medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Ancestry May Confound Genetic Machine Learning: Candidate-Gene Prediction of Opioid Use Disorder as an Example

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.

Conflicts of interest: 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.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Abstract

Background. Machine learning (ML) models are beginning to proliferate in psychiatry, however machine learning models in psychiatric genetics have not always accounted for ancestry. Using an empirical example of a proposed genetic test for OUD, and exploring a similar test for tobacco dependence and a simulated binary phenotype, we show that genetic prediction using ML is vulnerable to ancestral confounding.

Methods. 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 rerun analyses with 8 random sets of allele-frequency matched SNPs. We contrast findings with 11 genome-wide significant variants for tobacco smoking. To document generalizability, we generate and test a random phenotype.

Results. None of the 5 ML algorithms predict OUD 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. Genome-wide significant tobacco smoking variants were also confounded by ancestry. Finally, random SNPs predicting a random simulated phenotype show that the bias attributable to ancestral confounding could impact any ML-based genetic prediction.

Conclusions. Researchers and clinicians are encouraged to be skeptical of claims of high prediction accuracy from ML-derived genetic algorithms for polygenic traits like addiction, particularly when using candidate variants.

Key Words: Opioid Use Disorder, Machine Learning, Algorithmic Bias, Ancestry, Candidate Genes

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

1.Introduction. Machine learning (ML) applications are increasingly used to leverage big data from electronic health records to classify patient populations (Ellis et al., 2019). In the realm of direct to consumer (DTC) and physician-guided 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 (Karavani et al., 2019). 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. Alongside the potential ethical challenges of such predictions (Hooker, 2021) lie the scientific limitations. The genetic "inputs" that are used by these tests typically comprise of "candidate gene variants" that are scored using pattern recognition software, powered with "artificial intelligence" frameworks, such as machine learning (ML). Most candidate variants have not borne out in unbiased genome-wide association studies (GWASs) (Border et al., 2019; Johnson et al., 2017). Yet, they tend to be popular as genetic markers of disease risk in commercial assays (e.g., Keri Donaldson, et al. 2017) . Exacerbating the problem of false positive candidate variant findings, past work on ML algorithms in psychiatric genetics has shown that these models may not systematically account for genetic ancestry (Bracher-Smith et al., 2020). As a consequence, patients and physicians are now confronted with products that may be unrelated to disease, but rather serve as panels of ancestry informative markers.

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 (Saha et al., 2016). Predictive tools that aim to identify at-risk

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

individuals for prevention and early intervention are being developed (Ellis et al., 2019), and because OUDs are moderately heritable (h^2 = 30-70%; Sun et al., 2012), 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. 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 (Border et al., 2019; Duncan and Keller, 2011; Johnson et al., 2017). 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 only one genome-wide significant variant, rs1799971, in the gene encoding the mu opioid receptor (*OPRM1*) (Zhou et al., 2020); 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%)(Zhou et al., 2020), putting a limit on overall predictive ability using common variants. Thus, we hypothesized that when ML algorithms utilize unsubstantiated candidate variants and do not properly account for population stratification, they produce predictions of disease outcome that are spurious and can cause mis-diagnosis.

In contrast to OUD, loci for tobacco phenotypes are numerous (Liu et al., 2019). Similar to OUD, the top genome-wide significant variants for tobacco smoking were in candidate genes (e.g., variants in *CHRNA5*, *CYP2A6*). These GWAS-validated candidate variants may have measurable impact on smoking cessation (Chen et al., 2020). Whether the selection of GWAS-validated candidate variants ameliorates challenges of ancestrally confounded ML prediction

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

remains unevaluated. To investigate whether genome-wide significant variants that may also have been candidate variants would overcome limitations of the OUD genetic test, we selected the 11 lead SNPs from the largest GWAS of cigarettes smoked per day for a comparison analysis (Liu et al., 2019). To further demonstrate the generalizability of ancestral confounding in ML, we simulated random genotypes and 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.

OUD and tobacco smoking are leading contributors to mortality – precision medicine efforts to preempt progression to drug misuse or intervene with tailored treatments will likely continue to incorporate genetic data. While a good predictor could aid physicians by providing them with additional information on which to base personalized treatment options, inaccurate predictive tests pose substantial hazards. For OUD in particular, the possible harms attributable to a false positive result include both the withholding of beneficial medication and 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. First, we test the prediction accuracy of ML models with candidate variants that comprise a commercial genetic test (previously Life Kit Predict from Prescient Medicine, now available from Solvd Health: https://solvdhealth.com/oud/). We examine accuracy as genetic ancestry is progressively accounted for. We also test the impact of mismatched ancestry as a predictor of degree of admixture in African Americans. Next, to test if a better choice of genetic variants ameliorates the problem, we select 11 variants implicated by the largest GWAS of cigarettes per day and use the same ML models to predict tobacco dependence, as assessed with the Fagerström Test for Nicotine Dependence. We also use random

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

sets of genetic variants (that have been matched by allele frequency to the original variants) to see how these randomly selected SNPs compare to the candidate SNPs across different subsamples for a simulated trait. We hypothesize that ML algorithms will preferentially identify ancestral confounding over traits of interest, regardless of the quality of the genetic variants.

2.Methods

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

We identified current consumer/physician-oriented genetic 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 Solvd Health (previously Life Kit Predict®), a physician guided genetic testing kit for OUD. This test was selected because (a) it purports to predict, with 97% accuracy, risk for OUD, a complex trait that has been shown to be highly polygenic and influenced by environment, (b) is accompanied by a training procedure that was published and therefore, can be recapitulated. While not completely transparent, this was the only test with enough information to allow us to evaluate the test, as described below.

The genetic component of Solvd Health's prediction algorithm relies on 15 or 16 candidate single nucleotide polymorphisms (SNPs) depending on the version of the test (Donaldson, et al. 2017), 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" (Border et al., 2019; Johnson et al., 2017). Only one of these SNPs (*OPRM1**rs1799971) has been shown by GWAS to affect OUD risk; the very small effect size of this variant (beta= -0.066, p=1.51e-

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

10) is unlikely to predict OUD risk to a great degree. With the exception of rs1799971, none of these candidate SNPs have been associated at genome-wide significant levels (p<5E-8) (**Supplemental Table 2**) with *any* complex trait in the GWAS Atlas (Watanabe et al., 2019) (**Supplemental Table 3**). However, 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 a *trait*. Accordingly, it was our expectation that sets of these markers would also necessarily be associated to population rather than trait, regardless of the sophistication of the interposed statistical methodology.

