Predicting Substance Use Disorders: A Multifactorial Risk Index Combining Clinical, Environmental, and Genetic Risk Factors

Peter B. Barr 1,2, Morgan N. Driver 3, Sally I-Chun Kuo 4, Mallory Stephenson 5, Fazil Aliev 5,6, Richard Karlsson Linnér 7, Andrey P. Anokhin 8, Kathleen Bucholz 8, Grace Chan 9,10, Howard J. Edenberg 11,12, Alexis C. Edwards 5, Meredith W. Francis 8, K. Paige Harden 13,14, Chella Kamarajan 1, Jaakko Kaprio 15, Sivan Kinreich 1, John Kramer 10, Samuel Kuperman 10, Antti Latvala 16, Jacquelyn L. Meyers 1,2, Abraham A. Palmer 17,18, Martin H. Plawecki 19, Bernice Porjesz 1, Richard J. Rose 20, Marc A. Schuckit 17, Jessica E. Salvatore 4, and Danielle M. Dick 3,6

1 Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University, Brooklyn, NY, USA.
2 VA New York Harbor Healthcare System, Brooklyn, NY, USA.
3 Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA.
4 Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, Piscataway, NJ, USA.
5 Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA.
6 Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA.
7 Department of Economics, Leiden University, Leiden, The Netherlands.
8 Department of Psychiatry, School of Medicine, Washington University in St. Louis, St Louis, MO, USA.
9 Department of Psychiatry, School of Medicine, University of Connecticut, Farmington, CT, USA.
10 Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, IA, USA.
11 Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, IN, USA.
12 Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University, Indianapolis, IN, USA.
13 Department of Psychology, University of Texas at Austin, Austin, TX, USA.
14 Population Research Center, University of Texas at Austin, Austin, TX, USA.
15 Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland.
16 Institute of Criminology and Legal Policy, University of Helsinki, Helsinki, Finland.
17 Department of Psychiatry, University of California San Diego, La Jolla, CA, USA.
18 Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA.
19 Department of Psychiatry, School of Medicine, Indiana University, Indianapolis, IN, USA.
20 Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA.

Corresponding author:
Peter B. Barr
Department of Psychiatry and Behavioral Sciences
SUNY Downstate Health Sciences University
450 Clarkson Ave, MSC 1203, Brooklyn, NY 11203
e-mail: peter.barr@downstate.edu.

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

**Importance:** Substance use disorders (SUDs) incur serious social and personal costs. Screening techniques that identify persons at risk before problems develop can improve prevention efforts.

**Objective:** To examine whether models that include polygenic scores (PGS) and a clinical/environmental/risk index (CERI) are able to identify individuals as having a lifetime SUD.

**Design:** We tested the predictive power of PGS and CERI for lifetime diagnosis of DSM-IV substance dependence using four longitudinal cohorts.

**Setting:** The study included four samples: 1) the National Longitudinal Study of Adolescent to Adult Health (Add Health); 2) the Avon Longitudinal Study of Parents and Children (ALSPAC); 3) the Collaborative Study on the Genetics of Alcoholism (COGA); and 4) the Finnish Twin Cohort Study (FinnTwin12) for a combined sample of N = 15,134.

**Participants:** Participants in Add Health (N_{EUR} = 4,855; N_{AFR} = 1,605) and COGA (N_{EUR} = 1,878; N_{AFR} = 870) included individuals of both European and African ancestries. Participants in ALSPAC (N_{EUR} = 4,733) and FinnTwin12 (N_{EUR} = 1,193) were limited to individuals of European ancestries.

**Exposures:** A clinical/environmental risk index (CERI) composed of ten items and PGS for phenotypes with strong genetic overlap with SUDs (drinks per week, problematic alcohol use, externalizing problems, major depressive disorder, and schizophrenia).

**Main Outcomes:** Meeting lifetime criteria for DSM-IV: 1) alcohol dependence, 2) drug dependence, and 3) any substance dependence (alcohol, other drug, or nicotine).

**Results:** In the models containing the five PGS and CERI, the CERI was associated with all three outcomes (ORs = 1.35 – 1.64). PGS for problematic alcohol use was associated with alcohol dependence (OR = 1.14), PGS for externalizing was associated with drug dependence (OR = 1.14) and both were associated with any substance dependence (ORs = 1.11 – 1.19).
Including the five PGS, CERI, and covariates explained 6% - 13% of the variance in SUDs. Those in the top 10% of CERI and PGS had relative risk ratios of 3.82 - 9.13 for each SUD relative to the bottom 90%.

