Phenome-wide Association Analysis of Substance Use Disorders in a Deeply Phenotyped Sample

Rachel L. Kember, Emily E. Hartwell, Heng Xu, James Rotenberg, Laura Almasy, Hang Zhou, Joel Gelernter, and Henry R. Kranzler

ABSTRACT

BACKGROUND: Substance use disorders (SUDs) are associated with a variety of co-occurring psychiatric disorders and other SUDs, which partly reflects genetic pleiotropy. Polygenic risk scores (PRSs) and phenome-wide association studies are useful in evaluating pleiotropic effects. However, the comparatively low prevalence of SUDs in population samples and the lack of detailed information available in electronic health records limit these data sets’ informativeness for such analyses.

METHODS: We used the deeply phenotyped Yale-Penn sample (n = 10,610 with genetic data; 46.3% African ancestry, 53.7% European ancestry) to examine pleiotropy for 4 major substance-related traits: alcohol use disorder, opioid use disorder, smoking initiation, and lifetime cannabis use. The sample includes both affected and control subjects interviewed using the Semi-Structured Assessment for Drug Dependence and Alcoholism, a comprehensive psychiatric interview.

RESULTS: In African ancestry individuals, PRS for alcohol use disorder, and in European individuals, PRS for alcohol use disorder, opioid use disorder, and smoking initiation were associated with their respective primary DSM diagnoses. These PRSs were also associated with additional phenotypes involving the same substance. Phenome-wide association study analyses of PRS in European individuals identified associations across multiple phenotypic domains, including phenotypes not commonly assessed in phenome-wide association study analyses, such as family environment and early childhood experiences.

CONCLUSIONS: Smaller, deeply phenotyped samples can complement large biobank genetic studies with limited phenotyping by providing greater phenotypic granularity. These efforts allow associations to be identified between specific features of disorders and genetic liability for SUDs, which help to inform our understanding of the pleiotropic pathways underlying them.

https://doi.org/10.1016/j.biopsych.2022.08.010

Individuals with substance use disorders (SUDs) are at an increased risk of comorbid psychiatric and medical disorders (1). However, the etiologic factors underlying comorbidity are not well understood. Large-scale genome-wide association studies (GWASs) have identified common risk markers for many SUDs (2–6) and established a pattern of genetic correlations among SUDs and between SUDs and other traits. This growing body of evidence suggests that there are common loci or biological pathways that contribute to the risk for multiple SUDs and psychiatric disorders. Identifying pleiotropic loci and pathways could provide insight into the etiologies of co-occurring disorders, advancing efforts to categorize, prevent, and treat SUDs and co-occurring medical and psychiatric conditions.

The large samples required to identify variants of generally small effect are often characterized by phenotypic information that is neither purpose collected nor detailed. This trade-off between sample size and depth of phenotyping limits clinically meaningful insights into disease biology (7). Furthermore, the selection of a phenotype for GWAS—if not limited by the phenotype data available—requires assumptions regarding the most representative or informative traits. Thus, phenotype scans (8) or phenotype-wide association studies (PheWASs) (9) complement GWASs by testing the phenotype in a hypothesis-free manner.

PheWASs have been most commonly implemented using data from electronic health records (EHRs), where ICD codes are converted to a simplified dataset that contains case-control status for more than 1800 diseases (9). A recent PheWAS of genetic liability for SUDs, represented by polygenic risk scores (PRSs), identified cross-trait associations across multiple phenotypic domains in EHR data (10). However, as with GWASs, PheWASs of EHR data are limited by their reliance on ICD diagnoses and minimal phenotyping. PheWASs have also been performed on data extracted from epidemiological studies or clinical trials (11,12), an approach that allows...
testing of subthreshold (with respect to diagnosis) and non-diagnosis-based phenotypes.

The Yale-Penn sample, recruited for genetic studies of SUDs, was deeply phenotyped using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA). This comprehensive psychiatric interview schedule assesses physical, psychosocial, and psychiatric manifestations of SUDs and co-occurring psychiatric disorders (13,14). It includes more than 3500 items representing demographic information, lifetime diagnostic criteria for DSM-IV (15) and DSM-5 (16) SUDs and DSM-IV (15) psychiatric disorders, psychosocial history, medical history, and a detailed substance use history. The SSADDA yields reliable diagnoses and criterion counts for SUDs and psychiatric disorders (13,14).

