# Genetic Data Can Lead to Medical Discrimination: Opioid Use Disorder as a Cautionary Tale

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 J. 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, Department of Psychiatry, 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

Funding: This research is supported by MH109532. ASH acknowledges support from DA007261; AA acknowledges support from K02DA032573; FRW acknowledges support from F32 MH122058. 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, Education 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 Alkermes, Amygdala Neurosciences, Arbor, Dicerna, Ethypharm, Indivior, Lundbeck, Mitsubishi, Otsuka, 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.

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

## Abstract

Using genetics to predict the likelihood of future psychiatric disorders, such as opioid use disorder (OUD), poses scientific and ethical challenges. Machine learning models are beginning to proliferate in psychiatry, however, most machine learning models in psychiatric genetics to date have not accounted for ancestry. Using an empirical example of a proposed genetic test for OUD and by generating a simulated random binary phenotype, we show that ML genetic prediction is completely confounded by ancestry, potentially discriminatory, and of no benefit for clinical practice. In an empirical example, we examine results from five ML algorithms trained with brain reward-derived "candidate" SNPs proposed for commercial use and demonstrate that the algorithms do not predict OUD better than chance when ancestry is balanced but are highly confounded with ancestry in an out-of-sample test set. We show how such a test could also predict subpopulations in admixed samples. Random sets of variants matched to the candidate SNPs by allele frequency produced similarly flawed predictions, further questioning the plausibility of selecting candidate variants. Finally, using random SNPs that predict a random simulated phenotype we show that the bias attributable to ancestral confounding would impact any such ML-based genetic prediction algorithm. Given the small and distributed single-variant genetic effect sizes associated with most psychiatric disorders, researchers and clinicians are encouraged to be skeptical of claims of high prediction accuracy from the growing number of ML-derived genetic algorithms, particularly when models are naive to polygenicity and ancestral confounding.

Machine learning (ML) applications are increasingly used to leverage big data from electronic health records to classify patient populations[1]. In the realm of direct to consumer (DTC) genetic testing, ML approaches are gathering momentum, especially for psychiatric disorders. Currently, several commercial entities offer genetic testing for psychiatric disorders, and some have begun to offer controversial and scientifically disproven proposals for genetic embryo selection for behavioral and psychiatric traits[2]. While most genetic tests within psychiatry are aimed at medication efficacy in patients (e.g., pharmacogenetic or pharmacokinetic testing), a few recent tests target prediction of future psychiatric disorders, which has the potential of stigmatizing individuals. Alongside the ethical challenges of such predictions lie the scientific limitations. The genetic "inputs" that are the information used by these DTC tests typically comprise of "candidate gene variants" that are scored using pattern recognition software, powered with "artificial intelligence" (machine learning or ML) frameworks. While most of these candidate variants are not significantly associated with the disorders in genome-wide analyses, they continue to be used by some as if they were markers of disease risk[3]. Exacerbating the problem, past work on ML algorithms in psychiatric genetics has shown that these models often ignore common confounds, particularly ancestry[4]. As a consequence, patients and physicians are now confronted with an entrée assembled with indigestible ingredients and a flawed recipe.

One psychiatric illness that is being targeted by ML-based genetic algorithms is opioid use disorder (OUD), a complex trait associated with high disease burden, and estimated to affect 2% of the adult population[5]. Predictive tools that aim to identify at-risk individuals for prevention and early intervention are being developed[1], and because OUDs are moderately heritable ($h^2$ = 30-70%[6]), incorporating genetic variation into a predictive tool has great appeal.

In addition, because opioids comprise front-line pain management drugs, biomarkers that index risk of OUD in this setting are of potential interest. 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*)[7]; the effect size associated with this variant was small (β = -0.066 [SE = 0.012]). Current estimates of the total single nucleotide polymorphism (SNP)-based heritability of OUD is 11% (SE = 1.8%)[7], putting a limit on overall predictive ability using small numbers of common variants. Based on these best-to-date findings, it is unlikely that a genetic predictor of OUD would be clinically meaningful. Thus, we hypothesized that 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.

For psychiatric disorders such as OUD, inaccurate predictive tests pose substantial hazards; the harms attributable to a false positive result include both withholding beneficial medication and potential discrimination (e.g., employer bias). Such tests must be rigorously evaluated. Here, we examine two critical considerations in genetic prediction tools, particularly those developed using ML: population stratification and variant (feature) selection. We show that a combination of inappropriately selected genetic variants and inadequate consideration of population stratification has resulted in a flawed genetic prediction tool for OUD that is nevertheless currently being evaluated by the U.S. Food and Drug Administration, commercially advertised as LifeKit Predict® (https://labservices.prescientmedicine.com/testpanels/lifekit-predict; last accessed December 13th, 2020). We show that this tool predicts ancestry rather than risk for OUD.