2.2 Sample description

We tested the approach using data from subjects recruited at five sites across the eastern United States as part of the Yale-Penn study of the genetics of substance dependence and comorbid psychiatric and behavioral phenotypes (McCarthy et al., 2016). All participants were interviewed with the Semi-Structured Assessment for Drug Dependence and Alcoholism (Consortium et al., 2016) 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. Funding agencies did not play any role in the study.

2.3 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

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

genotype call rate < 98% or minor allele frequency < 0.01 were removed before imputation. Imputation was performed using Minimac3 (McCarthy et al., 2016) 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 (Zhou et al., 2017).

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 (Auton et al., 2015) using EIGENSOFT (Patterson et al., 2006; Price et al., 2006a). 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 (Alexander et al., 2009). 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 substantially reduce CV error with additional K, we chose K=2 as appropriate for these data. The mean European ancestry for the African-Americans included in this study was 23.2%.

2.4 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. (2017) using PLINKv1.90 over a total sample of 5,057 individuals (3286 African American and 1768 European American). For comparison, in addition to the SNPs used by Donaldson et al. (2017) we identified 8 lists of random SNPs - each SNP from Donaldson et

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

al. was replaced by a random SNP with matched MAF in the 2 populations; all random SNPs were unique.

2.5 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 iterations of the learning curve (see **Figure 2**). The training and test sets were initially analyzed in a way that assured that 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**).

For our analysis, we chose 5 broad algorithm-generating methods as a survey of supervised ML, because they have been used by academic and commercial entities to attempt to predict OUD. All were implemented in the Caret package in R version 3.6.110: (1) (extreme) Gradient Boosted Machines (GBM), 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; (4) Random Forests (RF) to represent more complex tree structures; and (5) Elastic Nets (EN) for representation of (flexible) linear regression models. All models were trained with 10-fold cross validation and a hyperparameter grid-search in the training set. All Area Under the Curve (AUC)

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

and pseudo r^2 were extracted from the non-overlapping test set. Learning curves were plotted at all iterations.

2.6 Generalizability of confounding tested using Tobacco Dependence and a random binary phenotype

To examine the generalizability of this confounding beyond the test case of OUD prediction, we examined prediction of tobacco dependence, measured by the Fagerstrom Test of Nicotine Dependence (FTND). Cases were defined as scores ≥4 and controls with scores <4 (Heatherton, et al., 1991). Eleven genome-wide significant lead variants for cigarettes per day were selected from the Liu et al. (2019) cigarettes per day GWAS and the Quach et al. (2020) GWAS of FTND. We used both the cigarettes per day and the FTND GWAS because they are genetically correlated at .95 and share some of the same lead SNPs (Quach et al., 2020): rs3743078, rs16969968, rs56113850, rs58379124, rs215600, rs3025383, rs2072659, rs7951365, rs7431710, rs11725618, rs45568238. We limited our learning curves to 140 subjects in each ancestral group and case/control group (instead of 250), as this sample size allowed for ancestral balance needed for the learning curve approach (as there are only 140 control of European ancestry available). We also trained the models within the European American (EA) and African American (AA) subsamples. Data on 658 cases (AA = 302, EA = 356) and 342 controls (AA = 198, EA = 142) were used in training and testing sets, respectively.

To explore the problem further, 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 genetic ancestry, such that we ended with the same split in cases and controls by ancestry as was

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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.

3. Results

3.1 Evaluation of a modern OUD prediction 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. (Donaldson, et al. 2017), the basis for Solvd Health's OUD test kit, 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 apparently high (AUC > 0.8) when the sample was fully confounded (that is, when predictions were essentially predictions of genetic 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.

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

All iterations with 8 permutations of random SNPs matched on minor allele frequency to those in Donaldson et al. (2017) performed similarly to the published candidate SNPs (**Figure**

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

3B shows 1 permutation**,** the other 7 are shown as **Supplemental Figure 1**). Across all iterations of all permutations the ML models using random SNPs were apparently predictive of OUD when confounded by ancestry, with decreasing prediction as ancestral balance improved. They were better predictors of ancestry than OUD. Therefore, the candidate variants perform no better than randomly selected variants with the same ancestral allele frequencies.

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

As African Americans include substantial European admixture (Jordan et al., 2019) we examined whether the 16 OUD variants used by Solvd Health predict the extent of European (genetic) admixture within the African American cases and controls. We chose the 15th 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 European admixture in African-American individuals than of OUD (i.e., case status) (**Figure 4**).

3.4 Genome-wide significant variants for Nicotine Dependence show similar bias.

Using variants discovered in a large, un-biased GWAS of tobacco smoking (Liu et al., 2019; Quach et al., 2020) showed the same pattern: when models were confounded by ancestry, genetic prediction of nicotine dependence appeared to be accurate but the apparent accuracy fell to zero as we balanced ancestral groups within cases and controls **(Supplemental Figure 2)**. At all iterations, the model was better at predicting ancestry than nicotine dependence, even with SNPs that were genome-wide significant. Finally, we trained the prediction model within each ancestral population and found that AUC values did not exceed .504 in the European sample

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

(Supplemental Table 4), suggesting no prediction even when training within the discovery ancestral population.

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

We next took the 8 random sets of SNPs (above) 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 3**), 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.

4. Discussion

ML models trained either on a handful of selected candidate variants or across the whole genome are strongly sensitive to confounding by genetic ancestry (Polimanti et al., 2015). 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 ML models offer no evidence of predictive ability greater than chance. 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.

In the field of ML, our results fall under the study of "algorithmic bias" (Hooker, 2021). Examples in healthcare research outside genetics (Obermeyer et al., 2019) show that pattern

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

recognition with little understanding of underlying phenomena (e.g., social population stratification) may produce biased results. 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 (Price et al., 2006b) that is yet to be widely dealt with by ML in psychiatric genetics (Bracher-Smith et al., 2020).

