

1 Clinical, Environmental, and Genetic Risk Factors for Substance Use Disorders:

2 Characterizing Combined Effects across Multiple Cohorts

3 Peter B. Barr, Ph.D.<sup>1,2</sup>, Morgan N. Driver, B.S.<sup>3</sup>, Sally I-Chun Kuo, Ph.D.<sup>4</sup>,  
4 Mallory Stephenson, M.S.<sup>5</sup>, Fazil Aliev, Ph.D.<sup>6</sup>, Richard Karlsson Linnér, Ph.D.<sup>7</sup>,  
5 Jesse Marks<sup>8</sup>, Andrey P. Anokhin, Ph.D.<sup>9</sup>, Kathleen Bucholz, Ph.D.<sup>9</sup>, Grace Chan, Ph.D.<sup>10,11</sup>,  
6 Howard J. Edenberg, Ph.D.<sup>12,13</sup>, Alexis C. Edwards, Ph.D.<sup>5</sup>, Meredith W. Francis, Ph.D.<sup>9</sup>,  
7 Dana B. Hancock<sup>8</sup>, K. Paige Harden, Ph.D.<sup>14,15</sup>, Chella Kamarajan, Ph.D.<sup>1</sup>,  
8 Jaakko Kaprio M.D., Ph.D.<sup>16</sup>, Sivan Kinreich, Ph.D.<sup>1</sup>, John Kramer, Ph.D.<sup>11</sup>,  
9 Samuel Kuperman, M.D.<sup>11</sup>, Antti Latvala, Ph.D.<sup>17</sup>, Jacquelyn L. Meyers, Ph.D.<sup>1,2</sup>,  
10 Abraham A. Palmer, Ph.D.<sup>18,19</sup>, Martin H. Plawecki M.D., Ph.D.<sup>20</sup>, Bernice Porjesz, Ph.D.<sup>1</sup>,  
11 Richard J. Rose, Ph.D.<sup>21</sup>, Marc A. Schuckit, M.D.<sup>18</sup>, Jessica E. Salvatore, Ph.D.<sup>4</sup>,  
12 and Danielle M. Dick, Ph.D.<sup>6</sup>

13  
14 <sup>1</sup> Department of Psychiatry and Behavioral Sciences, SUNY Downstate Health Sciences University,  
15 Brooklyn, NY, USA.

16 <sup>2</sup> VA New York Harbor Healthcare System, Brooklyn, NY, USA.

17 <sup>3</sup> Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA.

18 <sup>4</sup> Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, Piscataway, NJ,  
19 USA.

20 <sup>5</sup> Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia  
21 Commonwealth University, Richmond, VA, USA.

22 <sup>6</sup> Rutgers Addiction Research Center, Rutgers University, Piscataway, NJ, USA.

23 <sup>7</sup> Department of Economics, Leiden University, Leiden, The Netherlands.

24 <sup>8</sup> Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA

25 <sup>9</sup> Department of Psychiatry, School of Medicine, Washington University in St. Louis, St Louis, MO, USA.

26 <sup>10</sup> Department of Psychiatry, School of Medicine, University of Connecticut, Farmington, CT, USA.

27 <sup>11</sup> Department of Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, IA, USA

28 <sup>12</sup> Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis,  
29 IN, USA

30 <sup>13</sup> Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University,  
31 Indianapolis, IN, USA

32 <sup>14</sup> Department of Psychology, University of Texas at Austin, Austin, TX, USA.

33 <sup>15</sup> Population Research Center, University of Texas at Austin, Austin, TX, USA.

34 <sup>16</sup> Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

35 <sup>17</sup> Institute of Criminology and Legal Policy, University of Helsinki, Helsinki, Finland

36 <sup>18</sup> Department of Psychiatry, University of California San Diego, La Jolla, CA, USA.

37 <sup>19</sup> Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA.

38 <sup>20</sup> Department of Psychiatry, School of Medicine, Indiana University, Indianapolis, IN, USA

39 <sup>21</sup> Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA

40  
41 Corresponding author: Peter B. Barr

42 Department of Psychiatry and Behavioral Sciences

43 SUNY Downstate Health Sciences University

44 450 Clarkson Ave, MSC 1203, Brooklyn, NY 11203

45 e-mail: peter.barr@downstate.edu.