However, our observations are not limited to OUD – to demonstrate the generalizability of the peril of ancestral confounding in ML, we generate random genotypes and simulated phenotypes to document that ancestral confounding produces seemingly accurate prediction even when the phenotype is random noise and the variants are selected randomly from the genome. Finally, we demonstrate that, even within broadly-defined population groups, such ML genetic algorithms are predictors of subgroups within the population rather than predictors of diagnostic status.

## Methods

*Selection of direct-to-consumer test methodology for comparison*

We evaluated current DTC tests by conducting a web search within Google for "Genetic Testing for Psychiatry" and "Genetic Testing for Addiction" and selecting all tests from the first five pages. **Supplemental Table 1** shows a list of the 12 tests that were found, and known mechanisms for evaluation. The methodology presented in the current study was based on those used by Prescient Medicine's LifeKit Predict®, a DTC genetic testing kit for OUD. This test was selected because (a) it purports to predict, with 97% accuracy (https://labservices.prescientmedicine.com/testpanels/lifekit-predict) risk for OUD, a complex trait that has been shown to be highly polygenic, (b) is accompanied by a training procedure that was published and therefore, can be recapitulated. While still not completely transparent, this was the only test with enough information to provide a demonstration, which is described below.

The genetic component of the LifeKit Predict® prediction algorithm relies on 15 or 16 candidate single nucleotide polymorphisms (SNPs) depending on the version of the test[3], most of which are used in other DTC tests (see **Supplemental Table 1**) and have often been labeled in the psychiatric genetics literature as "candidate genes"[8,9]. Only one of these SNPs

(*OPRM1**rs1799971) has been shown by GWAS to affect OUD risk; the very small effect size of this variant (standardized beta= -0.066), although statistically significant, is unlikely to be clinically relevant. Industry-selected candidate variants (e.g., in dopamine and serotonin candidate genes) are routinely favored by those developing purported prediction tools for addiction, despite the scientific consensus regarding the weaknesses inherent to selection of candidate genes[8–10]. With the exception of rs1799971, none of the candidate SNPs in the LifeKit test has been associated at genome-wide significant levels (p<5E-8) (**Supplemental Table 2**) with *any* complex trait in the GWAS Atlas[11] (**Supplemental Table 3**). However, while these SNPs are not associated to psychiatric phenotypes, the MAFs of many of the candidate SNPs vary greatly across ancestral populations (**Figure 1**). That is, taken individually, they tend to be associated with one's *ancestral population*, but not to *trait*. Accordingly, it was our expectation that sets of these markers would also necessarily be associated to population but not to trait – regardless of the sophistication of the interposed statistical methodology.

*Sample description*

We used subjects recruited as part of the Yale-Penn study. These participants were recruited at five sites across the eastern United States to study the genetics of substance dependence and comorbid psychiatric and behavioral phenotypes[12]. All participants were interviewed with the Semi-Structured Assessment for Drug Dependence and Alcoholism[13] and provided written informed consent through a protocol approved by the institutional review board at each participating site – Yale Human Research Protection Program (protocols 9809010515, 0102012183, and 9010005841), University of Pennsylvania Institutional Review Board, University of Connecticut Health Center Institutional Review Board, Medical University of

South Carolina Institutional Review Board for Human Research, and the McLean Hospital Institutional Review Board.

*Genotyping and quality control*

The Yale-Penn phase 1 sample was genotyped using the Illumina HumanOmni1-Quad array. Individuals with mismatched sex or genotype call rate < 98% were removed; SNPs with genotype call rate < 98% or minor allele frequency < 0.01 were removed before imputation. Imputation was performed using Minimac3[12] and the Haplotype Reference Consortium reference panel implemented in the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html). More details on the Yale-Penn sample can be found elsewhere[14].

Genetic ancestral group was defined by principal component (PC) analysis on genotyped SNPs (pruning by linkage disequilibrium of $r^2 > 0.2$) and the 1000 Genome phase 3 reference panels[15] using EIGENSOFT[16,17]. The first 10 PCs were used to cluster the participants into African-American and European-American groups and to remove outliers from the 2 groups. European-ancestry proportions in African-American samples were estimated using ADMIXTURE[18]. SNPs were included in ancestry prediction following the developer's recommended independent SNP selection procedures. ADMIXTURE's cross-validation (CV) procedure was used to determine the most appropriate K – the most sensible number of component ancestries with which to model unknown sample ancestries. Based on lowest CV error and failure to reduce substantially CV error with additional K, we chose K=2 as appropriate for these data. The mean European ancestry for the African-Americans used in this investigation was 23.2%.