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 (Li et al., 2011). 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. To cut across algorithms here, we show stratified analysis with learning curves. In particular, learning curves should be purposefully developed with ancestral 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, is critical to reduce confounding in genetic testing practices. However, restricting analyses to one continental ancestry (e.g., Europeans only) is not the solution to ancestrally confounded analyses. While that may attenuate gross confounding, we show that cryptic admixture remains an issue. Instead,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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 (Martin et al., 2017).

Even with appropriate adjustment for admixture, it is unlikely that candidate variants that are not substantiated by well-powered GWAS will produce any meaningful prediction of OUD. Recent meta-analyses of depression (Border et al., 2019), schizophrenia (Johnson et al., 2017), and executive function (Hatoum et al., 2019) overwhelmingly show that the vast majority of candidate variants in psychiatry do not rise to levels of genome-wide significance. An exception is addiction, where GWAS have recapitulated candidate gene findings. Unfortunately, as shown by our analyses of nicotine dependence, even these genome-wide significant variants fail to circumvent the issue of ancestral confounding within ML models. An alternative might involve incorporating information across the genome. For instance, polygenic risk scores (PRS) for tobacco smoking have produced promising findings, including in clinical settings (Chen et al., 2018). However, even PRS with appropriate control for ancestry in the largest samples to date offer limited clinical utility (Liu et al., 2019).

Several limitations are noteworthy. First, the original publication by Donaldson et al. (2017), did not provide detailed characteristics of the samples in which the algorithms were developed and tested, nor the specific procedures used. This is not atypical for "proprietary" commercial products, but made it challenging to fully approximate their analytic pipeline; hence we tested 5 different ML approaches. Standards for reporting such product development, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)(Collins et al., 2015), would allow researchers to better evaluate claims of genetic prediction, particularly using methods such as machine learning. Second, we focused on

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

one existing product and did not evaluate all possible methods for genetic prediction, for example, polygenic risk scores (PRS). As noted above, PRS offer an opportunity to study aggregated genome-wide genetic susceptibility, but have their own caveats (e.g., additivity; see: Bogdan et al., 2018; Dudbridge, 2013; Martin et al., 2017). This work focuses squarely on one approach for genetic prediction as it increasingly gains momentum in allied health fields – use of candidate genes and ML. Limitations notwithstanding, this work shows that using a handful of candidate variants in a ML framework that is naïve to genetic confounders is likely to produce biased prediction that misclassifies individuals, especially in ancestrally mixed samples.

5. Conclusions

Opioids are useful for pain management, but are also highly addictive. Against the backdrop of the opioid epidemic, the desire for tests that can provide insight into the likelihood of patients developing OUD is understandable, and DTC or physician-guided testing seems appealing. However, to avoid problems of under-treatment, that might disproportionately affect people of admixed ancestry, it is critical that any proposed test be fully vetted to ensure that it properly accounts for potential confounding by ancestry and accurately predicts the trait of interest.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Reference Cited

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. https://doi.org/10.1101/gr.094052.109

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Yingrui, Liu, S., Liu, Xiao, Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.L., Fulton, L., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C.L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A.,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Huang, Z., Coin, L.J.M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E.P., Kural, D., Lee, W.P., Leong, W.F., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Banks, E., Bhatia, G., Del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Rausch, T., Fritz, M.H., Stütz, A.M., Beal, K., Datta, A., Herrero, J., Ritchie, G.R.S., Zerbino, D., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E. V., Stenson, P.D., Barnes, B., Bauer, M., Cheetham, R.K., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Herwig, R., Ding, L., Koboldt, D.C., Larson, D., Ye, Kai, Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Willems, T.F., Simpson, J.T., Shriver, M.D., Rosenfeld, J.A., Bustamante, C.D., Montgomery, S.B., De La Vega, F.M., Byrnes, J.K., Carroll, A.W., DeGorter, M.K., Lacroute, P., Maples, B.K., Martin, A.R., Moreno-Estrada, A., Shringarpure, S.S., Zakharia, F., Halperin, E., Baran, Y., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F.C.L., Craig, D.W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A.A., Sinari, S.A., Squire, K., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, Kenny, Burchard, E.G., Hernandez, R.D., Gignoux, C.R., Haussler, D., Katzman, S.J., Kent, W.J., Howie, B., Ruiz-Linares, A., Dermitzakis, E.T., Devine, S.E., Kang, H.M., Kidd, J.M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Wing, M.K., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Yun, Shi, X., Quitadamo, A., Lunter, G., Marchini,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

J.L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D.K., Oleksyk, T.K., Fu, Yunxin, Liu, Xiaoming, Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Browning, B.L., Browning, S.R., Hormozdiari, F., Sudmant, P.H., Khurana, E., Tyler-Smith, C., Albers, C.A., Ayub, Q., Chen, Y., Colonna, V., Jostins, L., Walter, K., Xue, Y., Gerstein, M.B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Yao, Harmanci, A.O., Jin, M., Lee, D., Liu, J., Mu, X.J., Zhang, J., Zhang, Yan, Hartl, C., Shakir, K., Degenhardt, J., Meiers, S., Raeder, B., Casale, F.P., Stegle, O., Lameijer, E.W., Hall, I., Bafna, V., Michaelson, J., Gardner, E.J., Mills, R.E., Dayama, G., Chen, K., Fan, X., Chong, Z., Chen, T., Chaisson, M.J., Huddleston, J., Malig, M., Nelson, B.J., Parrish, N.F., Blackburne, B., Lindsay, S.J., Ning, Z., Zhang, Yujun, Lam, H., Sisu, C., Challis, D., Evani, U.S., Lu, J., Nagaswamy, U., Yu, J., Li, W., Habegger, L., Yu, H., Cunningham, F., Dunham, I., Lage, K., Jespersen, J.B., Horn, H., Kim, D., Desalle, R., Narechania, A., Sayres, M.A.W., Mendez, F.L., Poznik, G.D., Underhill, P.A., Mittelman, D., Banerjee, R., Cerezo, M., Fitzgerald, T.W., Louzada, S., Massaia, A., Yang, F., Kalra, D., Hale, W., Dan, X., Barnes, K.C., Beiswanger, C., Cai, H., Cao, H., Henn, B., Jones, D., Kaye, J.S., Kent, A., Kerasidou, A., Mathias, R., Ossorio, P.N., Parker, M., Rotimi, C.N., Royal, C.D., Sandoval, K., Su, Y., Tian, Z., Tishkoff, S., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J.C., Mathias, R.A., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Deng, X., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Joof, F.S., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Hien, T.T., Dunstan, S.J., ThuyHang, N., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Schieffelin, J., Grant, D.S.,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brooks, L.D., Felsenfeld, A.L., McEwen, J.E., Vaydylevich, Y., Duncanson, A., Dunn, M., Schloss, J.A., 2015. A global reference for human genetic variation. Nature. https://doi.org/10.1038/nature15393

