Assessing Demographic Variability in Machine Learning Model Performance to Predict Alcohol Lapses

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

Abstract

Preliminary analyses of machine learning models that leverage ecological momentary assessments (EMA; 4x daily for three months) to predict future alcohol lapses among individuals in recovery for alcohol use disorder have demonstrated excellent performance overall, with areas under the receiver operating curves (auROCs) > .90 (Wyant et al, 2023). These models were trained and evaluated with 151 participants who were diverse with respect to sex (51% male), age (range=21-72), and income (range=0-200,000), but not race/ethnicity (87% White; 97% Non-Hispanic). Before implementing these models clinically, careful analyses of model performance in sub-groups that have been historically marginalized/under-served are necessary to avoid potentially exacerbating existing mental health disparities. The current study evaluated demographic subgroup heterogeneity in these models that predict future lapses with high temporal precision ranging from the next week to the next hour. Subgroup analyses were conducted using different performance metrics from 3 repeats of 10-fold nested, grouped, cross validation. We used Bayesian generalized mixed effect models to estimate and compare posterior probability distributions and 95% Bayesian confidence intervals (CIs) for auROCs, auPRs, specificities, sensitivities and ppvs of the models by race/ethnicity, sex, income, and age. Substantially poorer model performance was observed for participants of color, who were underrepresented in the training data. Clinically important reductions in model performance were also observed for women, individuals with lower income and people aged 55 and higher, despite adequate sample diversity on these characteristics. Potential implications of these algorithmic biases on mental health disparities and possible solutions to mitigate these problems in future real-world applications are discussed.

Aims

Predicting Alcohol Lapse via EMA

- Alcohol relapse is a chronic disease with a temporal, dynamic process and therefore necessitates constant risk monitoring
- Build machine learning models from EMA (proximal factors) to predict alcohol lapses
- Provides insight of *who* will lapse and *when*

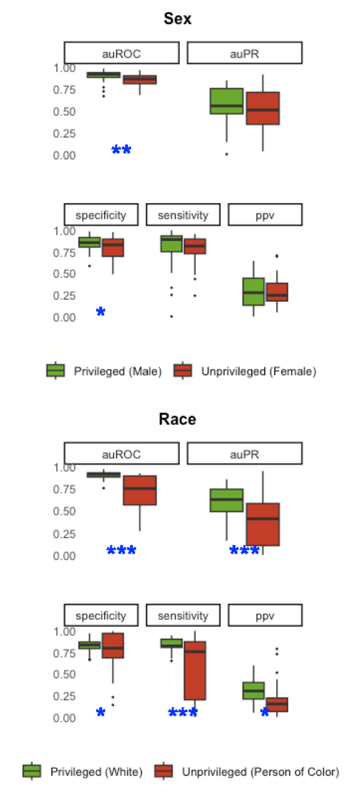
Algorithmic Bias

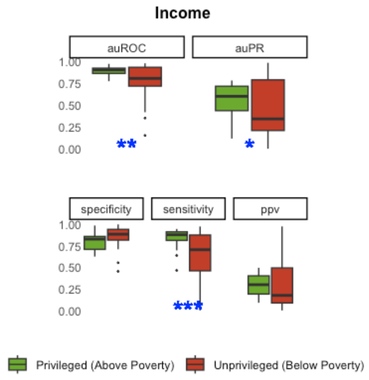
- Marginalized group face barriers accessing mental health resources due to provider availability, financial difficulties, stigma, etc.
- DTx has the advantage of potentially minimizing health inequity due to its higher accessibility. Yet differential performance across subgroups counter this goal
- The goal of this study is to compare algorithm performance between historically privileged and unprivileged groups.

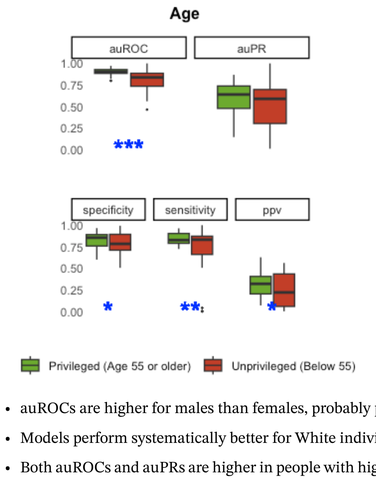
Methods

- Participants and Procedure
  - N = 151, in early recovery of AUD
  - Answer surveys up to 4x daily (craving, risky situations, stressful events, pleasant events)
- Algorithms
  - Uses XGBoost to do model training and leverages grouped, nested cross-validation to select the best hyperparameters
  - A mean of .90 auROC across the 30 held-out folds
- Sensitive Attributes
  - sex: Female vs. Male
  - race: White vs. People of Color
  - income: 50% of median personal income in Madison ($15k)
  - age: 55 or older
- Performance Metrics
  - auROC, auPRC, PPV, sensitivity, specificity
- Statistical Analysis
  - Bayesian generalized mixed effect models

Results







- auROCs are higher for males than females, probably primarily driven by specificity.
- Models perform systematically better for White individuals than People of Color.
- Both auROCs and auPRs are higher in people with higher income, probably due to significant differences in sensitivity.
- auROCs, sensitivity, and specificity are significantly higher for younger population. auPRs are not significantly different between the two groups, but ppv is higher in younger people.

Conclusions

- Substantially poorer model performance was discovered in people of color, while some reductions in model performance in certain performance metrics were observed in females, lower income groups, and older population.
- Possible explanations might include that racial minority is underrepresented in our dataset. The poorer performance might also stem from the fact that we selected predictors based on findings from decades of research that primarily recruited privileged samples.

Implications and Future Directions

- To reduce model bias, machine learning studies might need to recruit a more diverse sample and carefully select predictors generalizable to a broader population.
- Algorithmic bias in other minority groups such as people from more rural areas or with different education backgrounds should be examined.