The detailed information available on the Yale–Penn sample enables insights into the shared genetic etiology of a variety of substance use and psychiatric traits. The dataset has been used to conduct GWASs (17–21), gene-by-environment studies (22,23), and phenotypic investigations (24,25). Because the Yale–Penn dataset includes nearly equal numbers of African (AFR) and European (EUR) ancestry individuals, analyses can be conducted in both population groups.

Here, we describe the selection of a subset of the data points collected with the SSADDA to create a PheWAS dataset based on the Yale–Penn sample. Using PRSs for alcohol use disorder (AUD), opioid use disorder (OUD), smoking initiation (SMK), and lifetime cannabis use (CAN), we demonstrated the utility of this dataset for evaluating SUD pleiotropy. Further, we used an interactive consensus process among study clinicians (EEH, JR, HRK), a data scientist (HX), and a geneticist (RLK) to reduce the number of variables to 689 for use in PheWASs. Variables that were considered informative for genetic studies and nonduplicative were retained, and the data were cleaned to ensure consistency across categories. Full details of variable selection and data cleaning are available in Supplemental Methods in Supplement 1.

Variable Selection and Data Cleaning

We used an interactive consensus process among study clinicians (EEH, JR, HRK), a data scientist (HX), and a geneticist (RLK) to reduce the number of variables to 689 for use in PheWASs. Variables that were considered informative for genetic studies and nonduplicative were retained, and the data were cleaned to ensure consistency across categories. Full details of variable selection and data cleaning are available in Supplemental Methods in Supplement 1.

Case and Control Definitions

Participants who met diagnostic criteria for a disorder were coded as cases and those who met no diagnostic criteria for the disorder were coded as controls. Subthreshold cases, e.g., those meeting at least one, but less than the required number of criteria for a diagnosis, were excluded from further analyses for that disorder. For symptom variables, participants who endorsed an individual symptom were considered cases and those who met no diagnostic criteria for that disorder were coded as controls. Subthreshold cases, e.g., those meeting at least one, but less than the required number of criteria for a diagnosis, were excluded from further analyses for that disorder. For symptom variables, participants who endorsed an individual symptom were considered cases and those who met no diagnostic criteria for that disorder were coded as controls. Unanswered items were coded as “NA,” and individuals were considered as neither a case nor a control for that phenotype.

Genotyping, Imputation, and PRSs

Yale-Penn samples were genotyped in 3 batches using the Illumina HumanOmni1-Quad microarray, the Illumina HumanCoreExome array, or the Illumina Multi-Ethnic Global array (Illumina, Inc.). Genotyping quality control has been described in detail previously (19,20,26). Genotype data were imputed using the Michigan Imputation Server (27) with the 1000 Genomes phase 3 reference panel (28). Further details are available in Supplemental Methods in Supplement 1.

PRSs were calculated for AUD (5), OUD (3), SMK (2), and CAN (4) using PRS–Continuous Shrinkage software (29) (Table S1 in Supplement 2). We used the PRS–Continuous
Shrinkage software “auto” option to estimate the parameters of shrinkage priors and fixed the random seed to 1 for replicable results. We matched available ancestry summary statistics in all analyses (e.g., an AFR GWAS for AUD was used to calculate AUD PRS in AFR Yale-Penn individuals).

**Statistical Analysis**

For the PhEASE, we fitted logistic regression models for binary traits and linear regression models for continuous traits, adjusting for sex, age, and the top 10 principal components within each genetic ancestry. Binary phenotypes with fewer than 100 cases or 100 controls and continuous phenotypes with fewer than 100 individuals within each ancestral group were excluded for that group. A Bonferroni correction was applied within each ancestral group to account for multiple testing (AFR \( p < 8.7 \times 10^{-5} \), EUR \( p < 7.9 \times 10^{-5} \)). To further examine the pleiotropic effects identified in PhEASE analyses, we conducted supplementary PhEASE for each PRS: 1) in cases for the corresponding SUD, 2) in controls for the corresponding SUD, and 3) covarying for the other SUD PRSs.