Bogdan, R., Baranger, D.A.A., Agrawal, A., 2018. Polygenic Risk Scores in Clinical Psychology: Bridging Genomic Risk to Individual Differences. Annu. Rev. Clin. Psychol. 14. https://doi.org/10.1146/annurev-clinpsy-050817-084847

Border, R., Johnson, E.C., Evans, L.M., Smolen, A., Berley, N., Sullivan, P.F., Keller, M.C., 2019. 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. https://doi.org/10.1176/appi.ajp.2018.18070881

Bracher-Smith, M., Crawford, K., Escott-Price, V., 2020. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol. Psychiatry. https://doi.org/10.1038/s41380-020-0825-2

Chen, L.-S., Baker, T.B., Miller, J.P., Bray, M., Smock, N., Chen, J., Stoneking, F., Culverhouse, R.C., Saccone, N.L., Amos, C.I., Carney, R.M., Jorenby, D.E., Bierut, L.J., 2020. Genetic Variant in CHRNA5 and Response to Varenicline and Combination Nicotine Replacement in a Randomized Placebo-Controlled Trial. Clin. Pharmacol. Ther. 108, 1315–1325. https://doi.org/10.1002/CPT.1971

Chen, L.-S., Hartz, S.M., Baker, T.B., Ma, Y., Saccone, N.L., Bierut, L.J., 2018. Use of polygenic risk scores of nicotine metabolism in predicting smoking behaviors. https://doi.org/10.2217/pgs-2018-0081 19, 1383–1394. https://doi.org/10.2217/PGS-2018-0081

Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M., 2015. Transparent Reporting of a

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br. J. Surg. 102, 148–158. https://doi.org/10.1002/BJS.9736

Consortium, the H.R., McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L.J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato, C., Duijn, C.M. van, Gillies, C.E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J.C., Boomsma, D., Branham, K., Breen, G., Brummett, C.M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F.S., Corbin, L.J., Smith, G.D., Dedoussis, G., Dorr, M., Farmaki, A.-E., Ferrucci, L., Forer, L., Fraser, R.M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O.L., Hveem, K., Kretzler, M., Lee, J.C., McGue, M., Meitinger, T., Melzer, D., Min, J.L., Mohlke, K.L., Vincent, J.B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J.B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P.E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Berg, L.H. Van den, Rheenen, W. Van, Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M.G., Wilson, J.F., Frayling, T., Bakker, P.I.W. de, Swertz, M.A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C.A., Myers, R.M., Boehnke, M., McCarthy, M.I., Durbin, R., Abecasis, G., Marchini, J., 2016. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283. https://doi.org/10.1038/ng.3643

Donaldson, K., Laurence, D., Taylor, K., Lopez, J. and Chang, S. 2017. Multi-variant Genetic Panel for Genetic Risk of Opioid Addiction. Ann. Clin. Lab. Sci. 47, 452–456.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Dudbridge, F., 2013. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9, e1003348. https://doi.org/10.1371/journal.pgen.1003348

Duncan, L.E., Keller, M.C., 2011. A Critical Review of the First 10 Years of Candidate Gene-by-Environment Interaction Research in Psychiatry. Am. J. Psychiatry 168, 1041–1049. https://doi.org/10.1176/appi.ajp.2011.11020191

Ellis, R.J., Wang, Z., Genes, N., Ma'ayan, A., 2019. Predicting opioid dependence from electronic health records with machine learning. BioData Min. 12, 3. https://doi.org/10.1186/s13040-019-0193-0

Hatoum, A., Mitchell, E., Morrison, C.L., Evans, L., Keller, M., Friedman, N., 2019. 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. https://doi.org/10.1101/674515

Hooker, S., 2021. Moving beyond "algorithmic bias is a data problem." Patterns 2, 100241. https://doi.org/10.1016/J.PATTER.2021.100241

Johnson, E.C., Border, R., Melroy-Greif, W.E., de Leeuw, C.A., Ehringer, M.A., Keller, M.C., 2017. No Evidence That Schizophrenia Candidate Genes Are More Associated With Schizophrenia Than Noncandidate Genes. Biol. Psychiatry 82, 702–708. https://doi.org/10.1016/J.BIOPSYCH.2017.06.033

Jordan, I.K., Rishishwar, L., Conley, A.B., 2019. Native American admixture recapitulates population-specific migration and settlement of the continental United States. PLOS Genet. 15, e1008225. https://doi.org/10.1371/journal.pgen.1008225

Karavani, E., Zuk, O., Zeevi, D., Barzilai, N., Stefanis, N.C., Hatzimanolis, A., Smyrnis, N., Avramopoulos, D., Kruglyak, L., Atzmon, G., Lam, M., Lencz, T., Carmi, S., 2019.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Screening Human Embryos for Polygenic Traits Has Limited Utility. Cell 179, 1424-1435.e8. https://doi.org/10.1016/j.cell.2019.10.033

