Evaluating Algorithmic Fairness in Predicting Temporal Alcohol Lapse Risks with Machine Learning Models

Jiachen Yu, Kendra Wyant, Sarah Sant'Ana, Gaylen Fronk, John Curtin. (Department of Psychology)

Addiction Research Center
UNIVERSITY OF WISCONSIN – MADISON

Overview and Goals

Alcohol Lapse Prediction

- Developed an XGBoost machine learning model to predict alcohol lapses in the next 24 hours using ecological momentary assessments (EMA; Wyant et al., 2023)
- Model has exceptional performance when predicting lapses for new individuals (mean auROC = .90)
- Locally important features can identify the factors that contribute to lapse risk for any specific person and moment in time
- A "smart" digital therapeutic (smart DTx) could use algorithms based on this model to monitor lapse risk and recommend personalized, optimal interventions and other supports for momentary risks

Algorithmic Bias

- Less privileged, marginalized groups display mental health treatment disparities due to barriers related to affordability, accessibility, availability, and acceptability of mental healthcare
- Smart DTx can partially address these barriers by providing 24/7/365, affordable, personalized support
- However, if embedded algorithms perform relatively worse for less privileged groups, their use may exacerbate rather than reduce treatment disparities

GOAL: Evaluate potential algorithmic bias across historically privileged and unprivileged groups

Methods

Participants

- N = 151
- Early remission from Alcohol Use Disorder
- Endorsed abstinence goal

Procedure

- Personal sensing via smartphone for up to 3 months
- 4x daily (craving, affect, efficacy, risky situations, stressful events, pleasant events)
- Self-reported lapses back to alcohol use

Machine Learning Model

- XGBoost classification model
- Features based on previous EMA
- Provides hour-by-hour probability for future (next 24 hour) lapse

Sensitive Attributes

- **Race/Ethnicity**: Non-hispanic White vs. People of Color
- **Sex**: Male vs. Female
- **income**: above 50% of median personal income in Madison ($15k) vs. below
- **age**: 55 or younger vs. above 55

