HHS Public Access

Author manuscript

Nat Mach Intell. Author manuscript; available in PMC 2023 August 18.

Published in final edited form as:

Nat Mach Intell. 2023 May ; 5(5): 476–479. doi:10.1038/s42256-023-00651-3.

Translating Intersectionality to Fair Machine Learning in Health Sciences

Elle Lett, PhD, AM, MBiostat[1,2,3], **William G. La Cava, PhD**[1,4]

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, United States of America

[2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

[3]Center for Applied Transgender Studies, Chicago, Illinois, United States of America

[4]Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Fairness approaches in machine learning should involve more than 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.

INTRODUCTION

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 utility in preventing ML models from becoming an adverse digital health determinant. Populations subject to severe inequities in healthcare access, treatment, and outcomes experience multiple intersecting systems of power and oppression. Further, 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 impact of societal-level discrimination in modeling, and how, and for whom, (un)fairness

Corresponding Author: Elle Lett, PhD, AM, MBiostat, elle.lett@childrens.harvard.edu.

AUTHOR CONTRIBUTIONS

Elle Lett conceptualized the framework and wrote the original draft. William La Cava provided critical feedback and revisions.

COMPETING INTERESTS

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

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 Machine Learning

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 in concert. 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 like racism[3] and transphobia.[4] Additionally, 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," geographic regions in which residents have limited access to affordable and nutritious food (e.g. 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 United States has unique manifestations, particularly in the form of racialized police violence.[7] For our hypothetical task, beyond mortality and injury impacts 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 US. 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 like 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, wherein 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 fairnessaware 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 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 centers 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 allow non-academics to help define the prediction task and oversee the development and implementation pipeline.[13] CBPR approaches must include adequate compensation for labor 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 centers 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 patients 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 native and indigenous populations in the United States, 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] With regard to 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 (e.g. race and gender) and algorithms that satisfy group fairness constraints, imposing (near) equality of some metric across groups that share demographic traits. This approach flattens the multilevel interfaces of power and privilege (e.g. 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 multiple data sources on social inequalities (e.g. 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 conceptualizing 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 United States, 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 multiple 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 centers and even prioritizes the multiply marginalized—individuals who exist at the convergence of intersecting power systems. In 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. Additionally, 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 fail to maintain antiretroviral therapy for HIV in the US, 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 (PPV) 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 PPV rates.[12]

Post-deployment studies are necessary to determine the impact 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 actually 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 where model outputs have consequences that can compound health inequities. We adapt intersectionality to fair ML through its two arms: inquiry–emphasizing how we quantify and correct for algorithmic injustice in models; and 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.

ACKNOWLEDGMENTS

Elle Lett would like to thank Lisa Bowleg and Greta Bauer from the Intersectionality Training Institute and the E^2 Social Epidemiology Lab for their support of this work. The authors would like to thank Drs. Mya Roberson and Kellan Baker for their feedback on the manuscript.

EL was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Grant 5 T32 HD40128-19. WGL was supported by the National Institutes of Health National Library of Medicine grant R00-LM012926.

DATA AVAILABILITY

This piece does not include any original analyses so there are no data to be made available.

REFERENCES

1. Gianfrancesco MA, Tamang S, Yazdany J & Schmajuk G Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern. Med 178, 1544–1547 (2018). [PubMed: 30128552]
2. Collins PH & Bilge S Intersectionality. (John Wiley & Sons, 2020).
3. Bailey ZD et al. Structural racism and health inequities in the USA: evidence and interventions. The Lancet 389, 1453–1463 (2017).
4. White Hughto JM, Reisner SL & Pachankis JE Transgender stigma and health: A critical review of stigma determinants, mechanisms, and interventions. Soc. Sci. Med 147, 222–231 (2015). [PubMed: 26599625]
5. Morris AA et al. Relation of Living in a “Food Desert” to Recurrent Hospitalizations in Patients With Heart Failure. Am. J. Cardiol 123, 291–296 (2019). [PubMed: 30442360]