**Conclusions and Relevance:** Measures of clinical, environmental, and genetic, risk demonstrate modest ability to distinguish between affected and unaffected individuals for alcohol, drug, and any substance use disorders in young adulthood. These tools will continue to advance as we identify additional risk factors that can be incorporated into clinical practice and deliver on the goal of precision medicine.
INTRODUCTION

Substance use disorders (SUD) are associated with substantial costs to affected individuals, their families, and society. Approximately 73,000 Americans died as the result of an opioid overdose in the 12 months preceding May of 2021¹. In 2016, alcohol use contributed 4.2% to the global disease burden and other drug use contributed 1.3%². Excessive alcohol use and illicit drug use cost the United States an annual $250 billion³ and $190 billion⁴ respectively. Given the substantial human and economic costs of misuse and disorders, developing methods for identifying persons at heightened risk for SUD is a major public health concern.

There is evidence that targeting those at high risk is an effective strategy for interventions in substance misuse⁵. Ideally, screening tools for SUD risk would include measures of social, clinical, and genetic risk factors, each of which impacts the development of substance use disorders⁶–¹⁰. Prior research using an index of clinical and environmental risk factors (e.g. childhood disadvantage, family history of SUD, childhood conduct problems, childhood depression, early exposure to substances, frequent use during adolescence) found this risk index useful in identifying those with persistent SUDs¹¹. Recent analyses evaluating the potential for polygenic scores (PGS), which aggregate risk for a trait across the genome using information from genome-wide association studies (GWAS), show current PGS do poorly in identifying individuals affected by SUDs¹². Using individual genetic variants and clinical features outperformed clinical features alone¹³, but individual variants have limited predictive power. Overall, there is limited work on the combined genetic, environmental, and clinical risk factors for SUDs. For other medical conditions, such as melanoma¹⁴ or ischemic stroke¹⁵, combining clinical and genetic risk factors showed improvement over models using individual risk factors.

In the current study, we examine the joint association of early life clinical/environmental risk factors and polygenic scores with SUDs in early adulthood across four longitudinal cohorts: the National Longitudinal Study of Adolescent to Adult Health (Add Health); the Avon
Longitudinal Study of Parents and Children (ALSPAC); the Collaborative Study on the Genetics of Alcoholism (COGA); and the youngest cohort of the Finnish Twin Cohort Study (FinnTwin12). These samples include population-based cohorts from three countries (United States, England, and Finland) and a predominantly high-risk sample. Two of the samples (COGA and Add Health) are ancestrally diverse. We focus on early adulthood as this is a critical period for the development and onset of SUDs\textsuperscript{16}. Our research questions are guided by the understanding that risk factors for SUDs range from broader social conditions to genetic influences, and we must acknowledge all of these influences if we hope to deliver on the goal of an equitable precision medicine.

METHODS

Samples

Add Health is a nationally-representative longitudinal study of adolescents followed into adulthood in the United States\textsuperscript{17}. Data has been collected from Wave I when respondents were between 11-18 (1994-1995) to Wave V (2016-2018) when respondents were 35-42. The current analysis uses data from Waves I, II, and Wave IV.

ALSPAC is an ongoing, longitudinal population-based study of a birth cohort in the (former) Avon district of South West England\textsuperscript{18–21}. Pregnant female residents with an expected date of delivery between April 1, 1991, and December 31, 1992, were invited to participate (N = 14,541 pregnant women, 80\% of those eligible). This analysis uses data up to the age 22 assessment (details of all the data that is available through a searchable, web-based tool: http://www.bristol.ac.uk/alspac/researchers/our-data/).

COGA is a family-based sample consisting of alcohol dependent individuals (identified through treatment centers across the United States), their extended families, and community controls (N \sim 16,000). We use a prospective sample of offspring of the original COGA
participants (baseline ages 12-22, N = 3,573) and have been assessed biennially since recruitment (2004-2019).\(^2\)

FinnTwin12 is a population-based study of Finnish twins born 1983–1987 identified through Finland’s Central Population Registry. A total of 2,705 families (87% of all identified) returned the initial family questionnaire late in the year in which twins reached age 11.\(^2\) Twins were invited to participate in follow-up surveys when they were ages 14, 17, and approximately 22. Each cohort includes a wide range of social, behavioral, and phenotypic data measured across the life course. The measures of SUDs were derived from the corresponding young adult phases of data collection in each cohort (mean ages ~ 22 - 28). A full description of each sample is presented in the supplementary information.

