Submitter Name: Alexander S. Hatoum Submitted Email: ashatoum@wustl.edu

Machine learning Predictors of Substance Use disorder using Genetics Need Control for Ancestry

Alexander S. Hatoum[1,2], Frank R. Wendt[3], Marco Galimberti[3], Renato Polimanti[3,4], Benjamin Neale[5,6], Henry R. Kranzler[7,8], Joel Gelernter[3,4,9,10], Howard J. Edenberg[11,12], Arpana Agrawal[13]

[1]Washington University in St. Louis, Department of Psychological & Brain Sciences
[2]Washington University School of Medicine, The AI for Health Institute
[3]Department of Psychiatry, Division of Human Genetics, Yale School of Medicine, New Haven, CT;
[4]Veterans Affairs Connecticut Healthcare System, West Haven, CT;
[5]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA;
[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA;
[7]Center for Studies of Addiction, Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA;
[8]VISN 4 MIRECC, Crescenz VAMC, Philadelphia, PA;
[9]Department of Genetics, Yale School of Medicine, New Haven, CT;
[10]Department of Neuroscience, Yale School of Medicine, New Haven, CT;
[11]Department of Medical and Molecular Genetics, Indiana University School of Medicine;
[12]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine;
[13]Washington University in St. Louis, School of Medicine, Department of Psychiatry

Machine learning (ML) models are beginning to proliferate in psychiatry, however machine learning models in psychiatric genetics have not always accounted for ancestry. Using three examples, an empirical example based on a proposed genetic test for OUD, an empirical example using genome-wide significant hits for tobacco smoking, and simulation, we show that genetic prediction using ML generates algorithmic bias when ancestry is uncontrolled for. We utilize five ML algorithms trained with 16 brain reward-derived "candidate" SNPs proposed for commercial use and examine their ability to predict OUD vs. ancestry in an out-of-sample test set (N = 1000, stratified into equal groups of n = 250 cases and controls each of European and African ancestry). We contrast findings with 11 genome-wide significant variants for tobacco smoking. We rerun analyses with 8 random sets of allele-frequency matched SNPs. To document generalizability, we generate and test a random phenotype. None of the 5 ML algorithms predict OUD or tobacco dependence better than chance when ancestry was balanced but were confounded with ancestry in an out-of-sample test. In addition, the algorithms preferentially predicted admixed subpopulations. Random sets of variants matched to the candidate SNPs by allele frequency produced similar bias. Finally, random SNPs predicting a random simulated phenotype show that the bias attributable to ancestral confounding could impact any ML-based genetic prediction. Ancestry (and population stratification) are enduring confounds in genetics, generating algorithmic bias in machine learning models. Future approaches will need to control ancestral decent to appropriately predict risk for substance use disorders.

This work has already been published and most of this abstract is available here:
Hatoum, A. S., Wendt, F. R., Galimberti, M., Polimanti, R., Neale, B., Kranzler, H. R., ... & Agrawal, A. (2021). Ancestry may confound genetic machine learning: Candidate-gene prediction of opioid use disorder as an example. *Drug and alcohol dependence*, *229*, 109115. https://doi.org/10.1016/j.drugalcdep.2021.109115