Introductory Paragraph Length: 149

Letter Length: 1431

Genetic Data Can Lead to Medical Discrimination: Cautionary tale of Opioid Use Disorder

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

[1]Washington University in St. Louis, School of Medicine, Department of Psychiatry, USA
[2]Department of Psychiatry, Division of Human Genetics, Yale School of Medicine, New Haven, CT, USA
[3]Veterans Affairs Connecticut Healthcare System, West Haven, CT, USA
[4]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Center for Studies of Addiction, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[7]VISN 4 MIRECC, Crescenz VAMC, Philadelphia, PA, USA
[8]Department of Genetics, Yale School of Medicine, New Haven, CT, USA
[9]Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA
[10]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
[11]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA
*These authors contributed equally to the work

Key Words: Opioid Use Disorder, Machine Learning, Single-Nucleotide Polymorphism, Candidate Genes, Ancestry, Algorithmic Bias

Using genetics to predict the likelihood of future psychiatric disorders such as Opioid Use Disorder (OUD) poses scientific and ethical challenges. This report illustrates flaws in current machine learning (ML) approaches to such predictions using, as an example, a proposed genetic test for OUD derived from 16 candidate gene variants. In an independent sample of OUD cases and controls of European and African descent, results from five ML algorithms trained with purported "reward-system" candidate variants demonstrate that ML methods predict genomic ancestry rather than OUD. Further, sets of variants matched to the candidate SNPs by allele frequency produced similarly flawed predictions, questioning the plausibility of the selected candidate variants. We conclude that the genetic prediction of OUD (and by extension other highly polygenic psychiatric diseases) by ML has high potential to increase the likelihood of medical discrimination against population subgroups, with no benefit of accurate prediction for early intervention.

Opioid use disorder (OUD) is a complex trait associated with high disease burden, estimated to affect 2% of the U.S. population[1]. Predictive tools that might identify at-risk individuals for prevention and early intervention are being developed[2], and because OUDs are moderately heritable (h^2 = 30-70%[3]), incorporating genetic variation into a predictive tool has great appeal as one way to combat the ongoing opioid crisis. However, OUD is highly polygenic with a large number of variants of small effect contributing to its heritability. The largest genome-wide association study (GWAS) of OUD to date (15,756 OUD cases and 99,039 controls) identified one genome-wide significant variant, rs1799971, in the gene encoding the mu opioid receptor (*OPRM1*)[4]; the effect size associated with this variant was small (β = -0.066 [SE = 0.012]). Current estimates of total single nucleotide polymorphism (SNP)-based heritability of OUD is 11% (SE = 0.018)[4], putting a limit on overall predictive ability using common variants. Based on these findings, it is unlikely that a genetic predictor of OUD would be clinically meaningful.

For psychiatric disorders such as OUD, inaccurate predictive tests pose substantial hazards; the harms attributable to a false positive result include withholding beneficial medication and

potential discrimination (e.g., employer bias). Therefore, the selection of genetic content and the interpretation of such tests must be rigorously evaluated. Here, we examine two critical considerations in genetic prediction tools, particularly those developed using machine learning (ML): population stratification and variant (feature) selection. As a cautionary tale, we illustrate how such a combination of inappropriately selected genetic variants and lack of proper consideration of population stratification has resulted in a flawed genetic prediction tool for OUD that is currently being evaluated by the U.S. Food and Drug Administration as an Investigational Breakthrough Device, commercially advertised as LifeKit Predict® (https://prescientmedicine.com/technologies/lifekit-predict/; accessed September 8th, 2020).

The genetic component of the LifeKit Predict® prediction algorithm relies on 15 or 16 candidate SNPs depending on the version of the test[5]. Only one of these (rs1799971) has been confirmed by GWAS as affecting OUD risk; the very small effect size of this variant ($\beta = -0.066$), although statistically significant, is unlikely to be clinically relevant. Investigator-selected candidate variants (e.g., in dopamine system genes) are routinely favored by those developing purported prediction tools for addiction, despite the scientific consensus regarding the weaknesses inherent in the selection of candidate genes[6–8]. Current genetic research relies on hypothesis-free methods such as GWAS for variant discovery; most candidate variants have not been replicated in such unbiased approaches.