**Measures**

*Lifetime Diagnosis of Substance Use Disorder.*

We constructed measures of lifetime SUD diagnosis based on the data that were available in each of the samples, defined as meeting criteria for three, non-mutually exclusive categories of substance dependence: 1) alcohol dependence; 2) drug dependence (inclusive of cannabis, cocaine, opioids, sedatives, and other substances); and 3) any substance dependence (alcohol, nicotine, or illicit drug). Our analyses focused on DSM-IV as this diagnostic system was most consistently used across all samples. There were two exceptions: (1) The ALSPAC included measures of likely alcohol dependence based on scores from the Alcohol Use Disorder Identification Test (AUDIT); and (2) for all samples, nicotine dependence was measured using a cutoff of 7 or higher on the Fagerstrom Test for Nicotine Dependence.\(^2\) Where possible, we drew measures of substance dependence from the young adult waves of data collection to try and maintain temporal ordering between SUD diagnoses and measured risk factors.
Clinical/Environmental Risk Index

We created a clinical/environmental risk index (CERI) considering a variety of established risk factors for SUD (Table 1). The CERI included ten validated early life risk factors associated with later development of SUDs, including: low childhood socioeconomic status (SES), family history of SUD, early initiation of substance use, childhood internalizing problems, childhood externalizing problems, frequent drinking in adolescence, frequent smoking in adolescence, frequent cannabis use in adolescence, peer substance use, and exposure to trauma/traumatic experiences\(^\text{11,25,26}\). We dichotomized each risk factor (present vs not present) and summed them into an index for each person ranging from 0 to 10, providing a single measure of aggregate risk. A full list of how each measure is defined within each of the samples is available in the supplementary information.

Polygenic Scores

We constructed polygenic scores (PGS), which are aggregate measures of the number of risk alleles individuals carry weighted by effect sizes from GWAS summary statistics. We derived PGS from five recent GWAS of SUDs and comorbid conditions including: 1) externalizing problems (EXT)\(^\text{27}\); 2) major depressive disorder (MDD)\(^\text{28}\); 3) problematic alcohol use\(^\text{29}\) (ALCP); 4) alcohol consumption (drinks per week, ALCC)\(^\text{30}\); and 5) schizophrenia (SCZ)\(^\text{31}\). We focused on these PGS because: 1) SUD show strong genetic overlap with other externalizing\(^\text{32–34}\), internalizing\(^\text{28,35}\), and psychotic disorders\(^\text{29,36,37}\); 2) both shared and substance-specific genetic risk are associated with later SUDs\(^\text{38–40}\); and 3) substance use and SUDs have only partial genetic overlap\(^\text{41,42}\).

To date, GWAS have been overwhelmingly limited to individuals of European ancestries\(^\text{43,44}\). Importantly, polygenic scores derived from GWAS of one ancestry do not always transport into other ancestral populations\(^\text{45,46}\). We therefore used PRS-CSx\(^\text{47}\), a new method that combines information from well powered GWAS (typically of European ancestries) and
Running Header: Predicting SUDs in Young Adulthood

ancestrally matched GWAS to improve predictive power of polygenic scores in the African ancestry samples from Add Health and COGA. PRS-CSx integrates GWAS summary statistics across multiple input populations and employs a Bayesian approach to correct GWAS summary statistics for the non-independence of SNPs in linkage disequilibrium (LD) with one another. See the supplementary information for a detailed description.

**Analytic Strategy**

We pooled all the data for analysis using an integrative data analytic approach. In order to account for population stratification in the polygenic scores, we first regressed each PGS on age, age², sex, sex*age, sex*age², and the first 10 ancestral PC’s and saved the standardized residuals (Z-scores) as our PGS. Next, we pooled all of the data, including cohort as a fixed effect for each of the six cohorts (4 samples, of which two were split by ancestry) in subsequent analyses. Age of last observation and sex were also included as covariates. Finally, because COGA and FT12 included a large number of related individuals, we adjusted for familial clustering using cluster-robust standard errors.