6. Johnson JD et al. Racial and ethnic inequities in postpartum pain evaluation and management. Obstet. Gynecol 134, 1155–1162 (2019). [PubMed: 31764724]
7. Lett E, Asabor EN, Corbin T & Boatright D Racial inequity in fatal US police shootings, 2015–2020. J Epidemiol Community Health 75, 394–397 (2021).
8. Bor J, Venkataramani AS, Williams DR & Tsai AC Police killings and their spillover effects on the mental health of black Americans: a population-based, quasi-experimental study. The Lancet 392, 302–310 (2018).
9. Sewell AA et al. Illness spillovers of lethal police violence: the significance of gendered marginalization. Ethn. Racial Stud 44, 1089–1114 (2021).
10. Pleiss G, Raghavan M, Wu F, Kleinberg J & Weinberger KQ On fairness and calibration. Adv. Neural Inf. Process. Syst 30, (2017).
11. del Barrio E, Gordaliza P & Loubes J-M Review of mathematical frameworks for fairness in machine learning. ArXiv Prepr. ArXiv200513755 (2020).
12. Rodolfa KT, Lamba H & Ghani R Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nat. Mach. Intell 3, 896–904 (2021).
13. Prabhakaran V & Martin D Participatory Machine Learning Using Community-Based System Dynamics. Health Hum. Rights 22, 71–74 (2020). [PubMed: 33390696]
14. Sloane M, Moss E, Awomolo O & Forlano L Participation Is not a Design Fix for Machine Learning. in Equity and Access in Algorithms, Mechanisms, and Optimization 1–6 (Association for Computing Machinery, 2022). doi:10.1145/3551624.3555285.
15. Siegel SD et al. Racial disparities in triple negative breast cancer: toward a causal architecture approach. Breast Cancer Res. 24, 37 (2022). [PubMed: 35650633]
16. Huyser KR, Horse AJY, Kuhlemeier AA & Huyser MR COVID-19 Pandemic and Indigenous Representation in Public Health Data. Am. J. Public Health 111, S208–S214 (2021). [PubMed: 34709868]
17. Hardeman RR, Homan PA, Chantarat T, Davis BA & Brown TH Improving The Measurement Of Structural Racism To Achieve Antiracist Health Policy. Health Aff. (Millwood) 41, 179–186 (2022). [PubMed: 35130062]
18. Homan P, Brown TH & King B Structural intersectionality as a new direction for health disparities research. J. Health Soc. Behav 62, 350–370 (2021). [PubMed: 34355603]
19. Segar MW et al. Machine Learning-Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure. JAMA Cardiol. 7, 844–854 (2022). [PubMed: 35793094]
20. Mhasawade V & Chunara R Causal Multi-level Fairness. in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society 784–794 (Association for Computing Machinery, 2021). doi:10.1145/3461702.3462587.
21. Bellamy RK et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev 63, 4–1 (2019).
22. Kearns M, Neel S, Roth A & Wu ZS Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. in 2564–2572 (PMLR, 2018).
23. Hébert-Johnson U, Kim M, Reingold O & Rothblum G Multicalibration: Calibration for the (computationally-identifiable) masses. in 1939–1948 (PMLR, 2018).
24. Foulds JR, Islam R, Keya KN & Pan S Bayesian Modeling of Intersectional Fairness: The Variance of Bias*. in 424–432 (SIAM, 2020).
25. Sullivan PS et al. Epidemiology of HIV in the USA: epidemic burden, inequities, contexts, and responses. The Lancet 397, 1095–1106 (2021).

Table 1:

Intersectionality Core Ideas for Machine Learning Researchers

Intersectionality Core Idea	Implications for Machine Learning and Fairness
Social Inequalities	*Data Generating Mechanism*: Training data exhibits health inequities due to social inequalities (e.g. wealth, education, 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 (e.g. 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 United States, may not transport to another, like 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 center individuals from marginalized backgrounds throughout the ML pipeline *Impact*: Use post-deployment studies to determine if the benefits of ML tools are experienced equitably across groups and if corresponding health inequities are being decreased