Regardless of variant choice, a critical confound inherent to such genetic tests is population stratification, which can produce false positive (or negative) results stemming from differences in allele frequencies and linkage disequilibrium patterns across global populations[9]. A recent evaluation of ML methods in psychiatry found that none included procedures to account for population stratification, i.e., in samples of individuals of different ancestries, there is confounding of variants that predict an outcome with variants whose allele frequencies differ

across global populations[10]. When ML algorithms utilize unsubstantiated candidate variants and do not properly account for population stratification, they produce “predictions” that are not only spurious but also potentially discriminatory, and risk stigmatizing members of some ancestral populations. As shown in **Figure 1**, the 16 variants selected for LifeKit Predict® have large differences in allele frequency between African-American and European-American individuals and this pattern extends to many global populations. This raises the strong possibility of spurious findings should these population differences remain insufficiently accounted for in any genetic analysis.

To investigate the behavior of ML methods in this context, we implemented 5 different supervised ML models to predict opioid dependence (OD), in independent training and testing subsets of 1000 individuals from the Yale-Penn sample[4], which includes carefully-diagnosed OD cases and opioid-exposed non-OD controls. Subsets were completely balanced, with 250 OD cases of European (EUR) descent and 250 cases of African (AFR) descent and 250 each of their ancestrally matched controls (**Supplemental Figure S2C)**; ancestry was derived from whole genome data (see online methods)[4]. The ML models used were (i) (extreme) gradient boosted machines (GBMs), (ii) linear support vector machines (linear SVMs), (iii) radial basis support vector machines (nonlinear SVMs), (iv) random forests (RFs), and (v) elastic nets (ENs). All models were trained using the panel of 16 SNPs referenced in Donaldson et al.[5], the basis for LifeKit Predict®.

We first examined whether the accuracy of the model in predicting OD changed as the ancestry ratios in the test sample were varied. Learning curves were iteratively estimated, starting with complete confounding of ancestry – i.e., cases from one ancestry group and controls from the other (250 EUR cases and 250 AFR controls, see **Supplemental Figure S2A-2C**) – and

progressively balancing the case-control ancestry proportions in steps of 10 subjects until the 26th iteration, which reached perfect ancestral balance.

For all 5 ML methods, the prediction of case status was high (Area Under the Curve, AUC > 0.8) when the sample was fully confounded, and case-status prediction decreased as samples were better ancestrally balanced, until the prediction was no better than expected by chance alone in a balanced sample (AUC approached 0.5; **Figure 2A-2E**). At every iteration of every ML approach, the 16 variants predicted genomic ancestry much better than OD. Insufficiently controlled population stratification will confound genetic studies regardless of the complex trait under study. Our results document that ML algorithms are not exempt from this confound, and might even favor the identification of population stratification over case status.

Second, we examined whether the 16 candidate variants selected for their presumed neurobiological significance (i.e., "involved in the brain reward pathways"[5]) performed better than 8 (unique and non-overlapping) permuted sets of 16 SNPs selected at random from groups of SNPs matched on their minor allele frequencies in EUR and AFR to those in LifeKit Predict®. All iterations with 8 permutations of random (allele-frequency-matched) SNPs performed similarly to the chosen candidate SNPs in Lifekit Predict® variants (**Figure 3** for 1 permutation**,** the other 7 are available as **Supplemental Figure S3**). Specifically, across all iterations of all permutations the ML models were highly predictive of OD when highly confounded by ancestry, decreasing in prediction as ancestral balance improved, and generally, better predictors of ancestry than OD. Therefore, the LifeKit Predict® selected variants perform no better than randomly selected variants with the same ancestral variability in allele frequencies.

Third, as African Americans (and other minority groups in the U.S.) include substantial European admixture[11] we examined whether the 16 variants predicted the extent of EUR

admixture within the AFR cases and controls. We chose the 15th iteration (**Figure 2**) of the learning curve because this iteration had the greatest balance of ancestry that still offered some prediction of OD that was greater than chance. Across approaches, ML models designed to predict OD were up to 5 times better predictors of the percent of EUR admixture in African-American individuals, than of OD (i.e., case status) (**Figure 4**).