**RESULTS**

**Sample**

The PhEASE dataset comprises 689 variables in 25 phenotypic categories: 8 for substance use, 14 for psychiatric disorder, and 3 (demographics, environment, and medical) for other features. Table 1 shows the demographic and clinical features of the analytic sample and Table S2 in Supplement 2 shows the case counts for all diagnoses. The sample with genetic data available \( (n = 10,610) \) was 55.6% male (AFR: 54.9%, EUR: 56.2%) and included 4918 AFR participants (998 or 20.3% with no SUD diagnosis) and 5692 EUR participants (1570 or 24.1% with no SUD diagnosis). The mean number of SUD diagnoses in the sample was 2.44 (SD = 1.97) for DSM-IV and 2.28 (SD = 1.82) for DSM-5. We focused on individuals with ≥ 1 SUD diagnosis for DSM-IV alcohol dependence (AD); opioid dependence (OD); tobacco dependence (TD); or cannabis dependence; or DSM-5 AUD, OUD, or cannabis use disorder (CUD), comprising 3813 AFR (38.6% female) and 4294 EUR (37.1% female) individuals (average age = 40.1 years, SD = 11.0). There is a high degree of co-occurrence of SUD diagnoses in both population groups (Figure 1; Tables S3-S5 in Supplement 2).

**Primary Associations of SUD PRSs**

In AFR individuals, PRS for AUD (PRS\textsubscript{AUD}) was significantly associated with DSM-IV AD (odds ratio [OR] = 1.20, \( p = 7.0 \times 10^{-5} \), DSM-5 AUD diagnosis (OR = 1.21, \( p = 1.8 \times 10^{-5} \)), and DSM-5 AUD criterion count (\( \beta = 0.30, p = 5.4 \times 10^{-5} \)) (Figure 2A; Table S6 in Supplement 2). PRS for OUD (PRS\textsubscript{OUD}) and smoking initiation (PRS\textsubscript{SMMK}) were also significantly associated with either respective diagnosis or any other phenotypes (Figure 2A; Tables S7 and S8 in Supplement 2). We could not generate PRS for cannabis lifetime use (PRS\textsubscript{CAN}) in AFR individuals because the discovery data were limited to EUR ancestry.

In EUR individuals, PRS\textsubscript{AUD} and PRS\textsubscript{OUD} were significantly associated with their respective DSM-IV and DSM-5 diagnoses and DSM-5 criterion counts (PRS\textsubscript{AUD}: DSM-IV AD OR = 1.30, \( p = 1.3 \times 10^{-12} \), DSM-5 AUD OR = 1.29, \( p = 6.7 \times 10^{-13} \), DSM-5 criterion count \( \beta = 0.47, p = 2.3 \times 10^{-20} \); PRS\textsubscript{OUD}: DSM-IV OD OR = 1.28, \( p = 2.9 \times 10^{-16} \), DSM-5 OUD OR = 1.28, \( p = 3.1 \times 10^{-16} \), DSM-5 criterion count \( \beta = 0.49, p = 3.3 \times 10^{-15} \)) (Figure 2A; Tables S9 and S10 in Supplement 2). Similarly, PRS\textsubscript{SMMK} was significantly associated with the DSM-IV diagnosis of TD (OR = 1.67, \( p = 1.7 \times 10^{-49} \)) and the Fagerström Test for Nicotine Dependence score (\( \beta = 0.57, p = 7.3 \times 10^{-49} \)) (Figure 2A; Table S11 in Supplement 2). PRS\textsubscript{CAN} was only nominally associated with the respective criterion count (\( \beta = 0.16, p = 2.2 \times 10^{-7} \)) and DSM diagnoses (DSM-IV cannabis dependence: OR = 1.13, \( p = 5.0 \times 10^{-4} \); DSM-5 CUD: OR = 1.12, \( p = 3.0 \times 10^{-4} \)) (Figure 2A; Table S12 in Supplement 2).

**Associations With Phenotypes Involving the Same Substance**

PRSs associated with their respective substance use diagnosis were also associated with other phenotypes for the same substance. PRS\textsubscript{AUD} was associated with 11 alcohol phenotypes in AFR individuals (Figure 2B; Table S6 in Supplement 2) and 36 alcohol phenotypes in EUR individuals (Figure 2B; Table S9 in Supplement 2). In AFR, PRS\textsubscript{AUD} was associated with 3 of the 4 alcohol abuse criteria, including “continued use despite social/interpersonal problems” (OR = 1.20, \( p = 1.1 \times 10^{-7} \)), which was more significantly associated with PRS\textsubscript{AUD} compared with the diagnosis itself. PRS\textsubscript{AUD} was also significantly associated with frequent alcohol use, alcohol abuse, “ever had blackout,” and 2 of the 7 DSM-IV AD criteria (“unsuccessful efforts to decrease use” and “used more than intended”). In EUR, PRS\textsubscript{AUD} was associated with each AUD diagnostic criterion and with “sought treatment,” frequent use, age of first use, “ever had blackout,” 9 withdrawal symptoms (e.g., “depressed mood”), and 6 symptoms of heavy use (e.g., “depression”). Of these, 10 remained significant in a case-only analysis, including criterion count and “sought treatment.”