*Selection of random sets of 16 variants.*

We calculated in the Yale-Penn sample the minor allele frequency (MAF) of the SNPs in Donaldson et al.[3] using PLINKv1.90 over a total sample of 5057 individuals (3286 African American and 1768 European American). In addition to the SNPs used by Donaldson et al, for comparison, we identified 8 lists of random SNPs, within which each SNP from Donaldson et al. was replaced by a random SNP with matched MAF; all random SNPs were unique.

*Machine learning training procedure*

Each of the 16 alleles was dummy coded for homozygosity or heterozygosity status to allow for interactions among different levels of dosage of each minor allele from each SNP with minor allele dosages of other SNPs (i.e., to measure epistasis, if present). Each supervised ML algorithm was trained separately in the same training set, which varied from 500 to 1000 individuals based on k iteration of the learning curve (see **Figure 2**). The training and test sets were initially analyzed in a way that assured they were completely confounded by population differences, with all cases of European ancestry and all controls of African ancestry. At each iteration of the learning curve, we added 10 individuals of African descent to the cases and 10 individuals of European descent to the controls, through 26 steps to reach completely balanced samples (see **Figure 2**).

We chose 5 broad algorithm-generating methods for our analysis as a survey of supervised ML because they have been used by academics and commercial entities to attempt to predict OUD. All were implemented in the Caret package[19] in R version 3.6.110[20]: (1) (extreme) Gradient Boosted Machines (GBM)[21], which incorporate population stochastic gradient descent procedures that are ubiquitous in industry. (2) Linear and (3) nonlinear (radial basis function)

Support Vector Machines (SVM) to compare predictive accuracy from different kernels[22]. (4) Random Forests (RF) to represent more complex tree structures[23], and (5) Elastic Nets (EN) for representation of (flexible) linear regression models[24]. All models were trained with 10-fold cross validation and a hyperparameter grid-search in the training set. All AUC and pseudo r2 were extracted from the non-overlapping test set. Learning curves were plotted at all iterations.

*Generalizability of confounding using a random binary phenotype*

To examine the generalizability of this confounding, beyond the test case of OUD prediction, we generated a random binary phenotype by drawing from a binomial distribution. That is, the phenotype was essentially random noise and therefore not truly predictable. We matched the number of cases and controls for our random variable by geographic ancestry, such that we ended with the same split in cases and controls by ancestry as was used in our OUD demonstration. We then took the 8 random SNP set permutations and used them to predict the random noise. Because we used random SNPs with random outcomes, the effects provide an empirical NULL hypothesis: what the data look like when the result is by definition meaningless. We hypothesized that this empirical null will still show high effect sizes at high confounding, i.e. even random noise can seemingly be "accurately" predicted when the sample is confounded.

## Results

*Evaluation of a modern direct-to-consumer psychiatric diagnosis kit in the presence of confounding by ancestry.*

For our empirical test, all models were trained using the panel of 16 SNPs referenced in Donaldson et al.[3], the basis for LifeKit Predict®, and were trained to predict OUD in the Yale-

Penn sample. These 16 variants demonstrate substantial allele frequency differences across ancestries (**Figure 1)**. As shown in Figure 3A for all 5 ML methods, prediction of OUD case status was high (Area Under the Curve, AUC > 0.8) when the sample was fully confounded (that is, when predictions were essentially predictions of ancestry), 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). At every iteration of every ML approach, the 16 variants predicted genomic ancestry much better than they predicted OUD.

*Random SNPs predict OUD as well as biologically plausible SNPs due to confounding.*

All iterations with 8 permutations of random (minor allele frequency matched) SNPs performed similarly to the models with chosen candidate variants (**Figure 3B** shows 1 permutation**,** the other 7 are shown as **Supplemental Figure 1**). Across all iterations of all permutations the ML models were highly predictive of OUD only when confounded by ancestry, decreasing in prediction as ancestral balance improved. The models remained better predictors of ancestry than OD. Therefore, the selected variants perform no better than randomly selected variants with the same ancestral allele frequencies.

*Ancestral confounding leads to random genotypes making apparently accurate predictions of random phenotypes, mirroring results for OUD.*

We next took the 8 random subsets of SNPs and used them to predict a randomly generated phenotype. For a randomly generated phenotype, at perfect confounding by major geographic ancestral group, we get high apparent predictive accuracy of random noise, but as we

balanced the training and test sets by ancestry, our model performs as expected, with prediction no better than a coin flip **(Figure 3C)**. Across all iterations (**Supplemental Figure 2**), all models trained to predict random noise were stronger predictors of ancestry than random noise, suggesting that even if the outcome is meaningless, we can gain the appearance of meaningful results in the presence of ancestral confounding.

