Bias in Clinical Risk Prediction Models: Challenges in Application to Observational Health Data

Yoonyoung Park, Moninder Singh, Issa Sylla, Elaine Xiao, Jianying Hu, Amar Das
Center for Computational Health, IBM Research, USA
yoonyoung.park@ibm.com

Abstract

With the increasing use of machine learning and Artificial Intelligence (AI) in healthcare, ensuring the fairness of algorithms is paramount to prevent health disparities and inequities from being reproduced in algorithm-guided medical and policy decisions. In this work we investigate algorithmic bias in clinical prediction models and discuss challenges in analyzing bias in observational health data. We show that potential disparities in treatment opportunity exist between races in the data for patients with opioid use disorder, and that the direction of bias favoring one race over the other depends on the choice of outcome label or fairness metric. We further demonstrate how debiasing algorithms can effectively mitigate the apparent bias in most experimental settings. This study exemplifies the importance of thorough bias assessment in prediction tasks based on healthcare data.

Introduction

Now more than ever before, machine learning algorithms are used to guide decision making across various domains. Simultaneously, researchers and decision makers are becoming more alert to the presence of algorithmic biases in such models (Chen, Joshi, and Ghassemi 2020; Rajkomar et al. 2018; Gianfrancesco et al. 2018). Substantial work has been done in this field to define fairness metrics and devise approaches to detect, evaluate and mitigate bias (Makhlouf, Zhioua, and Palamidessi 2020; Friedler et al. 2018). Particularly, in the healthcare space, a recent landmark study showed that an algorithm widely used in the United States (US) to allocate care management resources is racially biased (Obermeyer et al. 2019). A similar type of bias was observed in a nationally representative US population dataset (Singh and Ramamurthy 2019), raising concerns for fairness in data- and model-driven decisions in healthcare.

While previous works have explored bias from population health management point of view, no one has investigated the problem for prediction tasks involving a specific disease cohort and clinical outcome. These tasks give rise to challenges including difficulty in bias ascertainment, effects of treatment options, and use of surrogate clinical outcomes. In this paper, we conduct a set of analyses using observational health data to assess fairness of clinical prediction algorithms and discuss our challenges. We consider patients diagnosed with opioid use disorders (OUD), a historically challenging population with highly complex physical and mental healthcare needs. Evidence-based treatment called medication-assisted treatment (MAT) exists for OUD, but the utilization is reported to be uneven across race groups (Lagisetty et al. 2019). We can visualize a setting where limited availability prevents prescribing MAT to all OUD patients, and a prediction model is developed to identify patients who are at greater risk of experiencing adverse outcomes. Fair outcome prediction for these patients is important from both clinical and public health point of views considering the history of disparity and inequity resulting in poor outcomes and distrust towards health systems.

The main contributions of this paper are (1) comprehensive bias analysis of a real-world clinical scenario, including application of debiasing approaches and (2) discussions around analytic challenges. In addition to reaffirming part of the findings in (Obermeyer et al. 2019), we demonstrate a transparent approach to bias analysis that will help gain trust from clinicians and practitioners and encourage more widespread use of similar efforts.

Data and Study Setting

We used longitudinal patient-level claim records from the IBM® MarketScan® Medicaid Databases (2013-2017). Patients were either white or black, continuously enrolled in Medicaid during one year prior to and after the initial diagnosis date (i.e. index date); patients did not receive any OUD treatment before the index date and did not have cancer or hospice for which use of opioid is justified. Selected characteristics of the patients are presented in Table 1.