Our findings indicate that the genetic component of the LifeKit Predict® test is unlikely to meaningfully contribute to OUD prediction. More generally, ML models trained either on a handful of selected variants or across the whole genome are highly sensitive to confounding by genetic ancestry[12]. Once ancestry is accounted for, these models offer evidence for prediction of OD that is no greater than chance. With insufficient control for population stratification, as is commonplace with ML approaches favored by developers of genetic prediction tests, tests such as LifeKit Predict® estimate an individual's genetic ancestry rather than case status, thus raising the strong possibility that individuals of some ancestral backgrounds will be disproportionately labeled as "at-risk" for development of OUD related to opioids prescribed for acute pain relief (as claimed by LifeKit Predict®) and potentially denied appropriate medical treatment.

Opioids are useful for pain management and analgesia. African-American and Latinx Americans are less likely to be prescribed opioids, an observation potentially attributable to physician bias[13]. Our analyses carry the caution that this racial disparity may be perpetuated and exacerbated by the use of genetic prediction tests for OUD that do not adequately account for allelic differences, not just between European and African Americans but also in other U.S. populations (e.g., Latinx Americans, who are genetically very diverse[11]) and even within broad self-identified racial groups (e.g., the extent of EUR admixture within African-Americans). These cautions generalize to other substance use disorders and other highly polygenic psychiatric illnesses, especially when there is a reliance on variants that lack robust support. Our findings

serve as a cautionary tale for efforts to develop genetic precision medicine, particularly for traits like addiction that carry stigma and potential discrimination.

Funding: This research is supported by MH109532. ASH acknowledges support from DA007261; AA acknowledges support from K02DA032573. Yale-Penn (phs000425.v1.p1; phs000952.v1.p1) was supported by National Institutes of Health Grants RC2 DA028909, R01 DA12690, R01 DA12849, R01 DA18432, R01 AA11330, and R01 AA017535 and the Veterans Affairs Connecticut and Philadelphia Veterans Affairs Mental Illness Research, Educational, and Clinical Centers.

Disclosures: HRK is an advisory board member for Dicerna and a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which was supported in the last three years by AbbVie, Alkermes, Dicerna, Ethypharm, Indivior, Lilly, Lundbeck, Otsuka, Pfizer, Arbor, and Amygdala Neurosciences. HRK and JG are named as inventors on PCT patent application #15/878,640 entitled: "Genotype-guided dosing of opioid agonists," filed January 24, 2018.

Reference Cited

1. Saha, T. D. *et al.* Nonmedical prescription opioid use and DSM-5 nonmedical prescription opioid use disorder in the United States. *J. Clin. Psychiatry* **77**, 772–780 (2016).
2. Ellis, R. J., Wang, Z., Genes, N. & Ma'ayan, A. Predicting opioid dependence from electronic health records with machine learning. *BioData Min.* **12**, 3 (2019).
3. Sun, J. *et al.* Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. *Addict. Behav.* **37**, 1138–1144 (2012).
4. Zhou, H. *et al.* Association of OPRM1 Functional coding variant with opioid use disorder: a genome-wide association study. *JAMA Psychiatry* (2020) doi:10.1001/jamapsychiatry.2020.1206.
5. Keri Donaldson, Laurence Demers, Joe Lopez, and S. C. Multi-variant genetic panel for genetic risk of opioid addiction. *Ann. Clin. Lab. Sci.* **47**, 452–456 (2017).
6. Johnson, E. C. *et al.* No evidence that schizophrenia candidate genes are more associated with schizophrenia than noncandidate genes. *Biol. Psychiatry* **82**, 702–708 (2017).
7. Border, R. *et al.* No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am. J. Psychiatry* **176**, 376–387 (2019).
8. Duncan, L. E. & Keller, M. C. A Critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am. J. Psychiatry* **168**, 1041–1049 (2011).
9. Hamer, D. H. & Sirota, L. Beware the chopsticks gene. *Mol. Psychiatry* **5**, 11–13 (2000).
10. Bracher-Smith, M., Crawford, K. & Escott-Price, V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry* 1–10 (2020) doi:10.1038/s41380-020-0825-2.
11. Jordan, I. K., Rishishwar, L. & Conley, A. B. Native American admixture recapitulates population-specific migration and settlement of the continental United States. *PLOS*

Genet. **15**, e1008225 (2019).