*Confounded models are better at detecting subpopulation within minority populations than diagnosis.*

As African Americans (and other minority groups in the United States) include substantial European admixture[25] we examined whether the 16 OUD variants used in the LifeKit test predicted the extent of EUR admixture within the AFR cases and controls. We chose the 15$^{th}$ iteration (**Figure 3**) of the learning curve as it had the greatest balance of ancestry that still offered some prediction of OUD that was greater than chance. Across all approaches, ML models designed to predict OUD were up to 5 times better predictors of the percent of EUR admixture in African-American individuals than of OUD (i.e., case status) (**Figure 4**).

## Discussion

ML models trained either on a handful of selected variants or across the whole genome are strongly sensitive to confounding by genetic ancestry[26]. They are supposed to offer high clinical prediction, but that can result from an imbalance of ancestries within cases and/or controls. We demonstrate these underlying problems for a specific genetic test for OUD, but our simulations demonstrate that the confound is generalizable – once ancestry is accounted for, these models offer no evidence of predictive ability greater than chance. This raises the strong

possibility that individuals of some ancestral backgrounds will be disproportionately labeled as "at-risk" for developing OUD. In the context proposed for the test we evaluated, related to oral opioid prescriptions for acute pain relief, African Americans stand a high risk of being denied appropriate medical treatment, of being stigmatized if results are introduced into their medical records, and of potential emotional trauma from such labeling. Our findings argue for great caution when evaluating results of other ML-based genomic analyses that do not explicitly and fully account for ancestral confounding. From the prescriber viewpoint, we conclude that given the high polygenicity of OUD, at this time, no genomic test claiming high accuracy in predicting the disorder is trustworthy.

In the field of ML, our results fall under the category of algorithmic bias. The uncritical reverence some seem to feel for ML, or colloquially "artificial intelligence", is due to these methods' powerful pattern-recognition capabilities. However, many examples in healthcare research outside genetics[27] show that pattern recognition with little understanding of underlying effects may recapitulate known health disparities. For example, algorithms that stratify patient health resources are biased by social population stratification[27], potentially discriminating against African-American communities. Here, we demonstrated that in the context of genomic data, this algorithmic bias was generated by population stratification, a well-characterized phenomenon in statistical genetics[28] that is yet to be widely dealt with by the ML field[4]. In particular, while standard genetic analyses are confounded, we assert that ancestral confounding by ML can render results unreliable and potentially discriminatory.

As human genetics has shown, this challenge is surmountable. Most ML algorithms allow for some form of de-confounding, typically as a multi-step or multi-model procedures. While typical ML pipelines employ multiple algorithms and simply select the best, more extensive

individual attention to the choice of algorithm is needed to evaluate confounding in the face of known covariates. Several avenues may be pursued based on the choice of a model. For example, in this paper (and in those models used by industry professionals) gradient boosted machines were used, which can (but have yet to) include sample weights in the model training procedure. Corrections for support vector machines also exist, and remove statistical dependence in the model training procedure[29]. Extensive work with each algorithm will best determine future routes for de-confounding, and needs to be an essential part of model training beyond just predictive accuracy. If one wishes to cut across algorithms, we show that learning curves should be purposefully developed with stratification in mind to ensure that lingering cryptic admixture does not confound predictions. Finally, statistical procedures will only take us so far, and careful considerations of samples and confounding, as is standard in GWAS literature, will only serve to reduce confounding in machine learning and genetic testing practices. However, restricting analyses to one continental ancestry (e.g., Europeans only) is not the solution to ancestrally confounded analyses. While such ancestral homogeneity may attenuate gross confounding, we show that cryptic admixture remains an issue. Instead, larger training and testing samples of diverse ancestral populations are needed to accelerate genomic discovery and ensure that when aggregated effect sizes are large enough, precision medicine will benefit all global communities[30].

Even with appropriate adjustment for admixture, it is unlikely that candidate variants will produce any meaningful prediction of OUD, or as we show, any other complex trait. Recent meta-analyses of depression[8], schizophrenia[9], and executive function[31] show that overwhelmingly, candidate variants do not rise to levels of genome-wide significance in psychiatry. Even in instances where GWAS has identified a variant in a previously nominated

gene, the genome-wide significant variant has been found to be distinct from candidate variants in the same gene. For instance, variants in *DRD2* are associated with schizophrenia[32], depression[33], problematic alcohol use[34], and tobacco smoking[35]. However, rs1800497, the most frequently studied candidate variant linked to the gene starting in the early 1990's (although it is in a neighboring gene, *ANKK1*) has not been implicated nor shown to be in linkage disequilibrium with the genome-wide significant signals for any of the indicated traits. Thus, a majority of psychiatric geneticists contend that the debate regarding the limited plausibility of most candidate variants in investigator-nominated genes (with a few exceptions in the field of substance use disorders – *ADH1B*, *ALDH2*, *CHRNA5*, *CYP2A6*, and *OPRM1*) has been settled. Yet, commercial entities have continued to rely on these variants, ostensibly because of their generalized appeal as variants in "reward-related" genes (e.g., *DAT1*, *DBH*), or neuroplasticity genes (e.g., *SLC6A4*). Awareness among healthcare providers who are likely to request such genetic testing is vital to ensure that the misguided deployment of these candidate variants is not viewed as being supported by the peer-review literature.