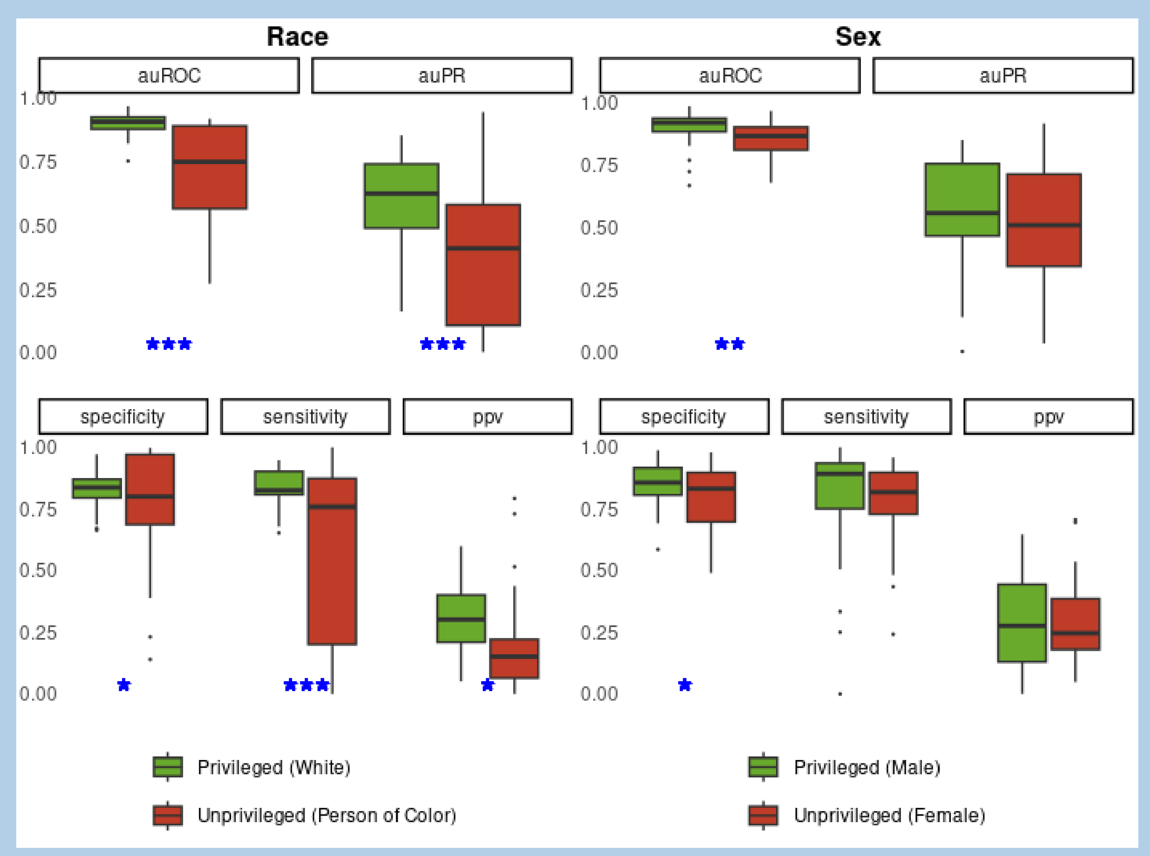
Performance Metrics

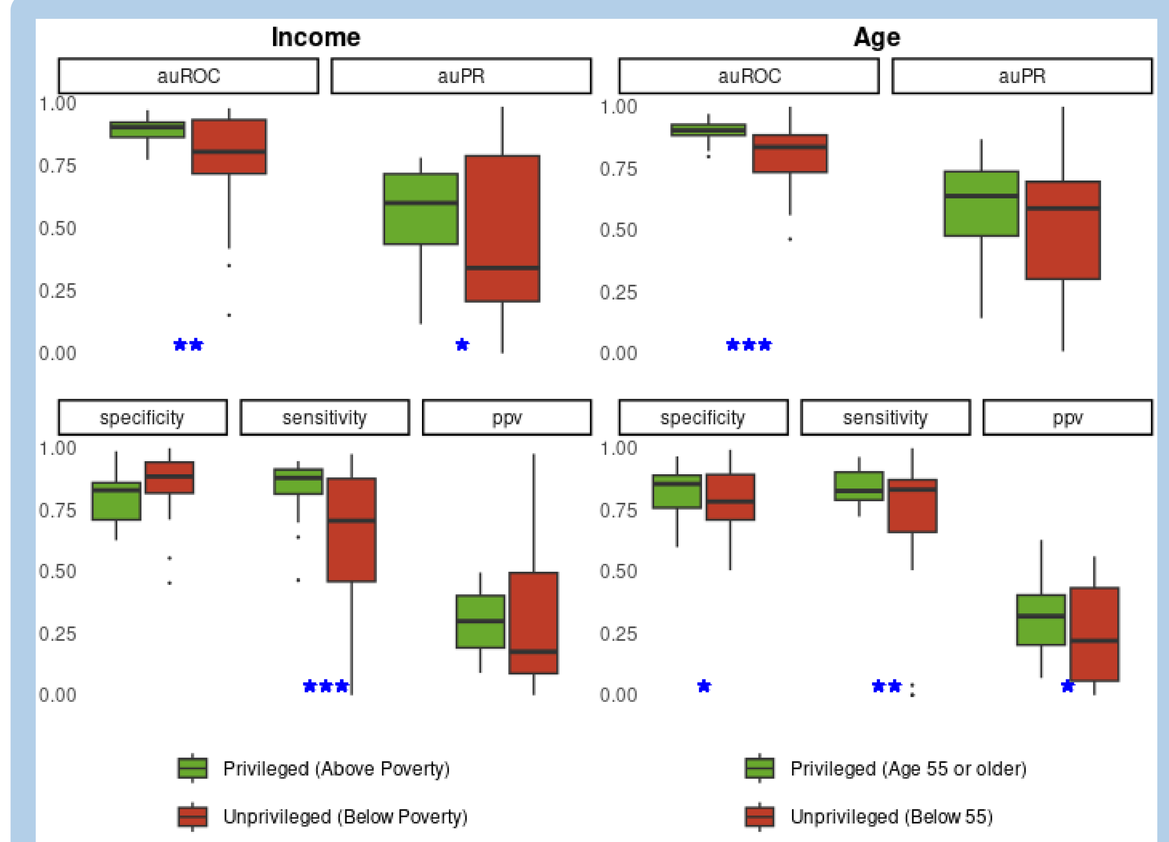
- auROC: true positive rate (sensitivity) against true negative rate (specificity) across thresholds
- auPR: sensitivity against positive predictive value (ppv) across thresholds

Statistical Analysis

- 30 held-out performance estimates (from nested k-fold cross validation) for each metric
- Posterior probabilities of group differences estimated using Bayesian generalized mixed effect models

Results





- Models perform systematically better for White individuals than People of Color.
- auROCs are higher for males than females, probably primarily driven by specificity.
- Both auROCs and auPRs are higher in people with higher income, probably due to sensitivity.
- auROCs, sensitivity, specificity and ppv are significantly higher for younger population.

Conclusions

- Substantially poorer model performance for people of color
- Modestly poorer model performance for groups defined by sex, income, and age
- Representation in training data is clearly important
- Bias may also result from selection of features using domain expertise based on decades of research focused on predominately white men

Next Steps

- Evaluate methods to reduce algorithmic bias
  - More representative training data (NIDA project)
  - Resampling to increase representation
  - Modified cost functions to differential penalize errors based on privilege
  - Control for privilege when predicting
  - More representative features
- Expand evaluation of privilege
  - Rural group, Education
  - AUD Severity
  - Intersectional analyses
  - Bias in treatment/support recommendations