We estimated a series of nested logistic regression models with the pooled data: 1) a baseline model (sex, age, and cohort), 2) a genetic risk model (baseline + PGS), 3) a clinical/environmental risk model (baseline + CERI), and 4) a combined risk model (baseline + PGS + CERI) to assess the predictive accuracy of each model using the difference in pseudo-\( \Delta \text{Pseudo-}R^2 \) \textsuperscript{50}, between the baseline and corresponding models. We also calculated the discriminatory power of the combined model using the area under the curve (AUC) from a receiver operating characteristic (ROC) curve. Our analytic strategy was preregistered on the Open Science Framework (https://osf.io/etbw8). Deviations from the preregistration are described and outlined in the supplementary information.
RESULTS

Table 2 presents the descriptive statistics and sample sizes across each of the cohorts and ancestries. All of the cohorts had similar proportions of females (~51% - 56%). The mean ages ranged from ~22 to ~29 years of age. The COGA cohorts (both European and African ancestries) reported the highest rates of SUD. These elevated rates are expected given the nature of the COGA sample (highly selected for substance use disorders). Add Health participants generally had higher rates of SUD than ALSPAC or FinnTwin12, but lower than COGA. Finally, ALSPAC and FinnTwin12 reported similar levels of alcohol, drug, and any substance dependence. COGA participants reported higher mean values on the CERI. The remainder of the cohorts report relatively similar rates of exposure to risk factors.

Table 3 presents the results from the PGS only, CERI only, and combined models for each outcome. Two of the five PGS were associated with the SUD outcomes in the PGS only model. PGS for externalizing (EXT OR = 1.19 – 1.35) and problematic alcohol use (ALCP OR = 1.08 – 1.15) were both associated with all SUD outcomes, though the association between ALCP and drug dependence was not significant after correcting for multiple testing. In the CERI only models, the CERI was consistently associated across each of the categories of SUD (ORs = 1.37 – 1.67). When we combined the PGS and risk index into the same model, the CERI remained significant across SUDs and was largely unchanged in magnitude (ORs = 1.35 – 1.64). EXT remained associated with drug dependence (OR = 1.14), ALCP remained associated alcohol dependence (OR = 1.14), and both remained associated with any substance dependence diagnosis (ORs = 1.11 – 1.19). Overall, the combined model explained 6.0%, 13.2%, and 12.8% of the variance in alcohol dependence, drug dependence, and any substance dependence, respectively.

To ensure the robustness of our results, we also performed: 1) a leave-one-out (LOO) analysis; 2) sex-stratified analyses, and 3) ancestry-specific analysis. The results from the LOO
and sex-stratified analyses were mostly identical to those from the full model. Results in the European ancestry cohorts mirrored the main results, while only the MDD PGS was associated with any of the SUDs in the African ancestry cohorts (see eTables 1-3).

Figure 1 (Panel A) presents the raw prevalence for each outcome across counts of the CERI. Those reporting 3 or more risk factors for drug dependence and 5 or more risk factors for both alcohol and any substance dependence, have a prevalence above lifetime prevalence estimates from nationally representative samples\textsuperscript{51}. Panel B depicts the prevalence of each category of SUD across several mutually exclusive categories: 1) those in the bottom 90% of both the CERI and all PGS (averaged across the five scores); 2) those in the top 10% of the CERI but the bottom 90% of the PGS distribution; 3) those in the top 10% of the PGS distribution and the bottom 90% of the CERI; and 4) those in the top 10% of both PGS and the CERI. There is a steady increase in risk across those with elevated genetic risk, clinical/environmental risk, and both. Those in the top 10% of both PGS and CERI had the highest prevalence of each of the SUDs, though the error bars overlap with the estimates from those in the top 10% of the risk index, alone. Compared to those in the bottom 90% on both, those in the to the top 10% of both have a relative risk of 3.82 (95% CI = 3.53, 4.14) for alcohol dependence, 9.13 (95% CI = 8.22, 10.15) for drug dependence, and 4.17 (95% CI = 3.63, 4.80) for any substance dependence.

Finally, we consider the AUC for the combined model for each of the SUD categories. Figure 2 presents the ROC curves for the full (CERI and PGS) and baseline (covariates only) models for each SUD category. The full AUC in each model ranged from 0.74 for alcohol dependence, 0.86 for drug dependence, and 0.78 for any substance dependence. The overall change in AUC (from the baseline to the full model) that we achieve when adding the risk index and PGS was modest (ΔAUC = 0.05 – 0.08), and this improvement was due in large part to the explanatory power of the CERI.
DISCUSSION

Substance use disorders remain a serious threat to public health. Developing screening protocols that can identify those at greater risk of developing problems has the potential to improve prevention efforts. Prior work on early, targeted interventions for substance misuse suggests that those at highest risk stand to benefit the most from these prevention efforts. In the current analysis, we examined the combination of clinical, environmental, and genetic risk factors for determining who is more likely to develop an SUD in early adulthood. We used previously validated measures of environmental and clinical risk and polygenic scores for externalizing problems, major depressive disorder, problematic alcohol use, alcohol consumption, and schizophrenia. The combination of genetic and social-environmental measures was significantly associated with the development of SUDs. The overall prediction was best for drug dependence, followed by any substance dependence, and lastly, alcohol dependence.