We follow the conventional terminology and assign a classification of 'high risk individual' as the *favorable label*, assuming that this will lead to treatment opportunity. The *protected attribute* is race, and *privileged value* is white as opposed to black. Using three classifiers (logistic regression, random forest, extreme gradient boosting trees (XGB)) we predict high risk patients for one of the outcomes introduced below. Patient features generated from the baseline data in-

Table 1: Characteristics of Medicaid patients with opioid use disorder, 2013-2017

	Overall (n=42,525)	**White (n=32,518)**	**Black (n=10,007)**
Pre-index (Mean (SD) or N(%))			
N (%)	100.0	76.5	23.5
Age	38.1 (12.8)	37.1 (12.4)	41.7 (13.4)
Female gender (N(%))	27,946 (65.7)	21,659 (66.6)	6,287 (62.8)
Comorbidity index	0.7 (1.3)	0.6 (1.1)	1.0 (1.7)
N psychiatric diagnosis	1.3 (1.3)	1.3 (1.3)	1.1 (1.3)
Had >0 outpatient ER visits ((N(%))	30,491 (71.7)	23,134 (71.1)	7,357 (73.5)
Had >0 emergency psychiatric admission (N(%))	2,059 (4.8)	1,562 (4.8)	497 (5.0)
Post-index (Mean (SD) or %)			
Total cost	$17.1K (46.8K)	$14.9K (41.4K)	$24.3K (60.5K)
Outpatient ER visit cost	$0.7K (2.5K)	$0.6K (1.9K)	$1.1K (3.7K)
N psychiatric diagnosis	1.3 (1.3)	1.3 (1.3)	1.1 (1.3)
Emergency psych admission cost	$1.1K (8.2K)	$1.1K (7.7K)	$1.3K (9.6)
MAT utilization rate (N(%))	8,700 (20.5)	7,837 (24.1)	863 (8.6)

N: number; ER: emergency room; MAT: medication assisted treatment

clude age, gender, race, medication use, comorbid condition diagnoses, mental health procedures, and utilization levels such as number of emergency room (ER) visits in baseline period. Data were split into train:validation:test sets (5:3:2) and models were mostly trained with default parameter values. While all three classifiers produced qualitatively consistent results, we focus our interpretation on the results from XGB models in the interest of space and clarity.

Bias Analysis and Challenges

Bias in our study context is discrepancy in health related measures between groups of people defined by the sensitive attribute, race, that cannot be justified. One 'fair' scenario with OUD patients would involve MAT being prescribed to patients who are most likely to benefit from treatment by avoiding adverse outcomes such as emergency room visits, overdose events, or hospitalizations, regardless of their race. We inspect the data and the prediction outcomes for bias before attempting to mitigate bias.

Bias assessment in data

The most fundamental step is determining whether there exists bias in the underlying data. If the training data is unbiased, the model is less likely to produce biased predictions. Because numerical discrepancies can be attributed to factors other than bias, an in-depth knowledge of the primary data generating process is critical. There may be 'admissible' differences between subgroups that can be explained by reasons such as comorbid conditions. On the other hand, the seemingly justifiable difference may actually arise from more deeply rooted disparities that manifest in differential access to care and diagnosis. One crucial step towards trustworthy bias analysis is explicitly stating all verifiable and non-verifiable assumptions.

If there is no bias in resource allocation, we would expect to see no differences by race in the data with respect to the characteristics of patients receiving MAT. There is no standard approach to assess data for presence of bias. We examine this by modeling the probability of receiving MAT, adjusted for clinical risk factors and baseline service utilization. Biological race is a non-modifiable factor, but as a social construct captures various determinants of health associated with race. As discussed in (VanderWeele and Robinson 2014), we wish to interpret the coefficient of race as the degree of inequality that would remain if risk factor distributions of the black population were set equal to that of the white population. We are assuming that 1) the numerical discrepancy we observe between white and black patients is indeed unjustifiable bias after adjusting for background determinants of health, and that 2) race's effect on the probability of receiving treatment is mediated through variations in clinical factors, health service utilization and related behaviors, and a composite effect of disparity.

