy across populations to facilitating the equitable distribution of health in the world.

Elle Lett[1,2,3] ✉ **& William G. La Cava**[1,4]

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Center for Applied Transgender Studies, Chicago, IL, USA. [4]Harvard Medical School, Boston, MA, USA.

✉e-mail: elle.lett@childrens.harvard.edu

Published online: 28 April 2023

References

1. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. *JAMA Intern. Med.* **178**, 1544–1547 (2018).
2. Collins, P. H. & Bilge, S. *Intersectionality* (John Wiley & Sons, 2020).
3. Bailey, Z. D. et al. *Lancet* **389**, 1453–1463 (2017).
4. White Hughto, J. M., Reisner, S. L. & Pachankis, J. E. *Soc. Sci. Med.* **147**, 222–231 (2015).
5. Morris, A. A. et al. *Am. J. Cardiol.* **123**, 291–296 (2019).
6. Johnson, J. D. et al. *Obstet. Gynecol.* **134**, 1155–1162 (2019).
7. Lett, E., Asabor, E. N., Corbin, T. & Boatright, D. *J. Epidemiol. Community Health* **75**, 394–397 (2021).
8. Bor, J., Venkataramani, A. S., Williams, D. R. & Tsai, A. C. *Lancet* **392**, 302–310 (2018).
9. Sewell, A. A. et al. *Ethn. Racial Stud.* **44**, 1089–1114 (2021).
10. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
11. del Barrio, E., Gordaliza, P. & Loubes, J.-M. Preprint at https://doi.org/10.48550/arXiv.2005.13755 (2020).
12. Rodolfa, K. T., Lamba, H. & Ghani, R. *Nat. Mach. Intell.* **3**, 896–904 (2021).
13. Prabhakaran, V. & Martin, D. *Health Hum. Rights* **22**, 71–74 (2020).
14. Sloane, M., Moss, E., Awomolo, O. & Forlano, L. P in *EAAMO '22: Equity and Access in Algorithms, Mechanisms, and Optimization* 1–6 (Association for Computing Machinery, 2022).
15. Siegel, S. D. et al. *Breast Cancer Res.* **24**, 37 (2022).
16. Huyser, K. R., Horse, A. J. Y., Kuhlemeier, A. A. & Huyser, M. R. *Am. J. Public Health* **111**, S208–S214 (2021).

17. Hardeman, R. R., Homan, P. A., Chantarat, T., Davis, B. A. & Brown, T. H. *Health Aff.* **41**, 179–186 (2022).
18. Homan, P., Brown, T. H. & King, B. *J. Health Soc. Behav.* **62**, 350–370 (2021).
19. Segar, M. W. et al. *JAMA Cardiol.* **7**, 844–854 (2022).
20. Mhasawade, V. & Chunara, R. C in *Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society* 784–794 (Association for Computing Machinery, 2021).
21. Bellamy, R. K. et al. *IBM J. Res. Dev.* **63**, 4–1 (2019).
22. Kearns, M., Neel, S., Roth, A. & Wu, Z. S. in *Proc. 35th International Conference on Machine Learning* **80**, 2564–2572 (2018).
23. Hébert-Johnson, U., Kim, M., Reingold, O. & Rothblum, G. in *Proc. 35th International Conference on Machine Learning* **80**, 1939–1948 (2018).
24. Foulds, J. R., Islam, R., Keya, K. N. & Pan, S. in *Proc. 2020 SIAM International Conference on Data Mining* 424–432 (Society for Industrial and Applied Mathematics, 2020).
25. Sullivan, P. S. et al. *Lancet* **397**, 1095–1106 (2021).

Acknowledgements

E.L. thanks L. Bowleg and G. Bauer from the Intersectionality Training Institute and the E^2 Social Epidemiology Lab for their support of this work. The authors thank M. Roberson and K. Baker for their feedback on the manuscript. E.L. was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) grant 5 T32 HD40128-19. W.G.L. was supported by the National Institutes of Health (NIH) National Library of Medicine grant R00-LM012926.

Competing interests

The authors declare no competing financial or non-financial interests.

Additional information

Peer review information Nature Machine Intelligence thanks the anonymous reviewers for their contribution to the peer review of this work.