12. Polimanti, R., Yang, C., Zhao, H. & Gelernter, J. Dissecting ancestry genomic background in substance dependence genome-wide association studies. *Pharmacogenomics* **16**, 1487–1498 (2015).

13. Meghani, S. H., Byun, E. & Gallagher, R. M. Time to take stock: A meta-analysis and systematic review of analgesic treatment disparities for pain in the united states. *Pain Medicine* vol. 13 150–174 (2012).

14. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

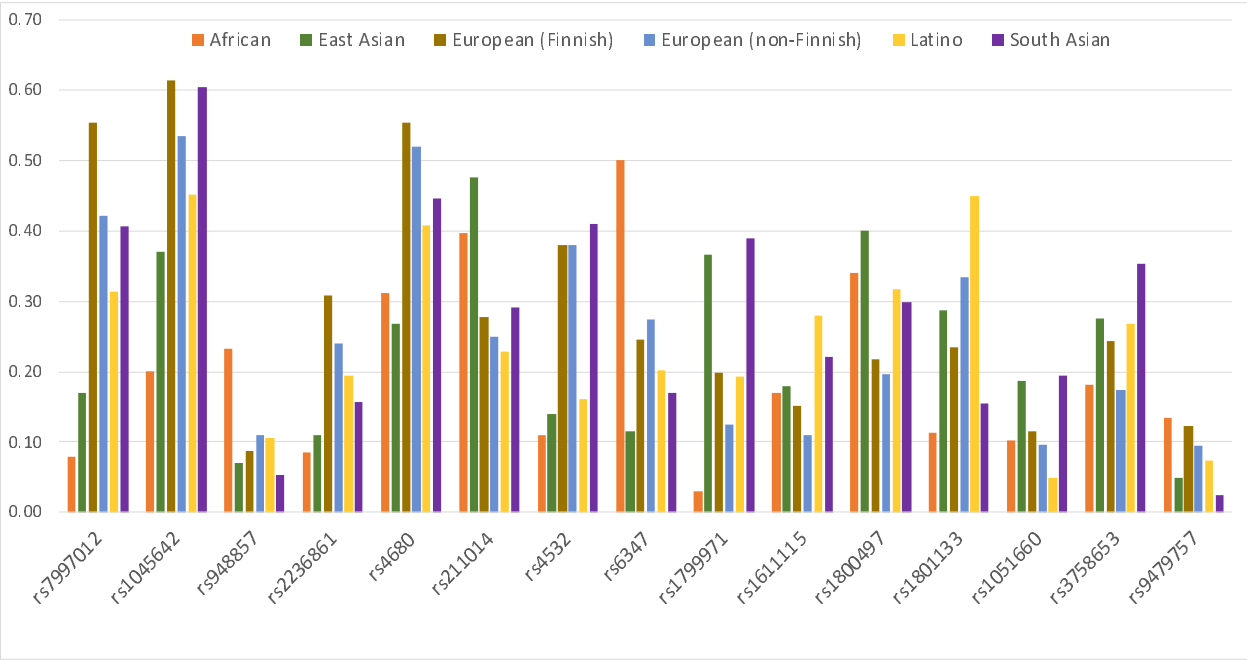

Figure 1. Population allele frequencies (from GNomad[14]) for the candidate alleles in Donaldson et al.[5] LifeKit Predict® (https://prescientmedicine.com/technologies/lifekit-predict/; accessed September 8th, 2020) across different major geographic ethnic groups showing substantial variation in frequency across global ancestral populations.