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Aguilar Salinas, C.A., Ahmad, T., Albert, C.M., Ardissino, D., Atzmon, G., Barnard, J., Beaugerie, L., Benjamin, E.J., Boehnke, M., Bonnycastle, L.L., Bottinger, E.P., Bowden, D.W., Bown, M.J., Chambers, J.C., Chan, J.C., Chasman, D., Cho, J., Chung, M.K., Cohen, B., Correa, A., Dabelea, D., Daly, M.J., Darbar, D., Duggirala, R., Dupuis, J., Ellinor, P.T., Elosua, R., Erdmann, J., Esko, T., Färkkilä, M., Florez, J., Franke, A., Getz, G., Glaser, B., Glatt, S.J., Goldstein, D., Gonzalez, C., Groop, L., Haiman, C., Hanis, C., Harms, M., Hiltunen, M., Holi, M.M., Hultman, C.M., Kallela, M., Kaprio, J., Kathiresan, S., Kim, B.J., Kim, Y.J., Kirov, G., Kooner, J., Koskinen, S., Krumholz, H.M., Kugathasan, S., Kwak, S.H., Laakso, M., Lehtimäki, T., Loos, R.J.F., Lubitz, S.A., Ma, R.C.W., MacArthur, D.G., Marrugat, J., Mattila, K.M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Meigs, J.B., Melander, O., Metspalu, A., Neale, B.M., Nilsson, P.M., O'Donovan, M.C., Ongur, D.,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Orozco, L., Owen, M.J., Palmer, C.N.A., Palotie, A., Park, K.S., Pato, C., Pulver, A.E., Rahman, N., Remes, A.M., Rioux, J.D., Ripatti, S., Roden, D.M., Saleheen, D., Salomaa, V., Samani, N.J., Scharf, J., Schunkert, H., Shoemaker, M.B., Sklar, P., Soininen, H., Sokol, H., Spector, T., Sullivan, P.F., Suvisaari, J., Tai, E.S., Teo, Y.Y., Tiinamaija, T., Tsuang, M., Turner, D., Tusie-Luna, T., Vartiainen, E., Ware, J.S., Watkins, H., Weersma, R.K., Wessman, M., Wilson, J.G., Xavier, R.J., Neale, B.M., Daly, M.J., MacArthur, D.G., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Li, L., Rakitsch, B., Borgwardt, K., 2011. ccSVM: Correcting Support Vector Machines for confounding factors in biological data classification. Bioinformatics 27, i342. https://doi.org/10.1093/bioinformatics/btr204

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., Zhan, X., Agee, M., Alipanahi, B., Auton, A., Bell, R.K., Bryc, K., Elson, S.L., Fontanillas, P., Furlotte, N.A., Hinds, D.A., Hromatka, B.S., Huber, K.E., Kleinman, A., Litterman, N.K., McIntyre, M.H., Mountain, J.L., Northover, C.A.M., Sathirapongsasuti, J.F., Sazonova, O. V., Shelton, J.F., Shringarpure, S., Tung, J.Y., Vacic, V., Wilson, C.H., Pitts, S.J., Mitchell, A., Skogholt, A.H., Winsvold, B.S., Sivertsen, B., Stordal, E., Morken, G., Kallestad, H., Heuch, I., Zwart, J.A., Fjukstad, K.K., Pedersen, L.M., Gabrielsen, M.E., Johnsen, M.B., Skrove, M., Indredavik, M.S., Drange, O.K., Bjerkeset, O., Børte, S., Stensland, S.Ø., Choquet, H., Docherty, A.R., Faul, J.D., Foerster, J.R., Fritsche, L.G., Gordon, S.D., Haessler, J., Hottenga, J.J., Huang, H., Jang, S.K., Jansen, P.R., Ling, Y., Mägi, R., Matoba, N., McMahon, G., Mulas, A., Orrù, V., Palviainen, T., Pandit, A., Reginsson, G.W., Smith, J.A., Taylor, A.E., Turman, C.,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Willemsen, G., Young, H., Young, K.A., Zajac, G.J.M., Zhao, W., Zhou, W., Bjornsdottir, G., Boardman, J.D., Boehnke, M., Boomsma, D.I., Chen, C., Cucca, F., Davies, G.E., Eaton, C.B., Ehringer, M.A., Esko, T., Fiorillo, E., Gillespie, N.A., Gudbjartsson, D.F., Haller, T., Harris, K.M., Heath, A.C., Hewitt, J.K., Hickie, I.B., Hokanson, J.E., Hopfer, C.J., Hunter, D.J., Iacono, W.G., Johnson, E.O., Kamatani, Y., Kardia, S.L.R., Keller, M.C., Kellis, M., Kooperberg, C., Kraft, P., Krauter, K.S., Laakso, M., Lind, P.A., Loukola, A., Lutz, S.M., Madden, P.A.F., Martin, N.G., McGue, M., McQueen, M.B., Medland, S.E., Metspalu, A., Mohlke, K.L., Nielsen, J.B., Okada, Y., Peters, U., Polderman, T.J.C., Posthuma, D., Reiner, A.P., Rice, J.P., Rimm, E., Rose, R.J., Runarsdottir, V., Stallings, M.C., Stančáková, A., Stefansson, H., Thai, K.K., Tindle, H.A., Tyrfingsson, T., Wall, T.L., Weir, D.R., Weisner, C., Whitfield, J.B., Yin, J., Zuccolo, L., Bierut, L.J., Hveem, K., Lee, J.J., Munafò, M.R., Saccone, N.L., Willer, C.J., Cornelis, M.C., David, S.P., Jorgenson, E., Kaprio, J., Stitzel, J.A., Stefansson, K., Thorgeirsson, T.E., Abecasis, G., Liu, D.J., Vrieze, S., 2019. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet. https://doi.org/10.1038/s41588-018-0307-5

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., Kenny, E.E., 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. 100, 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L.J., Zhang, H., Mahajan, A., Veldink, J., Peters, U., Pato,

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

C., Van Duijn, C.M., Gillies, C.E., Gandin, I., Mezzavilla, M., Gilly, A., Cocca, M., Traglia, M., Angius, A., Barrett, J.C., Boomsma, D., Branham, K., Breen, G., Brummett, C.M., Busonero, F., Campbell, H., Chan, A., Chen, S., Chew, E., Collins, F.S., Corbin, L.J., Smith, G.D., Dedoussis, G., Dorr, M., Farmaki, A.E., Ferrucci, L., Forer, L., Fraser, R.M., Gabriel, S., Levy, S., Groop, L., Harrison, T., Hattersley, A., Holmen, O.L., Hveem, K., Kretzler, M., Lee, J.C., McGue, M., Meitinger, T., Melzer, D., Min, J.L., Mohlke, K.L., Vincent, J.B., Nauck, M., Nickerson, D., Palotie, A., Pato, M., Pirastu, N., McInnis, M., Richards, J.B., Sala, C., Salomaa, V., Schlessinger, D., Schoenherr, S., Slagboom, P.E., Small, K., Spector, T., Stambolian, D., Tuke, M., Tuomilehto, J., Van Den Berg, L.H., Van Rheenen, W., Volker, U., Wijmenga, C., Toniolo, D., Zeggini, E., Gasparini, P., Sampson, M.G., Wilson, J.F., Frayling, T., De Bakker, P.I.W., Swertz, M.A., McCarroll, S., Kooperberg, C., Dekker, A., Altshuler, D., Willer, C., Iacono, W., Ripatti, S., Soranzo, N., Walter, K., Swaroop, A., Cucca, F., Anderson, C.A., Myers, R.M., Boehnke, M., McCarthy, M.I., Durbin, R., Abecasis, G., Marchini, J., 2016. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283. https://doi.org/10.1038/ng.3643