Among EUR individuals, PRS\textsubscript{OUD} was associated with 41 opioid phenotypes (Figure 2B; Table S10 in Supplement 2), including “time spent obtaining/using” (OR = 1.30, \( p = 5.34 \times 10^{-18} \)) and “ever used opioids” (OR = 1.28, \( p = 1.9 \times 10^{-15} \)—both associations more significant than with the diagnosis. PRS\textsubscript{OUD} was significantly associated with 10 OD and abuse criteria (“legal problems” being the exception [OR = 1.04, \( p = .44 \)].) PRS\textsubscript{OUD} was also significantly associated with “sought treatment”; frequent use; 4 symptoms of heavy use; and 16 withdrawal symptoms, the most significant being “depressed mood” (OR = 1.26, \( p = 3.7 \times 10^{-15} \)). None of these remained significant in a case-only analysis.

Among EUR individuals, PRS\textsubscript{SMMK} was associated with 25 tobacco phenotypes (Figure 2B; Table S11 in Supplement 2), including all 7 TD criteria. Following the top associations with the Fagerström Test for Nicotine Dependence score and the DSM-IV diagnosis of TD, the most significant association was “smoked over 100 cigarettes lifetime” (OR = 1.62, \( p = 8.6 \times 10^{-49} \)). Associations were also found with frequent tobacco use; “sought use”; “ever used tobacco”; age at first use; “health problems”; and 8 withdrawal symptoms, the most
significant being “irritability” ($OR = 1.44, p = 5.7 \times 10^{-33}$). In case-only analysis, no phenotypes survived correction.

Although only nominally associated with the diagnosis of CUD in EUR individuals, PRS$_{CAND}$ (based on a lifetime measure of cannabis use) was significantly associated with 3 other cannabis phenotypes (Figure 2B; Table S12 in Supplement 2): “ever used” ($OR = 1.20, p = 1.3 \times 10^{-8}$), “regularly use” ($OR = 1.15, p = 4.9 \times 10^{-8}$), and cannabis abuse ($OR = 1.14, p = 2.0 \times 10^{-5}$). None of these survived correction in the case-only analysis.

### Phenome-wide Analyses

The PheWAS of PRS in AFR individuals identified no significant associations that passed Bonferroni correction in other phenotypic domains (Figure 3; Tables S6–S8 in Supplement 2). However, we identified multiple significant associations across phenotypic domains in EUR individuals (Figure 3; Tables S9–S12 in Supplement 2), and for all 4 PRSs, the largest number were with other substance use phenotypes.

PRS$_{AUD}$ was associated with 126 phenotypes in 12 categories, including all 7 substance use categories, the most significant of which was DSM-IV TD ($OR = 1.35, p = 6.0 \times 10^{-15}$) and “ever used cocaine” ($OR = 1.34, p = 3.1 \times 10^{-19}$). Of these, 21 remained significant in a case-only analysis and 36 remained significant in the control-only analysis. The “ever used cocaine” phenotype was significant in the case-only analysis, but not in the control-only analysis. Both case-only and control-only analyses showed significant associations with DSM-IV TD and DSM-IV OD.

PRS$_{OLD}$ was associated with 76 phenotypes in 12 categories, including all 7 substance use categories, with the most significant being DSM-IV TD ($OR = 1.28, p = 6.0 \times 10^{-15}$), “sought treatment for cocaine use” ($OR = 1.25, p = 7.7 \times 10^{-14}$), and Fagerström Test for Nicotine Dependence score ($\beta = 0.27, p = 1.4 \times 10^{-15}$). Although none of these remained significant in the case-only analysis, in the control-only analysis, there were 5 significant substance use phenotypes, including DSM-IV TD.