As expected, the CERI was the strongest association with each outcome. The proportion of those meeting criteria for each SUD in persons with 3 or more risk factors for drug dependence, and 5 or more risk factors for alcohol or any substance dependence, surpassed lifetime estimates of these SUD, respectively. Overall, the discriminatory power of the full model (AUC = .74 - .86) was similar to AUC estimates published in the original paper from which many of the risk index items were derived (AUC ~ 0.80). Interestingly, this risk index was originally developed for identifying persons with persistent SUD through early mid-life (~age 40). In the current analysis we demonstrated that the CERI in conjunction with demographic covariates and polygenic scores does equally well for those who meet criteria for any SUD by young adulthood.

In terms of measures of genetic risk, the overall predictive power of the PGS alone was in the range of 1.3 – 2.4%. Only the PGS for externalizing problems and problematic alcohol use were consistently associated with SUD outcomes. The externalizing PGS was associated
Running Header: Predicting SUDs in Young Adulthood

drug dependence, the problematic alcohol use was associated with alcohol dependence, and both were associate with any substance dependence. Overall, these results support prior evidence that genetic risk for SUD consists of a both shared and substance-specific variance\textsuperscript{27,34,40}. Interestingly, even though the effect sizes for the PGS were attenuated in the model, the PGS for externalizing and problematic alcohol use remained significantly associated even when we included the risk index in the model. Since the risk index also included many of the phenotypes each of the PGS measured (e.g., childhood conduct disorder for externalizing, childhood depression for major depressive disorder; and frequent alcohol use for alcohol consumption), part of this attenuation is likely due to the inclusion of the actual phenotypes through which risk for some of these disorders is expressed. PGS are also confounded, at least in part, by environmental variance\textsuperscript{52} and the reduction in effect sizes could be accounting for some of that confounding. Regardless, PGS may add information beyond well-known risk factors, which could eventually prove useful in settings when patients are hesitant to reveal information about potentially stigmatized behaviors, or when information about patients’ early life exposures is unavailable. Additionally, PGS may provide information about risk before behaviors manifest, allowing for earlier intervention.

This analysis has several important limitations. First, although we included individuals of diverse ancestries, the PGS for our African Ancestry samples were still severely underpowered due to the small size of the discovery sample. Large-scale GWAS in diverse cohorts are vital to ensuring that any benefit of precision medicine is delivered in an equitable manner\textsuperscript{53}. Second, while distinct, ancestry is closely related to race-ethnicity, one of the most profound social determinants of health\textsuperscript{54}. Our measure of environmental risk may not fully capture risk factors and other social determinants that contribute to SUDs in populations beyond non-Hispanic whites. Future studies should include racially relevant measures of risk (e.g., experiences of interpersonal racism/discrimination, racial residential segregation) as well as other social and environmental measures that are known risk factors for SUDs (e.g., neighborhood social.
Running Header: Predicting SUDs in Young Adulthood

conditions, alcohol outlet density). Further refinement of known risk factors may allow for even
greater prediction of those at risk of SUD. Finally, while we tried to ensure time order between
risk factors and onset of disorder, some risk factors (particularly adolescent substance use)
could have occurred concurrently with diagnosis. Future work in samples with risk factors
measured before the initiation of substance use (such as the Adolescent Brain Cognitive
Development Study) will be important for replication efforts.