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science (80-.). 366, 447–453. https://doi.org/10.1126/science.aax2342

Patterson, N., Price, A.L., Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2, 2074–2093. https://doi.org/10.1371/journal.pgen.0020190

Polimanti, R., Yang, C., Zhao, H., Gelernter, J., 2015. Dissecting ancestry genomic background in substance dependence genome-wide association studies. Pharmacogenomics 16, 1487–1498. https://doi.org/10.2217/pgs.15.91

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006a. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909. https://doi.org/10.1038/ng1847

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006b. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909. https://doi.org/10.1038/ng1847

Quach, B.C., Bray, M.J., Gaddis, N.C., Liu, M., Palviainen, T., Minica, C.C., Zellers, S., Sherva, R., Aliev, F., Nothnagel, M., Young, K.A., Marks, J.A., Young, H., Carnes, M.U., Guo, Y., Waldrop, A., Sey, N.Y.A., Landi, M.T., McNeil, D.W., Drichel, D., Farrer, L.A., Markunas, C.A., Vink, J.M., Hottenga, J.J., Iacono, W.G., Kranzler, H.R., Saccone, N.L., Neale, M.C., Madden, P., Rietschel, M., Marazita, M.L., McGue, M., Won, H., Winterer, G., Grucza, R., Dick, D.M., Gelernter, J., Caporaso, N.E., Baker, T.B., Boomsma, D.I., Kaprio, J., Hokanson, J.E., Vrieze, S., Bierut, L.J., Johnson, E.O., Hancock, D.B., 2020. Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. Nat. Commun. 11, 1–13. https://doi.org/10.1038/s41467-020-19265-z

Saha, T.D., Kerridge, B.T., Goldstein, R.B., Chou, S.P., Zhang, H., Jung, J., Pickering, R.P., Ruan, W.J., Smith, S.M., Huang, B., Hasin, D.S., Grant, B.F., 2016. Nonmedical prescription opioid use and DSM-5 nonmedical prescription opioid use disorder in the United States. J. Clin. Psychiatry 77, 772–780. https://doi.org/10.4088/JCP.15m10386

Sun, J., Bi, J., Chan, G., Oslin, D., Farrer, L., Gelernter, J., Kranzler, H.R., 2012. Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors. Addict. Behav. 37, 1138–1144. https://doi.org/10.1016/j.addbeh.2012.05.010

TF, H., LT, K., RC, F., KO, F., 1991. The Fagerström Test for Nicotine Dependence: a revision

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

of the Fagerström Tolerance Questionnaire. Br. J. Addict. 86, 1119–1127. https://doi.org/10.1111/J.1360-0443.1991.TB01879.X

Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., Posthuma, D., 2019. A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. https://doi.org/10.1038/s41588-019-0481-0

Zhou, H., Polimanti, R., Yang, B.Z., Wang, Q., Han, S., Sherva, R., Nunez, Y.Z., Zhao, H., Farrer, L.A., Kranzler, H.R., Gelernter, J., 2017. Genetic risk variants associated with comorbid alcohol dependence and major depression. JAMA Psychiatry 74, 1234–1241. https://doi.org/10.1001/jamapsychiatry.2017.3275

Zhou, H., Zhou, H., Rentsch, C.T., Rentsch, C.T., Rentsch, C.T., Cheng, Z., Cheng, Z., Kember, R.L., Kember, R.L., Nunez, Y.Z., Nunez, Y.Z., Sherva, R.M., Tate, J.P., Tate, J.P., Dao, C., Dao, C., Xu, K., Xu, K., Polimanti, R., Polimanti, R., Farrer, L.A., Farrer, L.A., Farrer, L.A., Farrer, L.A., Farrer, L.A., Justice, A.C., Justice, A.C., Justice, A.C., Kranzler, H.R., Kranzler, H.R., Gelernter, J., Gelernter, J., Gelernter, J., Gelernter, J., 2020. Association of OPRM1 Functional Coding Variant with Opioid Use Disorder: A Genome-Wide Association Study. JAMA Psychiatry. https://doi.org/10.1001/jamapsychiatry.2020.1206

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Figure Legends

Figure 1. Population allele frequencies (from GNomad (Karczewski et al., 2020)) for the candidate alleles in Donaldson et al.(Donaldson et al., 2017) Solvd Health® (https://solvdhealth.com/oud/; accessed July 12th, 2021) 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 (Donaldson et al., 2017). 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