A generalized linear model predicting receipt of MAT adjusted for baseline and mediating factors had the odds ratio (OR) of 3.18 (95% confidence interval 2.95-3.43) for race variable, suggesting existence of inequality in treatment provided that our assumptions hold. Since we adjust for many background variables, we expect that at least part of the results is attributed to disparity. In a model adjusting for age, gender, race, and comorbidity, use of MAT was associated with reduced risk of ER visit and reduced cost (total, ER, inpatient cost). Lower adjusted probability of any ER visits (OR 0.82, 0.78-0.87) and adjusted mean cost (total, ER cost) among white patients suggest treatment inequality can lead to more undesirable events (e.g. ER visit) and higher costs in black patients. MAT use was not associated with overdose risk, and the probability of overdose or psychiatric admission was slightly higher among white patients.

Bias in predictions

Determining MAT allocation by simply following historical distribution of MAT treatment is problematic because current practice appears to be associated with racial disparity as we have shown above. The alternative is to identify at-risk patients using predictive algorithms. Unlike those explicitly targeting cost, clinical risk prediction models often need to opt for surrogate outcomes due to limited data. The

Table 2: Predicted high risk subcohorts (XGB) without debiasing

	Total Cost		**ER Visit Cost**		**Psych IP Cost**		**N Psych Dx**	
(% or Mean*)	White	Black	White	Black	White	Black	White	Black
Age	48.0	46.4	42.2	43.5	33.5	37.6	34.7	39.4
Female gender (%)	59.5	57.8	68.2	65.1	62.9	50.5	65.7	48.8
Pre-index comorbidity index	2.6	3.4	1.7	2.5	0.8	1.3	1.0	1.6
Total cost ($)	49.4K	77.2K	45.0K	71.8K	33.6K	38.2K	27.9K	39.7K
Outpatient ER visit cost ($)	2.4K	3.5K	3.1K	4.4K	1.4K	2.7K	1.2K	2.8K
Emer psych admission cost ($)	5.0K	4.7K	4.3K	4.9K	6.2K	9.0K	6.2K	9.6K
N psychiatric diagnosis	2.2	1.6	2.4	1.8	2.8	2.6	3.1	3.1
MAT utilization (%)	11.8	7.0	16.3	8.0	14.5	6.4	15.5	3.5
Overdose event (%)	4.5	2.5	4.7	2.2	4.5	3.5	5.0	3.5

*Leaving out N and standard deviations for space

All post-index measures except age, gender, and comorbidity index

challenge here is choosing the most relevant yet least biased proxy in high dimensional healthcare data. Measurement error and human bias affect the label data which is often treated as the ground-truth in model building. Bias is defined with respect to a specific outcome, therefore evaluating bias is sensitive to outcome label.

A number of surrogate labels exist for OUD patients. The shared goal is to find high risk patients for MAT who will benefit from treatment by lowering the risk. We compare four models trained with different labels to classify high risk patients with varying likelihoods of reflecting biases in the underlying data. These labels were chosen for clinical and public health relevance based on prior studies.

- Being in the top decile of total healthcare cost (Total Cost): while total cost is often used to identify potentially ‘expensive’ patients for subsequent interventions, it is correlated with access to care or utilization pattern, differences that may arise from underlying disparities.
- Being in the top decile of outpatient ER visit cost (ER Visit Cost): we hypothesized that the occurrence of health related issues leading to ER utilization would be less associated with bias arising from access to or quality of care, but more with the severity of physical and mental illness.
- Being in the top decile of emergency psychiatric admission cost (Psych IP Cost): we expected that occurrence of acute exacerbation of mental health issues resulting in emergency admissions would be less associated with race related bias compared to all-cause healthcare needs.
- Being in the top decile of the number of psychiatric diagnoses (N Psych Dx): the count of chronic conditions has been used as a measure of underlying health status as an alternative to cost. We used the total number of psychiatric diagnoses as a measure of psychiatric comorbidity.