Recognizing that multiple social, clinical, and genetic factors contribute to risk for SUDs
is important as we move towards the goal of an equitable precision medicine that benefits all
segments of the population. The results of this integrative data analysis provide initial evidence
that each of these risk factors contribute unique information for SUDs in early adulthood.
Expanding our sources of information (such as electronic health records, census data from
home of record) and making use of increasingly well-powered PGS will continue to improve our
ability to identify and intervene for those who have the greatest risk of developing SUDs.
ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Institute on Alcohol Abuse and Alcoholism and the National Institute of Drug Abuse of the National Institutes of Health under award numbers R01AA015416, R01DA050721, and K02AA018755; the Academy of Finland (grants 100499, 205585, 118555, 141054, 265240, 308248, 308698 and 312073); and the Scientific and Technological Research Council of Turkey (TÜBİTAK) under award number 114C117 (FA); and the Sigrid Juselius Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding bodies. This research also used summary data from the Psychiatric Genomics Consortium (PGC), the Million Veterans Program (MVP), the GWAS and Sequencing Consortium for Alcohol and Nicotine (GSCAN), UK Biobank, the Genomic Psychiatry Cohort (GPC) and 23andMe, Inc. We would like to thank the many studies that made these consortia possible, the researchers involved, and the participants in those studies, without whom this effort would not be possible. We would also like to thank the research participants and employees of 23andMe.

The Externalizing Consortium: Principal Investigators: Danielle M. Dick, Philipp Koellinger, K. Paige Harden, Abraham A. Palmer. Lead Analysts: Richard Karlsson Linnér, Travis T. Mallard, Peter B. Barr, Sandra Sanchez-Roige. Significant Contributors: Irwin D. Waldman. The Externalizing Consortium has been supported by the National Institute on Alcohol Abuse and Alcoholism (R01AA015416 - administrative supplement), and the National Institute on Drug Abuse (R01DA050721). Additional funding for investigator effort has been provided by K02AA018755, U10AA008401, P50AA022537, as well as a European Research Council Consolidator Grant (647648 EdGe to Koellinger). The content is solely the responsibility of the authors and does not necessarily represent the official views of the above funding bodies.
Hummer) at the University of North Carolina at Chapel Hill. Waves I-V data are from the Add Health Program Project, grant P01 HD31921 (Harris) from Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.

**ALSPAC:** We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors, and Peter Barr and Danielle Dick will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf); This research was specifically funded by the Medical Research Council (MRC) under grants MR/L022206/1, MR/M006727/1, and G0800612/86812; the Wellcome Trust under grant 086684; and the National Institute on Alcohol Abuse and Alcoholism under 5R01AA018333-05. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. **COGA:** We thank The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, T. Foroud; Scientific Director, A. Agrawal; Translational Director, D. Dick, includes eleven different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, T. Foroud, Y. Liu, M. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman, J. Kramer); SUNY Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan, A. Pandey); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, R. Hart, J. Salvatore);
The Children’s Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Virginia Commonwealth University (D. Dick); Icahn School of Medicine at Mount Sinai (A. Goate, P. Slesinger); and Howard University (D. Scott). Other COGA collaborators include: L. Bauer (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X., Xuei, D. Lai, S. O’Connor, (Indiana University); G. Chan (University of Iowa; University of Connecticut); D.B. Chorlian, J. Zhang, P. Barr, S. Kinreich, G. Pandey (SUNY Downstate); N. Mullins (Icahn School of Medicine at Mount Sinai); A. Anokhin, S. Hartz, E. Johnson, V. McCutcheon, S. Saccone (Washington University); J. Moore, Z. Pang, S. Kuo (Rutgers University); A. Merikangas (The Children’s Hospital of Philadelphia and University of Pennsylvania); F. Aliev (Virginia Commonwealth University); H. Chin and A. Parsian are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting- Kai Li, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). All code necessary to replicate this study is available upon request.
REFERENCES


Running Header: Predicting SUDs in Young Adulthood


27. Karlsson Linner R, Mallard TT, Barr PB, et al. Multivariate genomic analysis of 1.5 million people identifies genes related to addiction, antisocial behavior, and health. *Nat Neurosci.* Published online January. h


33. Krueger RF, Hicks BM, Patrick CJ, Carlson SR, Iacono WG, McGue M. Etiological connections among substance dependence, antisocial behavior and personality:


