IBM **Research**

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

IBM Research, USA

Introduction

Prior literature showed racial bias in a commercial algorithm used to allocate patient management resources in the US (Obermeyer et al. 2019) and in a similar machine-learning (ML) application study (Singh and Ramamurthy 2019). In this work we investigate algorithmic bias in clinical prediction models for patients with opioid use disorder (OUD) and discuss challenges in analyzing bias in observational health data.

Setup

Setting. Utilization of treatment for OUD patients called medication-assisted treatment (MAT) is reported to be uneven across race groups. We hypothesize 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 for experiencing the outcome (see below)

Data. We used longitudinal patient-level claim records from the IBM® MarketScan® Medicaid Databases (2013-2017)t to identify patients with OUD.

Experiments. We compared 3 classifiers and 2 bias mitigation methods for predicting 4 different binary outcomes (labels)

1. Models: logistic regression, random forest, gradient boosted trees (XGB; extreme gradient boosting)
2. Mitigation methods: reweighing, Prejudice Remover (logistic regression only)
3. Prediction labels (all binary, predict top decile)
   - Total healthcare cost (Total Cost)
   - Emergency room visit cost (ER Visit Cost)
   - Emergency psychiatric admission cost (Psych IP Cost)
   - Psychiatric comorbidity (N Psych Dx)

	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

Table. Predicted high risk subcohorts without debiasing (XGB)

Bias in Data

- Context dependent
- Varying definitions of bias or disparity
- Knowledge of data generating process

Bias in Prediction Outcomes

- Clinical prediction often needs surrogate labels
- Label choice crucial for evaluating algorithmic fairness - each have varying likelihood of reflecting biases in the underlying data
- Clinical and public health relevance

Bias Measurement

- Choice of fairness metric represents our belief on 'what is fair'
- Metrics are not always compatible with each other

Bias Mitigation

- Different classes of mitigation actions available
- Many of the existing debiasing methods are sensitive to fluctuations in the input data (Friedler et al. 2018)

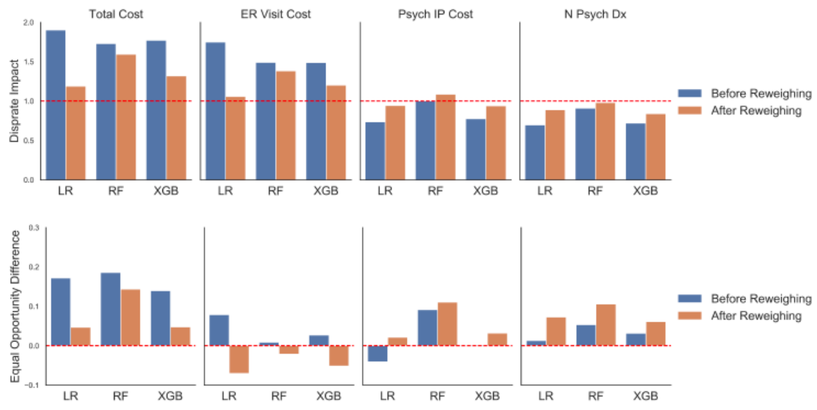

Figure. Bias metrics before and after debiasing by reweighing

A generalized linear model for receipt of MAT had the odds ratio (OR) of 3.18 (95% CI 2.95-3.43) for race, adjusting for confounding/mediating factors. We are assuming the increased OR represents unjustifiable bias or disparity (VanderWeele and Robinson 2014).

Total/ER cost label classified as high risk older and sicker patients compared to the other two labels (Table) → With the same purpose of identifying the most at-risk patients, two labels will 'favor' white patients; the other two 'favor' black patients. Depending on the label, follow-up actions will have the opposite impact on patients.

Disparate impact (DI): statistical parity may still be an acceptable goal in the presence of historic disparity
Equal opportunity difference (EOD): when the treatment is beneficial, finding true positives matter more than avoiding false positives. It would not be true if the treatment is not always beneficial or potentially harmful.

Reweighing reduce DI values in most experimental settings (Figure1). Debiasing through reweighing did not have negative impact on the balanced accuracy.
The small magnitude of EOD values has a less practical implication than does the magnitude of DI values. Prejudice Remover reduced DI for all but one target.

Discussion

We show that potential disparities in treatment opportunity exist between races in the data for patients with OUD, 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 methods can be used mitigate the apparent bias. 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; this highlights the need to rigorously evaluate bias and proactively deploy debiasing measures when developing risk models.

Contact: yoonyoung.park@ibm.com