Finally, our empirical work looked at OUD. Opioids are useful for pain management and analgesia, but are also highly addictive. Against the backdrop of the opioid epidemic, the urgency for tests that can provide any insights into the likelihood of patients developing OUD is understandable. In genetics, a field which is notorious for very small effect sizes, DTC or physician-engaged testing that offers prediction accuracies above 80% seem like a breakthrough. It is unlikely that genetic testing will provide this panacea, and it raises serious ethical problems. African Americans and Latinx Americans are already less likely to be prescribed opioids for pain relief, potentially attributable to physician bias[36]. Our findings 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 Americans and African Americans but also in other U.S. populations (e.g., Latinx Americans, who are genetically very diverse[25]) 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 reliance on variants that lack robust empirical support. Our findings serve as a cautionary tale for efforts to advance genetic precision medicine, particularly for traits like addiction that carry stigma and are potential sources of discrimination.

## Reference Cited

1. 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).
2. Karavani, E. *et al.* Screening Human Embryos for Polygenic Traits Has Limited Utility. *Cell* **179**, 1424-1435.e8 (2019).
3. 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).
4. 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.
5. 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).
6. Sun, J. *et al.* Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. *Addict. Behav.* **37**, 1138–1144 (2012).
7. 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.
8. 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).
9. 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).

10. 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).

11. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0481-0.

12. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

13. Consortium, the H. R. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

14. Zhou, H. *et al.* Genetic risk variants associated with comorbid alcohol dependence and major depression. *JAMA Psychiatry* **74**, 1234–1241 (2017).

15. Auton, A. *et al.* A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).

16. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).

17. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

18. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

19. Kuh, M. caret: Classification and Regression Training. (2020).

20. Team, R. C. R: A language and encironment for statistical computing. (2019).

21. Chen, Tianqi. He, Tong. Benesty, Mchiael. Khotilovich, Vadim. Tang, Yuan. Cho, Hyunsu. Chen, Kailong. Mitchell, Rory. Cno, Ignacio. Zhou, Tianyi. Li, Mu. Junyuan,

Xie. Lin, Min. Geng, Yifeng& Li, Y. xbgoost: Extreme Gradient Boosting. (2020).

22. Kaatzoglou, Alexandros. Smola, Alex. Hornik, Kurt. Zeileis, A. kernlab - An S4 Package fo Kernel Metods in R. **9**, 1–20 (2004).

23. Liaw. Wiener, M. Classification adn Regression by randomForest. 18–22 (2002).

24. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

25. 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).

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

27. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (80-. ).* **366**, 447–453 (2019).

28. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

29. Li, L., Rakitsch, B. & Borgwardt, K. ccSVM: Correcting Support Vector Machines for confounding factors in biological data classification. *Bioinformatics* **27**, i342 (2011).

30. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

31. Hatoum, A. *et al.* GWAS of Over 427,000 Individuals Establishes GABAergic and Synaptic Molecular Pathways as Key for Cognitive Executive Functions. *GWAS Over 427,000 Individ. Establ. GABAergic Synaptic Mol. Pathways as Key Cogn. Exec. Funct.*

674515 (2019) doi:10.1101/674515.

32. Schizophrenia Working Group of the Psychiatric Genomics Consortium, S. W. G. of the P. G. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–7 (2014).

33. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).

34. Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.* 1–10 (2020) doi:10.1038/s41593-020-0643-5.

35. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics* vol. 51 237–244 (2019).

36. 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).