Running Header: Predicting SUDs in Young Adulthood


Table 1: Items included in the Clinical/Environmental Risk Index (CERI)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Low childhood SES</td>
<td>Parent(s) report having less than basic level of education [culturally dependent]; having a low-skill or menial occupation; income at or below the poverty line; or receipt of government assistance.</td>
</tr>
<tr>
<td>2) Family history of SUD</td>
<td>Biological parent self-reports history of SUD for themselves or other biological parent or meets criteria for SUD from clinical interview/AUDIT threshold of 8 or higher.</td>
</tr>
<tr>
<td>3) Childhood behavior problems</td>
<td>Respondent meets criteria for conduct disorder or oppositional defiant disorder from a clinical interview or computer-based prediction; or has a behavior problems score at or above the 90th percentile at 15 or younger.</td>
</tr>
<tr>
<td>4) Childhood internalizing problems</td>
<td>Respondent reports diagnosis of depression/anxiety or panic disorder; meets criteria for internalizing disorder in clinical interview/computer-based prediction; or has a CES-D score above a threshold of 16 at 15 or younger.</td>
</tr>
<tr>
<td>5) Early initiation of substance use</td>
<td>Respondent reports age of first whole alcoholic drink, smoked whole cigarette, or tried cannabis before the age of 15.</td>
</tr>
<tr>
<td>6) Adolescent alcohol use</td>
<td>Frequency of self-reported use 5 or more days per week at age 18 and below.</td>
</tr>
<tr>
<td>7) Adolescent tobacco use</td>
<td>Frequency of self-reported use at daily use at age 18 and below.</td>
</tr>
<tr>
<td>8) Adolescent cannabis use</td>
<td>Frequency of self-reported use 5 or more days per week at age 18 and below.</td>
</tr>
<tr>
<td>9) Peer substance use</td>
<td>Respondent reports the majority of their best friends use alcohol/tobacco/cannabis; their three best friends smoke daily/drink once a month/use cannabis once a month; or more than one friend smokes/drinks alcohol/has tried other drugs.</td>
</tr>
<tr>
<td>10) Traumatic events</td>
<td>Respondent reports exposure to any traumatic event.</td>
</tr>
</tbody>
</table>

Full description of sample specific definitions available in the supplementary information.
Table 2: Descriptive statistics of SUDs risk factors across each cohort

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Cohort</th>
<th>Ancestry</th>
<th>N*</th>
<th>Mean/%</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>55.26%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>53.59%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>56.71%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>51.38%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>51.33%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>53.73%</td>
<td>-</td>
</tr>
<tr>
<td>Age (at last observation)</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>28.89</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>28.84</td>
<td>1.70</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>22.47</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>24.13</td>
<td>5.12</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>24.24</td>
<td>5.26</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>22.44</td>
<td>0.72</td>
</tr>
<tr>
<td>Alcohol dependence</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>3.93%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>12.75%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>5.92%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>11.49%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>21.14%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>8.55%</td>
<td>-</td>
</tr>
<tr>
<td>Drug dependence</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>6.73%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>10.79%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>0.78%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>26.44%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>23.59%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>1.34%</td>
<td>-</td>
</tr>
<tr>
<td>Any substance dependence†</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>11.21%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>25.81%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>8.87%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>30.69%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>34.66%</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>10.98%</td>
<td>-</td>
</tr>
<tr>
<td>CERI</td>
<td>Add Health</td>
<td>AFR</td>
<td>1,605</td>
<td>1.95</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>Add Health</td>
<td>EUR</td>
<td>4,855</td>
<td>2.07</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>ALSPAC</td>
<td>EUR</td>
<td>4,733</td>
<td>2.08</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>AFR</td>
<td>870</td>
<td>3.98</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>COGA</td>
<td>EUR</td>
<td>1,878</td>
<td>3.65</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td>FinnTwin12</td>
<td>EUR</td>
<td>1,193</td>
<td>2.62</td>
<td>1.27</td>
</tr>
</tbody>
</table>