We compared the characteristics of high-risk subcohorts to see how they differ across the four label choices (Table 2). As expected, the subcohorts in general had higher disease burden and utilized more health services. Black patients were still much less likely to receive MAT. Using total cost or ER cost label classified as high risk older and sicker patients compared to using the other two labels, with black patients having higher disease burden and much higher cost. For example, using total cost as the label, the average total cost ($) was 49.4K and 77.2K and average comorbidity score was 2.6 and 3.4 for high risk white and black patients, respectively. This is similar to what was observed in (Obermeyer et al. 2019). Using the emergency psychiatric admission cost label, the average total costs were much more comparable at 33.6K and 38.2K, and the average comorbidity scores were 0.8 and 1.3. However, discrepancy in the average emergency psychiatric admission cost was much lower when using total cost (5.0K and 4.7K) than when using psychiatric admission cost label (6.2K and 9.0K). This observation suggests that the factors affecting cost, such as clinical characteristics or patient preferences, vary for total cost and psychiatric service-specific cost. With the same purpose of identifying the most at-risk patients, two labels will ’favor’ white patients by classifying more white patients as high risk; the other two labels ’favor’ black patients.

One remaining question pertains to the seemingly contradictory message on fairness across the target labels. Depending on the choice of label, actions following the prediction outcomes will have the opposite impact on patients. We propose that in this case, one should choose the label directly associated with the specific issue at hand - substance abuse and mental health related care - rather than all-encompassing labels which may be associated with factors that are unrelated or related in opposite ways. If we choose a mental health specific outcome, we conclude that using that prediction models will unfairly favor white patients.

Bias measurement

The choice of fairness metric represents our belief on ’what is fair’. The existing metrics can conflict with one another and cannot be fulfilled simultaneously (Kleinberg, Mullainathan, and Raghavan 2017). Two commonly used metrics are disparate impact (DI), defined as the difference in proportions of predicted positive label, and equal opportunity difference (EOD), the difference in true positive rates (TPR) between privileged and unprivileged groups.

We used DI and EOD in our analysis to measure the level of fairness (Figure 1). Using DI as a metric assumes that equal proportions of positive labels produces fairness. This

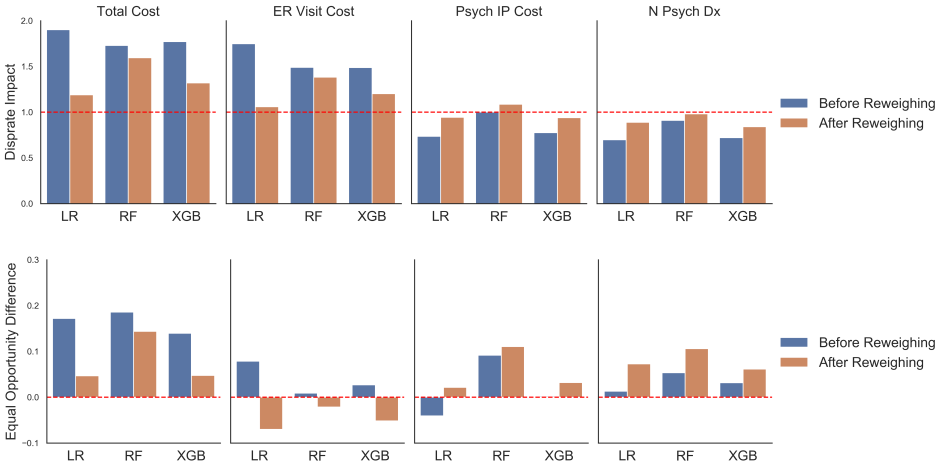

Figure 1: Bias metrics before and after reweighing

may not always be the case as the accuracy may differ between the two groups. But if we believe some outcome measures are associated with historical biases (e.g. a white person is more likely to be diagnosed for the same disease), achieving demographic parity can still be a way to improve fairness. Use of EOD assumes that we care about true positives only and not false positives. This approach can be justified in the study setting because the decision maker will care more about finding all true positives and less about having false positives, which will likely not cause harm to patients. We saw that while the magnitudes of bias quantified by DI had consistent results, the magnitudes measured by EOD was more variable. If we take EOD as the metric, we might conclude that there is not enough evidence of bias in the models predicting ER visit cost, whereas using DI as the metric will lead to the opposite conclusion.