PRS$_{SMK}$ was associated with 168 phenotypes in 15 categories, including all 7 substance use categories. The most significant substance use phenotype was “ever used cocaine” ($OR = 1.47, p = 5.1 \times 10^{-29}$), which remained significant in the control-only analysis. PRS$_{CAND}$ was associated with 23 phenotypes in 7 categories, though unlike the other PRSs, it was associated with phenotypes in only 4 of the 7 substance use categories. The most significant substance use phenotypes were “ever injected stimulants” ($OR = 1.19, p = 6.1 \times 10^{-8}$) and “ever used” stimulants (OR = 1.18, $p = 6.6 \times 10^{-8}$), hallucinogens (OR = 1.18, $p = 8.8 \times 10^{-8}$), or sedatives (OR = 1.16, $p = 6.5 \times 10^{-7}$). Both “ever injected” and “ever used” stimulants remained significant in the control-only, but not the case-only, analysis.

The psychiatric phenotype most significantly associated with PRS$_{AUD}$, PRS$_{OLD}$, and PRS$_{SMK}$ was “truancy, suspended or expelled from school” in the conduct disorder domain (PRS$_{AUD}$: OR = 1.27, $p = 7.7 \times 10^{-12}$; PRS$_{OLD}$: OR = 1.22, $p = 4.5 \times 10^{-10}$; PRS$_{SMK}$: OR = 1.44, $p = 7.9 \times 10^{-27}$). This finding is driven by the association in controls for each substance, with a stronger association in control-only analyses.
than case-only analyses. Both PRSAUD and PRSOUD were associated with multiple depression-related phenotypes, including the major depressive disorder criterion count (PRSAUD: $b = 0.24, p = 4.5 \times 10^{-5}$; PRSOUD: $b = 0.25, p = 5.3 \times 10^{-5}$). The second most significant phenotype for PRSSMK was the criterion count for PTSD ($b = 0.21, p = 3.1 \times 10^{-3}$), which was more significant in the control-only analysis ($b = 0.15, p = 8.9 \times 10^{-5}$) than the case-only analysis ($b = 0.12, p = 5.9 \times 10^{-3}$). PRSSMK was also associated with phenotypes for depression, ASPD, and attention-deficit/hyperactivity disorder.

PRSCAN was associated with 4 phenotypes in the ASPD domain, including the ASPD diagnosis (OR = 1.21, $p = 3.1 \times 10^{-2}$) and 2 in the conduct disorder domain: “stealing (without confrontation)” (OR = 1.18, $p = 3.4 \times 10^{-5}$) and “persistent pattern of behavior” (OR = 1.18, $p = 2.3 \times 10^{-5}$).

For PRSAUD, PRSOUD, and PRSSMK, the most significant association with a demographic phenotype was a negative association with educational attainment (PRSAUD: $b = -0.21, p = 3.2 \times 10^{-21}$; PRSOUD: $b = -0.16, p = 1.7 \times 10^{-14}$; PRSSMK: $b = -0.31, p = 3.8 \times 10^{-45}$), evident in both the case-only and control-only analyses for all PRSs except the PRSOUD case-only analysis. These 3 PRSs were also negatively associated with household income and positively associated with the number of outpatient psychiatric treatments. Both PRSAUD and PRSSMK were positively associated with childhood environmental variables, including “aware of household members using drugs or alcohol” and “frequent use of drugs/alcohol in household.” PRSAUD was also associated with phenotypes in the medical section, such as “alcohol used to intoxication” (OR = 1.05, $p = 2.5 \times 10^{-5}$) and “health rating” (higher score = poorer health; $b = 0.08, p = 2.4 \times 10^{-2}$). PRSMK (OR = 1.18, $p = 3.0 \times 10^{-3}$) and PRSCAN (OR = 1.13, $p = 3.0 \times 10^{-5}$) were positively associated with lifetime trauma, although these were not significant in case-only or control-only analyses. PRSOUD was associated with not having had a parent as the main caregiver (PRSOUD: OR = 0.84, $p = 5.5 \times 10^{-5}$).

In a supplementary analysis of PheWAS associations for each PRS, with other SUD PRSs as additional covariates (Tables S13–S19 in Supplement 2), the specificity of most PRSs for the corresponding substances increased. For instance, PRSAUD in EUR was associated with 162 phenotypes at a Bonferroni-corrected $p$ value, with 36 (22%) alcohol related. Covarying for the other PRSs yielded 36 significant associated phenotypes, of which 31 (86%) were alcohol related. Similar proportional increases were seen for PRSOUD (35% of associated phenotypes were opioid related compared with 81% when covarying for other SUD PRSs), with smaller...
proportional increases for PRS_{SMK} (13%–17%) and PRS_{CAN} (12%–17%).