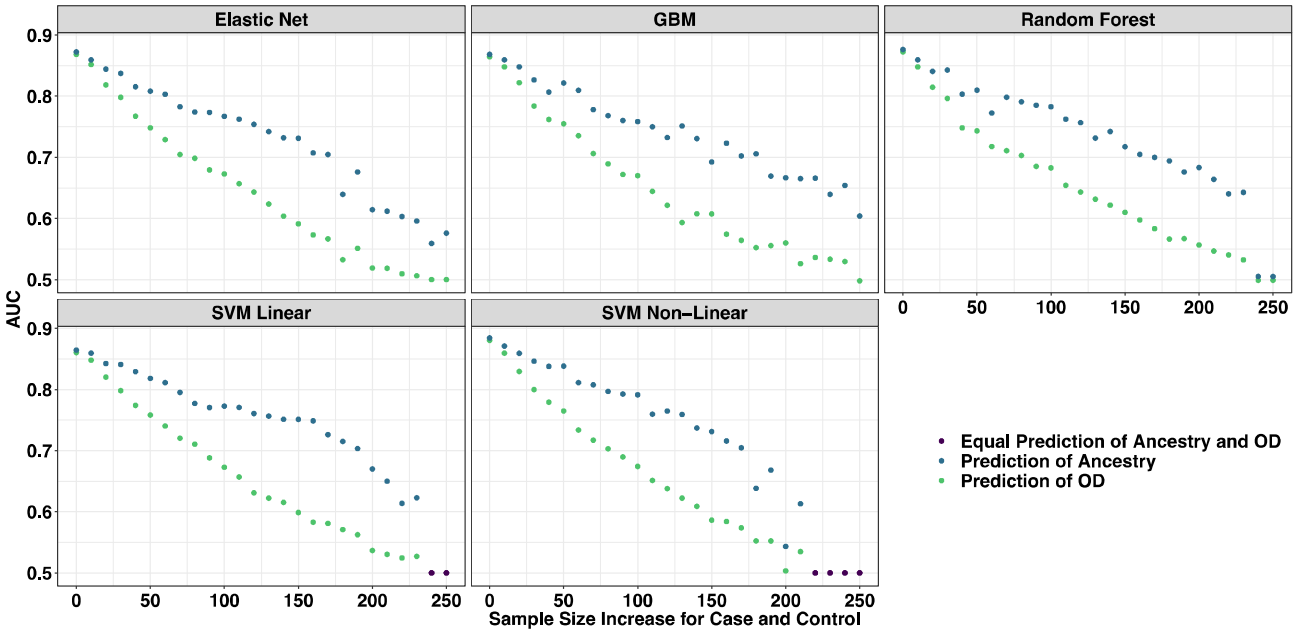

Figure 2. Learning curves from models trained to predict opioid dependence from 16 "reward-related" SNPs[5]. The curves are plotted by AUC based on their prediction of opioid dependence (orange) and geographic ancestry (blue) as the samples start from complete population confounding become more balanced by major geographic ancestry (European American or African American) until completely balanced. Each data point represents a larger and more balanced sample size by adding 20 individuals, 10 African-American cases and 10 European-American controls (as measured on the x-axis).