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

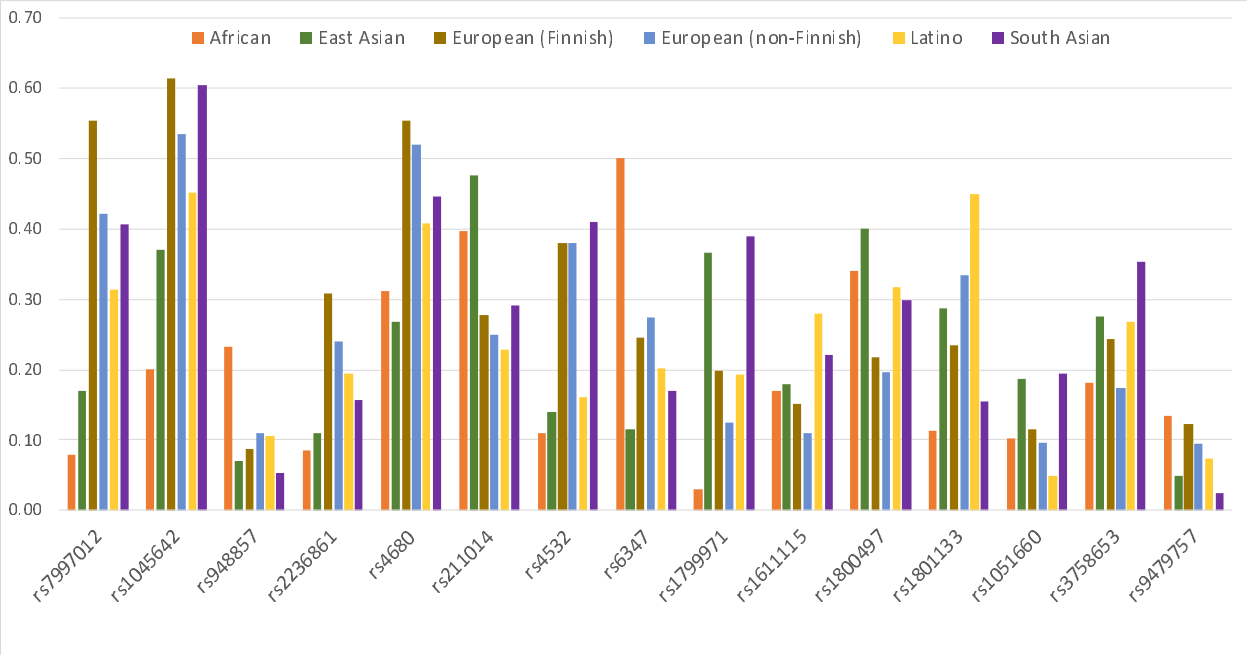


**Figure 1.** Population allele frequencies (from GNomad[37]) for the candidate alleles in Donaldson et al.[3] LifeKit Predict® (https://prescientmedicine.com/technologies/lifekit-predict/; accessed September 8th, 2020) across different major geographic ethnic groups showing substantial variation in frequency across global populations.