* Available samples with genotypic, phenotypic, and environmental risk data
† Any substance dependence includes those who meet criteria for alcohol, nicotine, or drug dependence.
CERI = clinical/environmental risk index
<table>
<thead>
<tr>
<th></th>
<th>Alcohol Dependence</th>
<th>Drug Dependence</th>
<th>Any Substance Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>PGS Only Model*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALCC PGS</td>
<td>1.06 (1.00, 1.11)</td>
<td>1.04 (0.97, 1.10)</td>
<td>1.00 (0.96, 1.05)</td>
</tr>
<tr>
<td>ALCP PGS</td>
<td>1.15 (1.10, 1.21)</td>
<td>1.08 (1.01, 1.14)</td>
<td>1.12 (1.07, 1.16)</td>
</tr>
<tr>
<td>EXT PGS</td>
<td>1.19 (1.13, 1.25)</td>
<td>1.35 (1.28, 1.41)</td>
<td>1.34 (1.29, 1.39)</td>
</tr>
<tr>
<td>MDD PGS</td>
<td>0.98 (0.92, 1.04)</td>
<td>1.05 (0.99, 1.12)</td>
<td>1.02 (0.97, 1.06)</td>
</tr>
<tr>
<td>SCZ PGS</td>
<td>1.07 (1.01, 1.13)</td>
<td>1.06 (0.99, 1.13)</td>
<td>1.05 (1.00, 1.10)</td>
</tr>
<tr>
<td>ΔPseudo-R²</td>
<td>0.013</td>
<td>0.018</td>
<td>0.024</td>
</tr>
<tr>
<td>CERI Only Model*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERI</td>
<td>1.37 (1.33, 1.41)</td>
<td>1.67 (1.61, 1.72)</td>
<td>1.58 (1.54, 1.63)</td>
</tr>
<tr>
<td>ΔPseudo-R²</td>
<td>0.054</td>
<td>0.129</td>
<td>0.120</td>
</tr>
<tr>
<td>Combined Model*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CERI</td>
<td>1.35 (1.32, 1.38)</td>
<td>1.64 (1.61, 1.67)</td>
<td>1.55 (1.52, 1.58)</td>
</tr>
<tr>
<td>ALCC PGS</td>
<td>1.05 (0.99, 1.11)</td>
<td>1.04 (0.97, 1.11)</td>
<td>1.00 (0.95, 1.04)</td>
</tr>
<tr>
<td>ALCP PGS</td>
<td>1.14 (1.09, 1.20)</td>
<td>1.06 (0.99, 1.13)</td>
<td>1.11 (1.06, 1.16)</td>
</tr>
<tr>
<td>EXT PGS</td>
<td>1.08 (1.01, 1.14)</td>
<td>1.14 (1.07, 1.22)</td>
<td>1.19 (1.14, 1.24)</td>
</tr>
<tr>
<td>MDD PGS</td>
<td>0.95 (0.89, 1.01)</td>
<td>0.99 (0.93, 1.06)</td>
<td>0.97 (0.92, 1.02)</td>
</tr>
<tr>
<td>SCZ PGS</td>
<td>1.06 (0.99, 1.12)</td>
<td>1.02 (0.95, 1.09)</td>
<td>1.04 (0.99, 1.09)</td>
</tr>
<tr>
<td>ΔPseudo-R²</td>
<td>0.06</td>
<td>0.132</td>
<td>0.128</td>
</tr>
</tbody>
</table>

* All models included age, sex, and cohort as covariates. PGS residualized on age, sex, and first 10 ancestral principal components. Bolded estimates = \( p < .05 \) after correction for multiple testing (\( p < .05/3 = .017 \)). \( \Delta \text{Pseudo-R}² \) denotes pseudo-\( R² \) above model including age, sex, and cohort. CI = confidence interval; PGS = polygenic score; CERI = clinical/environmental risk index.
FIGURE CAPTIONS

Figure 1: SUD Prevalence Across Genetic and Environmental Risk Factors

Panel A: Prevalence (and 95% confidence intervals) of those who meet criteria for alcohol, drug, any substance dependence across counts for items in the risk index. Panel B: Prevalence (and 95% confidence intervals) of those who meet criteria for alcohol, drug, or any substance dependence across four categories: 1) those below the 90th percentile for all PGS and the CERI; 2) those at or above the 90th percentile for the CERI; 3) those at or above the 90th percentile for all PGS; and 4) those at or above the 90th percentile for both the CERI and PGS. PGS and risk index were first residualized on sex, age, age$^2$, cohort, sex$*$age, sex$*$age$^2$, sex$*$cohort, cohort$*$age, cohort$*$age$^2$, sex$*$cohort$*$age, and sex$*$cohort$*$age$^2$. Dotted colored lines represent corresponding lifetime prevalence estimates for alcohol dependence (red), drug dependence (green), and any substance use disorder (blue) from nationally representative data$^{51}$.

Figure 2: ROC Curves for Combined and Baseline Models

Receiver operating characteristic (ROC) curves for baseline models (red line, covariates only) and the full models (blue line, PGS + CERI + covariates) for each substance use disorder. Area under the curve (AUC) is presented for the PGS model in each cell. Change in AUC represents value of the difference between AUC from the full model and AUC from the base model.
Alcohol Dependence

Drug Dependence

Any Substance Dependence

Full AUC = 0.74
Change in AUC = 0.05

Full AUC = 0.86
Change in AUC = 0.07

Full AUC = 0.78
Change in AUC = 0.08