Bias mitigation

After quantifying bias, we applied debiasing algorithms to mitigate the bias in prediction outcomes. It does not, however, 'fix' or remove bias in the data. There are three classes of bias mitigation algorithms depending on where the bias occurs in a machine learning pipeline (d'Alessandro, O'Neil, and LaGatta 2017). A recent review paper illustrated how sensitive these debiasing algorithms are to fluctuations in input data (Friedler et al. 2018). Such sensitivity will discourage clinicians from trusting and adopting the new ways of approaching model biases.

We used Reweighing, a pre-processing method that modifies the training data by generating weights for (group, label) combinations. We also tested Prejudice Remover, an in-processing method that uses a discrimination-aware regularization term in the objective function to remove bias. We focus our presentation on the results from Reweighing. Both were implemented using AI Fairness 360 (Bellamy et al. 2019), an open source python toolkit for fairness research.

Implementing the reweighing algorithm successfully reduced DI values in most experimental settings (Figure1). Debiasing through reweighing did not have negative impact on the balanced accuracy. EOD values fluctuated and sometimes became worse than before debiasing. However, the small magnitude of EOD values has a less practical implication than does the magnitude of DI values, which can change the proportion of high risk patients. For the total cost label with the largest difference in EOD (based on XGB), the decrease in value from 0.14 to 0.05 means fewer black (n=8, 3% of the true positives) and more white (n=37, 7% of the true positives) patients who actually have positive outcomes will be classified as positive (true positives). Prejudice Remover reduced DI for all but one target, number of psychiatric diagnosis, whose values were similar before and after debiasing (data not shown).

Discussions

Healthcare data is rife with known and unknown biases, and varying sources of bias make it difficult to have one-size-fits-all approaches for understanding algorithmic bias. As illustrated in our work, the choice of target label significantly affects the degree and magnitude of bias present. Considering the well-known incompatibility of fairness metrics and the sensitivity of interpretation to the metrics, researchers caution against purely technical approaches to debiasing algorithms in clinical settings (McCradden et al. 2020). The first step toward successful bias mitigation is to have a thorough understanding of the study population, health service utilization patterns, data collection mechanisms, and quality of surrogacy of target measures. We did not consider any deep models, but the three models we examined are more frequently used in medicine due to the ease of implementation and better explainability which are important for clinical adaptation. Additionally, we show that even with careful selection of target measures, the lack of unbiased outcome surrogate or gold standards to confirm unfairness makes it very difficult to completely avoid bias in machine learning models; a certain degree of bias was present across all experiments. This highlights the need to rigorously evaluate bias and proactively deploy debiasing measures when developing risk models.

References

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J. T.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63:4:1–4:15.

Chen, I. Y.-P.; Joshi, S.; and Ghassemi, M. 2020. Treating health disparities with artificial intelligence. *Nature Medicine* 26:16–17.

d'Alessandro, B.; O'Neil, C.; and LaGatta, T. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5 2:120–134.

Friedler, S. A.; Scheidegger, C. E.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2018. A comparative study of fairness-enhancing interventions in machine learning. *FAT* '19*.

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine* 178:1544–1547.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent trade-offs in the fair determination of risk scores. *ArXiv* abs/1609.05807.

Lagisetty, P.; Ross, R.; Bohnert, A. S. B.; Clay, M. A.; and Maust, D. T. 2019. Buprenorphine treatment divide by race/ethnicity and payment. *JAMA psychiatry*.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. On the applicability of ml fairness notions. *ArXiv* abs/2006.16745.

McCradden, M.; Joshi, S.; Mazwi, M.; and Anderson, J. A. 2020. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digital Health*.

Obermeyer, Z.; Powers, B. W.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366:447–453.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G. S.; and Chin, M. H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 169:866–872.

Singh, M., and Ramamurthy, K. N. 2019. Understanding racial bias in health using the medical expenditure panel survey data. *ArXiv* abs/1911.01509.

VanderWeele, T., and Robinson, W. 2014. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 25 4:473–84.