**DISCUSSION**

PheWAS is valuable for exploring cross-trait associations of phenotypes with genetic liability for specific disorders, though to date, most PheWASs have used high-level EHR phenotypes typically limited to clinical diagnosis. Here, we describe a dataset for PheWAS derived from the Yale-Penn sample, ascertained using a detailed psychiatric interview whose administration included multiple quality control procedures (13). We selected features to reduce 3727 variables to 689 variables that are informative for genetic analysis. We refined cases and controls for each binary variable by applying methods commonly used in EHR PheWASs (9), removing subthreshold individuals who met ≥1 criteria, but not the full diagnosis. We identified novel phenotypic associations with PRSs for 4 substance use traits, particularly for subthreshold criteria or pertinent symptoms available only in a deeply phenotyped sample such as that derived from the Yale-Penn study.

The SSADDA is a useful assessment tool. Diagnoses made using a semi-structured interview following careful training procedures, with prespecified criteria and strict quality control methods, yield valid diagnoses (30) that likely are more accurate than those derived from EHR billing codes. As expected, many of the PRSs were associated with their respective primary diagnoses, supporting the validity of the approach. Likely due to the comparatively small GWAS discovery sample, PRSCAN was not associated with the primary diagnoses. Despite similar AFR and EUR target sample sizes, there were few associations for PRSs in the AFR sample, reflecting the lack of power for AFR in the parent GWASs.

The detailed information obtained with the SSADDA makes it possible to evaluate the impact of genetic risk for substance use traits on a variety of other substance-related traits not typically available in EHRs. In AFR, the criterion most strongly associated with PRS_{AUD} was "continued use despite social/interpersonal problems," which was the second strongest association in EUR. Although in DSM-IV, this is an alcohol abuse criterion, factor analysis has shown that it loads on the same factor as the 7 AD criteria, and in item response theory analysis, it is among those with the greatest information value (31).
The DSM-IV substance abuse criterion “legal problems” had the fewest significant associations with any of the respective PRSs. This is consistent with the results of twin and epidemiological studies in which the legal criterion has the lowest loading of the DSM-IV criteria and low discriminatory power (32,33), supporting its omission from DSM-5 (34). Our results also showed several associations with craving, a criterion that was added to DSM-5, which although not reported by all individuals with SUDs, in some studies is predictive of relapse (35,36) and thus a target of pharmacological and psychosocial treatment.
treatments (37–39). Furthermore, in our case-only analyses of PRS_{AUD}, many of the associations with phenotypes related to the primary substance remained, suggesting that greater PRS is associated with greater severity of the phenotype.

In addition to the association of PRSs with primary substance-related phenotypes, each PRS also showed multiple associations with other substance use traits. The high genetic correlation among substance use traits (2–6) suggests that this likely reflects true shared genetic effects. However, given the high levels of substance-related comorbidity in the Yale-Penn sample, the results likely also reflect phenotypic correlation and ascertainment bias. To address this, we ran a control-only analysis, which for many SUD diagnoses showed associations of the PRS for one substance (e.g., AUD) with diagnoses for other SUDs (e.g., DSM-IV TD) even among controls (e.g., without AD or AUD). This suggests that the association is not driven by phenotypic correlation. Whereas covarying the other SUD PRSs yielded greater specificity for the primary substance, at least some of the cross-trait findings are likely due to shared genetic etiology, with each PRS contributing additional substance-specific risk.

We replicated associations between genetic risk for SUDs and other traits, including psychiatric diagnoses. The association between AUD and major depressive disorder has been identified in a PheWAS of AUD PRS (10) and problematic alcohol use PRS (6) and in the analysis of genetic correlations between AUD and depression (6). Here, we dissect this by identifying associations of both PRS_{AUD} and PRS_{SUD} and specific features of major depression, including low mood and difficulty concentrating. The psychiatric phenotype most significantly associated with PRS_{AUD}, PRS_{SUD}, and PRS_{SMK} was “truancy, suspended or expelled from school” in the conduct disorder domain. Interestingly, in our control-only analysis, this association remains, again suggesting a potential direct association between PRS for substance use and this conduct disorder phenotype even when SUDs are absent. Consistent with phenotypic studies showing positive correlations between substance use and antisocial behaviors (40), we also observed significant associations for all 4 EUR PRSs with other criteria of ASPD and conduct disorder, such as shoplifting, fraud, and cheating. Previous studies have identified genetic correlations between PTSD and SUDs (41). Among EUR subjects, we observed a significant association of PRS_{SMK} with PTSD criteria, lifetime trauma assessment, and seeking treatment for PTSD. Similarly, PRS_{CAN} was significantly associated with lifetime trauma assessment. These relationships help to elucidate the features that underlie these common co-occurring symptoms and disorders.