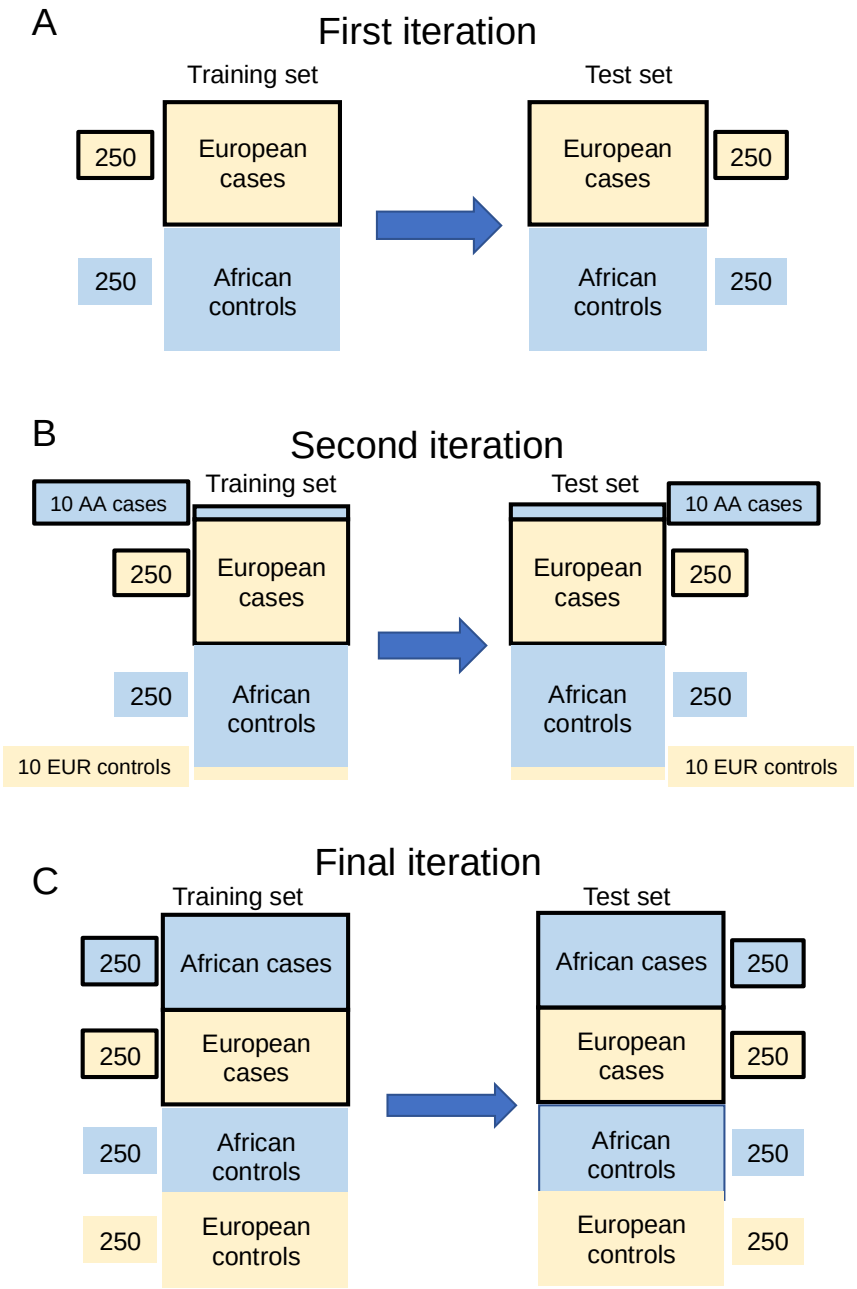


**Figure 2.** Learning curve procedure for all analyses. We had a complete and non-overlapping training and test set, each of 1000 subjects with 250 cases and controls of European and African descent. (A) For the first iteration we started with 250 subjects of European descent (tan) that were OUD cases, and 250 subjects of African descent (blue) that were controls. (B) At each iteration, we added 10 OUD individuals of African descent to the cases and 10 controls of European descent to the controls. We estimated the model in the training data and used it to predict OUD status in the non-overlapping test set. (C) By the final iteration, we had a training and test set that was balanced by major geographic ancestry and OUD status.