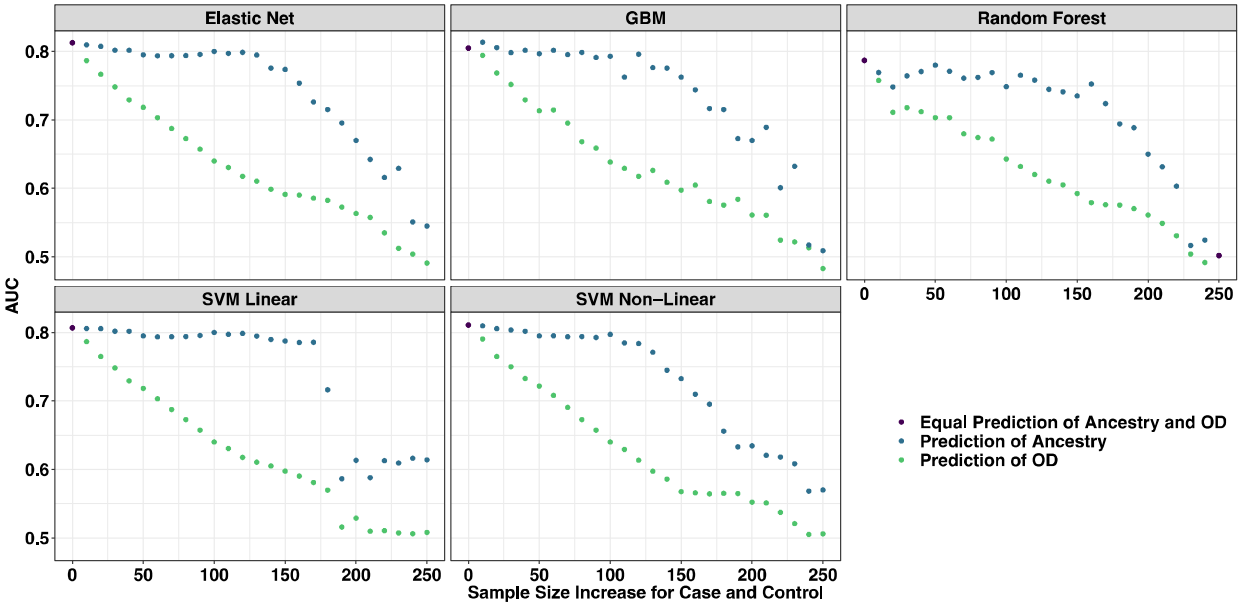

Figure 3. Learning curves from a model trained to predict opioid dependence from 16 MAF matched SNPs. The curves are plotted by AUC based on their prediction of opioid dependence (orange) and geographic ancestry (blue) as the samples (each data point) start from complete population confounding become more balanced by major geographic ancestry (European American or African American) until completely balanced, as in Figure 2.

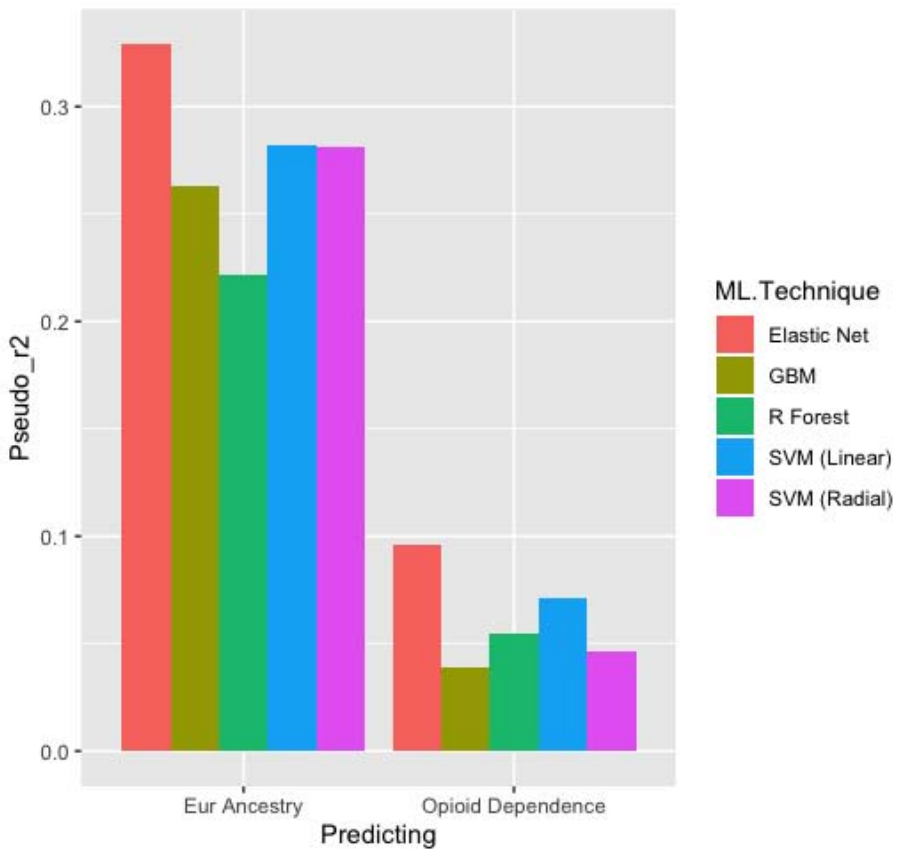

Figure 4. Bar plots of the pseudo r^2 from a logistic regression comparing the predictions of opioid dependence and percentage of European ancestry in a sample of 250 African American individuals from the Yale-Penn Test set. Pseudo r^2 was used instead of AUC because the percentage of European descent is a continuous variable and this put both predictions on the same scale.