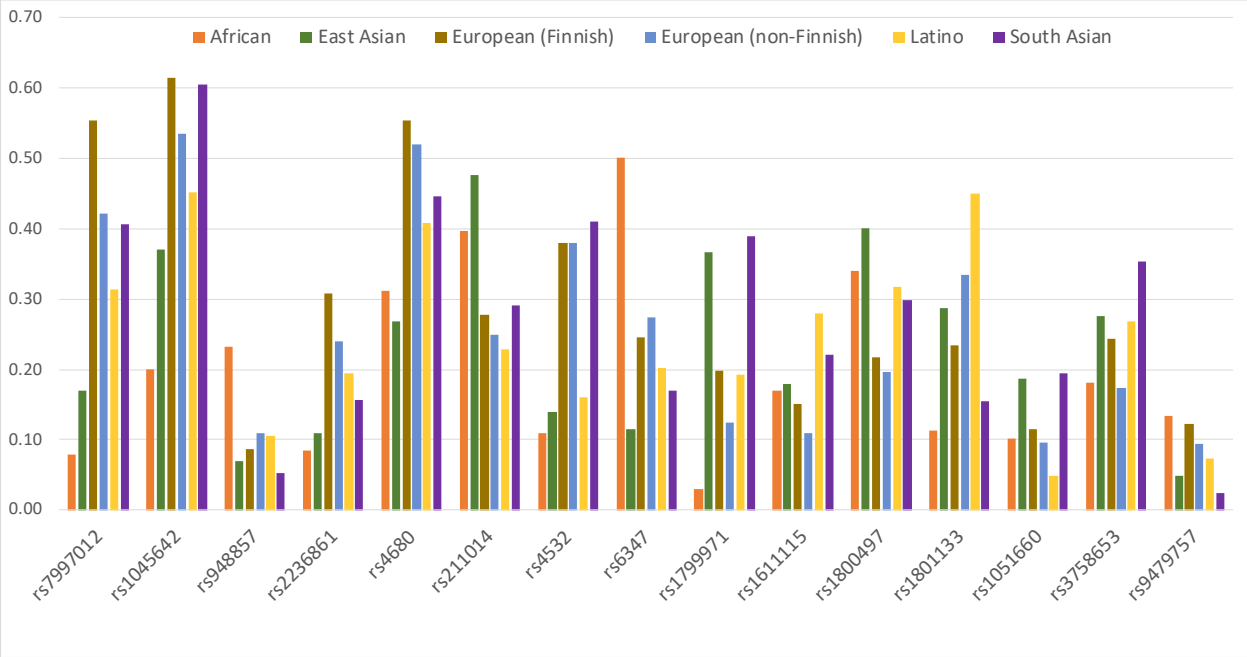


Figure 1. Population allele frequencies (from GNomad(Karczewski et al., 2020)) for the candidate alleles in Donaldson et al.(Keri Donaldson, Laurence Demers, Joe Lopez, 2017) LifeKit Predict® (https://prescientmedicine.com/technologies/lifekit-predict/; accessed September 8th, 2020) across different major geographic ethnic groups showing substantial variation in frequency across global populations.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

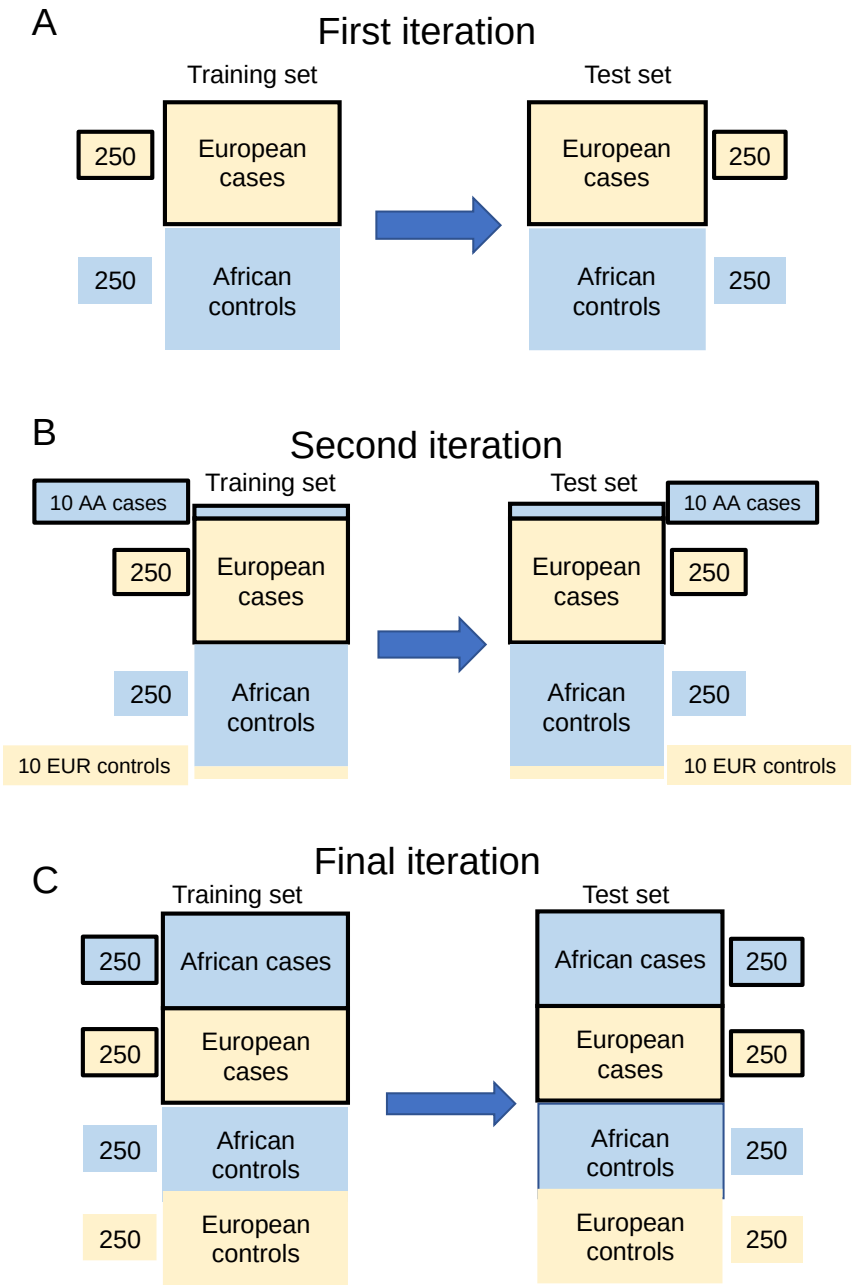
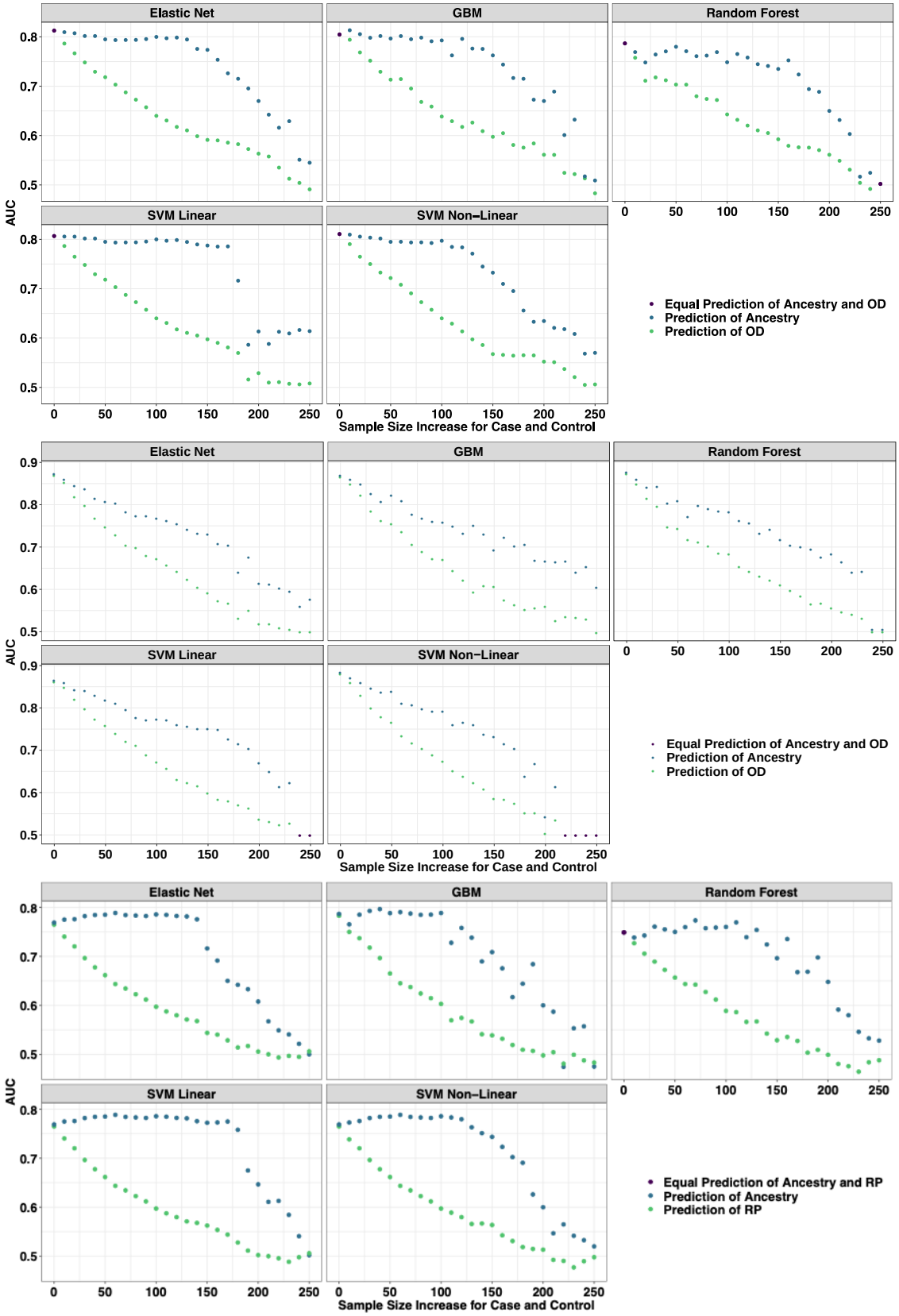