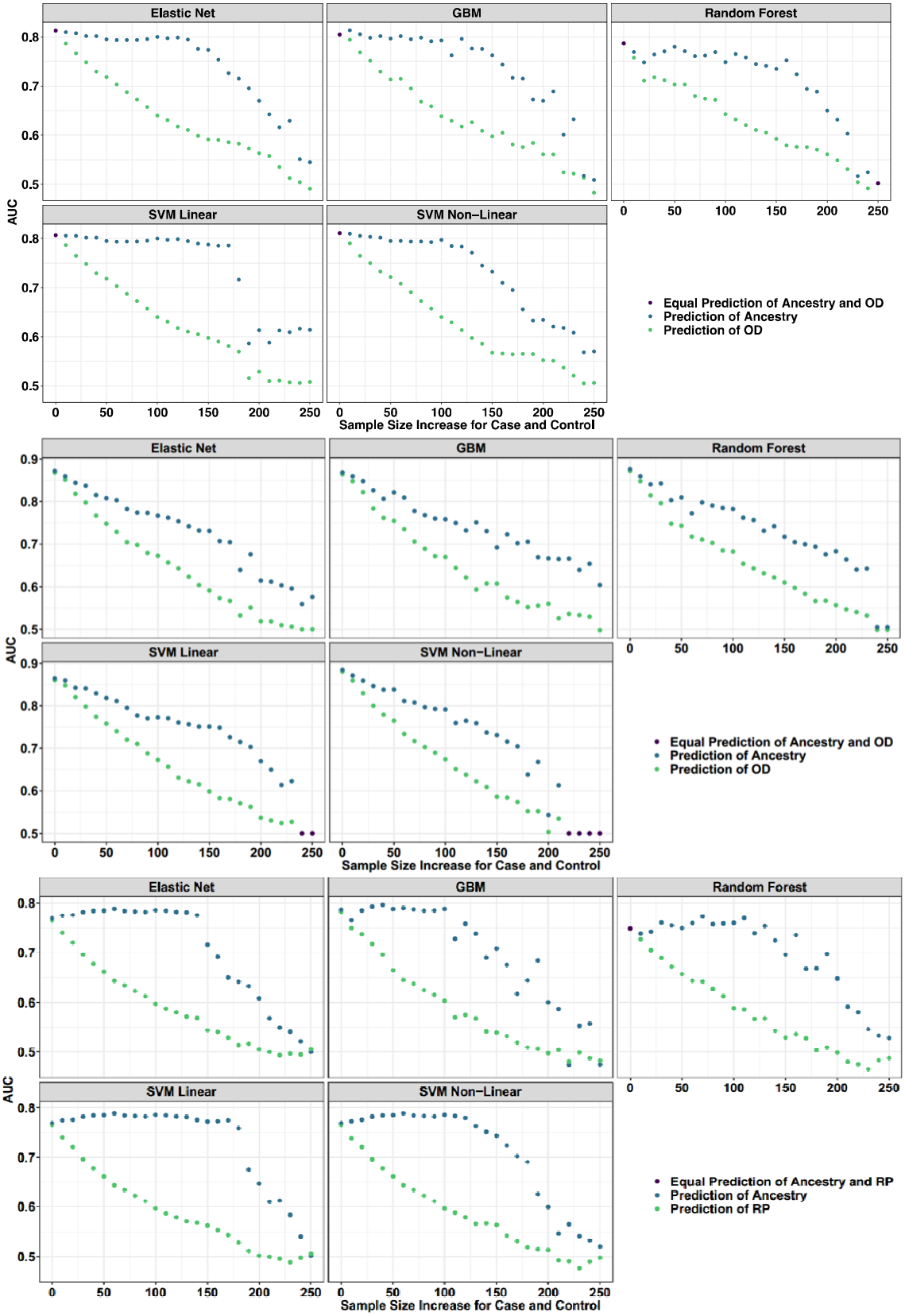

**Figure 3.** Learning curves from models trained to predict opioid dependence from 16 "reward-related" SNPs[3]. The curves are plotted by AUC based on their prediction of opioid dependence (orange) and geographic ancestry (blue) as the samples start from complete population confounding become more balanced 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). **(A)** *A priori* Candidate SNPs predicting Opioid Use Disorder. **(B)** Set 1 of randomly selected (MAF matched) SNPs predicting Opioid Use Disorder. **(C)** Set 1 of Random (MAF match) SNPs predicting a random phenotype binary phenotype. Across all perspectives, the prediction is entirely driven by major geographic ancestry.

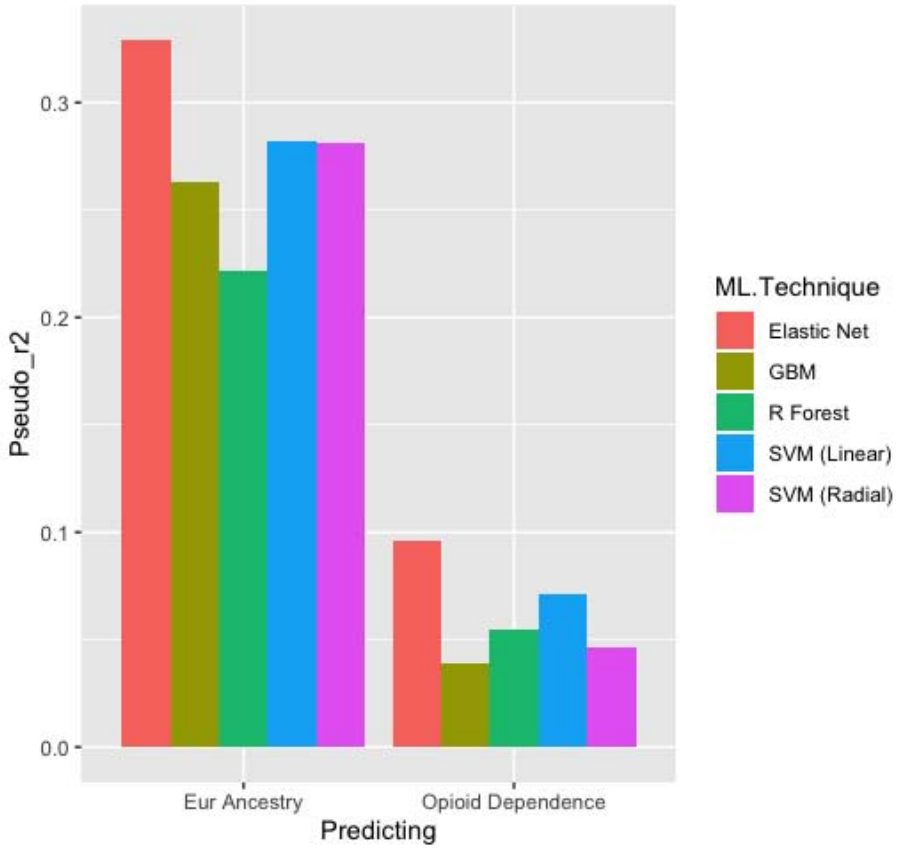


**Figure 4.** Bar plots of the pseudo $r^2$ from a logistic regression comparing the predictions of opioid dependence and percentage of European ancestry in a sample of 250 African American individuals from the Yale-Penn Test set. Pseudo $r^2$ was used instead of AUC because the percentage of European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