In EUR individuals, both PRS_{AUD} and PRS_{SMK} were positively associated with childhood environmental variables reflecting substance use at home, whereas PRS_{SUD} was associated with not having a parent as the main caregiver, which capture aspects of a family history of substance use. However, although a family history of an SUD is associated with many substance use outcomes, it is not wholly overlapping with genetic risk, and the use of both sources of information can yield a fuller measure of risk (42). These findings raise an important theoretical question that cannot be answered with the data available here: namely, do associations of PRSs with features such as trauma, truancy, education, and parental substance use reflect intergenerational effects (i.e., “genetic nurture”) or do they refute typical assumptions that the genetic and environmental components in gene-by-environment interactions are uncorrelated (i.e., adverse environments are evenly distributed across the range of PRS)?

A limitation of the Yale-Penn sample is its comparatively small size given the resource-intensive recruitment and ascertainment activities. Thus, we believe that the use of deeply phenotyped samples is complementary to that of biobank data, which are more amenable to gene discovery. Unlike EHR-based genetic studies, our study is cross-sectional and therefore lacks a longitudinal perspective. We were able to calculate PRSs and conduct a PheWAS in both AFR and EUR ancestry individuals by selecting the majority of the GWASs from the Million Veteran Program, a large and diverse biobank. However, the Million Veteran Program comprises veterans who are predominantly male and older and who have high rates of medical comorbidity and thus differ from the target sample. Our prioritization of large GWASs also led to differences in the phenotypes selected—2 for SUDs (AUD and OUD) and 2 for substance use (SMK and CAN). Because SUDs and substance use have related but distinct genetic etiologies (43), this may have led to differences in associations between the PRSs.

GWASs of SUDs and related traits are limited by a lack of deep phenotyping, a trade-off with the large samples needed to provide adequate statistical power to identify common variants of small effect. This study relied on a carefully constructed diagnostic interview that enabled us to conduct analyses of both primary substance use traits, results of which validated the effort, and a wealth of phenotypic data not captured in EHR-based biobanks (e.g., individual diagnostic criteria and symptoms, age of onset, and environmental variables). The continued growth of ancestrally diverse biobanks will provide opportunities to compare the performance of PRSs with those reported here. Additional enriched, deeply phenotyped samples are needed to support such efforts.

ACKNOWLEDGMENTS AND DISCLOSURES

This study was supported by National Institutes of Health grants (Grant Nos. AA028292 [to RLK], DA046345 [to HRK], and AA028364 [to JG]) and the Veterans Integrated Service Network 4 Mental Illness Research, Education and Clinical Center. A previous version of this article was published as a preprint on medRxiv: https://www.medrxiv.org/content/10.1101/2022.02.09.22270737v1. HRK is a member of advisory boards for Dicerna Pharmaceuticals, Sophrosyne Pharmaceuticals, and Enthorn Pharmaceuticals; a consultant to Sobrena Pharmaceuticals; the recipient of research funding and medication supplies for an investigator-initiated study from Alkermes; and a member of the American Society of Clinical Psychopharmacology’s Alcohol Clinical Trials Initiative, which was supported in the past 3 years by Alkermes, Dicerna, Ethypharm, Lundbeck, Mitsubishi, and Otsuka. JG and HRK hold U.S. patent 10,900,082 titled: “Genotype-guided dosing of opioid agonists,” issued January 26, 2021. All other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania (RLK, EEH, HK, HRK); Mental Illness Research, Education and Clinical Center, Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania (RLK, EEH, JR, HKR); Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania (LA); Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia,


42. Schaefer JD, Jang S-K, Clark DA, Deak JD, Hicks BM, Iacono WG, et al. (2021): Associations between polygenic risk of substance use and use disorder and alcohol, cannabis, and nicotine use in adolescence and young adulthood in a longitudinal twin study. Psychol Med 1–11.