Figure 2. Learning curve procedure for all analyses. We had a complete and non-overlapping training and test set, each of 1000 subjects with 250 cases and controls of European and African descent. (A) For the first iteration we started with 250 subjects of European descent (tan) that were OUD cases, 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Figure 3. Learning curves from models trained to predict opioid dependence from 16 "reward-related" SNPs(Keri Donaldson, Laurence Demers, Joe Lopez, 2017). 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.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.12.20193342; this version posted August 6, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

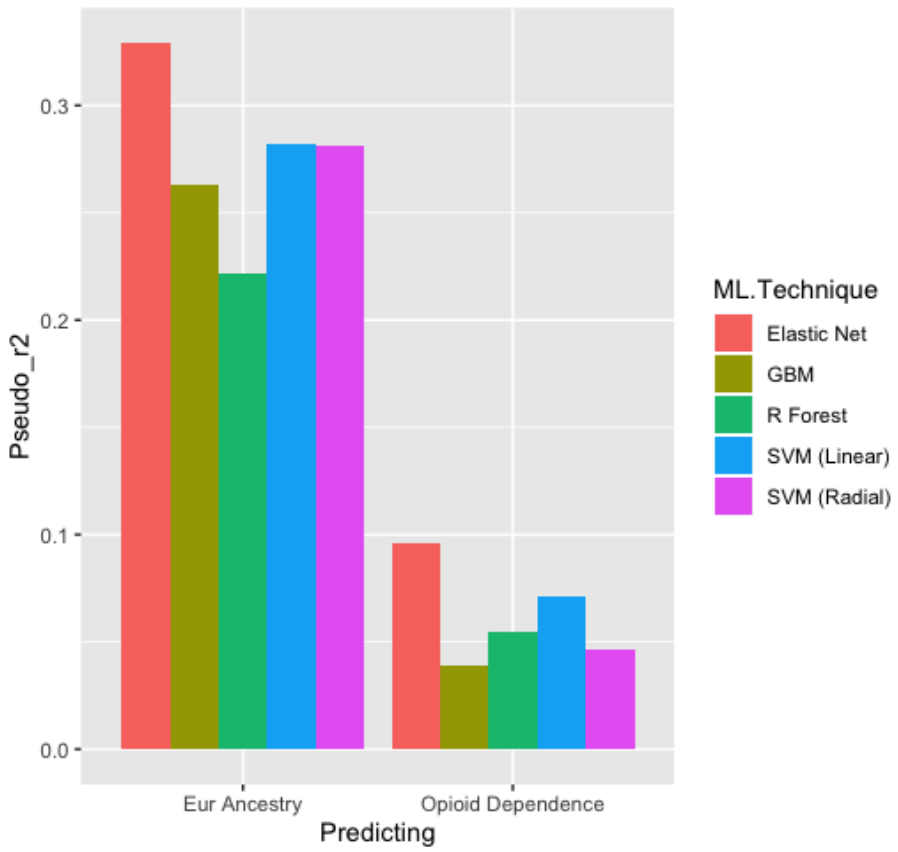


Figure 4. Bar plots of the pseudo r^2 from a logistic regression comparing the predictions of opioid dependence and percentage of European ancestry in a sample of 250 African American individuals from the Yale-Penn Test set. Pseudo r^2 was used instead of AUC because the percentage of European descent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