46  
47 Submission for: Molecular Psychiatry

48 Running Header: Characterizing Risk for SUDS Across Cohorts

49 Word count abstract: 253; Word count text body: 3,537

50 Figures and Tables: 3 Tables, 2 Figures

51 **ABSTRACT**

52           Substance use disorders (SUDs) incur serious social and personal costs. Risk for SUDs  
53 is complex, ranging from social conditions to individual genetic variation. We examined whether  
54 models that include a clinical/environmental risk index (CERI) and polygenic scores (PGS) are  
55 able to identify individuals at increased risk of SUD in young adulthood across four longitudinal  
56 cohorts for a combined sample of  $N = 15,134$ . Our analyses included participants of European  
57 ( $N_{EUR} = 12,659$ ) and African ( $N_{AFR} = 2,475$ ) ancestries. SUD outcomes included: 1) alcohol  
58 dependence, 2) nicotine dependence; 3) drug dependence, and 4) any substance dependence.  
59 In the models containing the PGS and CERI, the CERI was associated with all three outcomes  
60 (ORs = 1.37 – 1.67). PGS for problematic alcohol use, externalizing, and smoking quantity were  
61 associated with alcohol dependence, drug dependence, and nicotine dependence, respectively  
62 (OR = 1.11 – 1.33). PGS for problematic alcohol use and externalizing were also associated  
63 with any substance dependence (ORs = 1.09 – 1.18). The full model explained 6% - 13% of the  
64 variance in SUDs. Those in the top 10% of CERI and PGS had relative risk ratios of 3.86 - 8.04  
65 for each SUD relative to the bottom 90%. Overall, the combined measures of clinical,  
66 environmental, and genetic risk demonstrated modest ability to distinguish between affected  
67 and unaffected individuals in young adulthood. PGS were significant but added little in addition  
68 to the clinical/environmental risk index. Results from our analysis demonstrate there is still  
69 considerable work to be done before tools such as these are ready for clinical applications.

## 70 INTRODUCTION

71 Substance use disorders (SUDs) are associated with substantial costs to affected  
72 individuals, their families, and society. An estimated 107,000 Americans died as the result of an  
73 overdose in 2021 <sup>1</sup>. In 2016, alcohol use contributed 4.2% to the global disease burden and  
74 other drug use contributed 1.3% <sup>2</sup>. Excessive alcohol use and illicit drug use cost the United  
75 States an annual \$250 billion <sup>3</sup> and \$190 billion <sup>4</sup> respectively. Given the substantial human and  
76 economic costs of substance misuse and disorders, understanding the combined impact of  
77 important risk factors across multiple levels of analysis has important public health implications.

78 Substance use disorders are complex phenomena, and the development of substance  
79 related problems can be attributed to factors ranging from broader social and economic  
80 conditions to individual genetic variation <sup>5-10</sup>. Prior research using a multifactorial index of  
81 clinical and environmental risk factors (e.g., childhood disadvantage, family history of SUD,  
82 childhood conduct problems, childhood depression, early exposure to substances, frequent use  
83 during adolescence) found it useful in identifying those with persistent SUDs <sup>11</sup>.

84 More recently, polygenic scores (PGS), which aggregate risk for a trait across the  
85 genome using information from genome-wide association studies (GWAS), were robustly  
86 associated with substance use <sup>12</sup> and substance related problems <sup>13</sup> across adolescence and  
87 into young adulthood. However, though robustly associated, current PGS do poorly in identifying  
88 individuals affected by SUDs <sup>14</sup>. To date, there is limited work on the combined impact of  
89 genetic, environmental, and clinical risk factors for SUDs. Prior work combining individual  
90 genetic variants and clinical features outperformed clinical features alone <sup>15</sup>, but individual  
91 variants have limited predictive power. In other medical conditions, such as melanoma <sup>16</sup> or  
92 ischemic stroke <sup>17</sup>, combining clinical and genetic risk factors showed improvement predicting  
93 risk for a specific outcome over models using individual risk factors.

## Running Header: Characterizing Risk for SUDs Across Cohorts

94            In the current study, we examine the joint association of early life clinical/environmental  
95 risk factors and PGS with SUDs in early adulthood across four longitudinal cohorts: the National  
96 Longitudinal Study of Adolescent to Adult Health (Add Health); the Avon Longitudinal Study of  
97 Parents and Children (ALSPAC); the Collaborative Study on the Genetics of Alcoholism  
98 (COGA); and the youngest cohort of the Finnish Twin Cohort Study (FinnTwin12). These  
99 samples include population-based cohorts from three countries (United States, England, and  
100 Finland) and a predominantly high-risk sample. Two of the samples (COGA and Add Health) are  
101 ancestrally diverse. We focus on early adulthood as this is a critical period for the development  
102 and onset of SUDs <sup>18</sup>. Our research questions are guided by the understanding that risk factors  
103 for SUDs range across multiple levels of analysis.

## 104    **METHODS**

### 105    *Samples*

106            *Add Health* is a nationally representative longitudinal study of adolescents followed into  
107 adulthood in the United States <sup>19</sup>. Data have been collected from Wave I when respondents  
108 were between 11-18 (1994-1995) to Wave V (2016-2018) when respondents were 35-42. The  
109 current analysis uses data from Waves I, II, and Wave IV.

110            *ALSPAC* is an ongoing, longitudinal population-based study of a birth cohort in the  
111 (former) Avon district of Southwest England <sup>20-23</sup>. Pregnant female residents with an expected  
112 date of delivery between April 1, 1991 and December 31, 1992 were invited to participate (N =  
113 14,541 pregnant women, 80% of those eligible). This analysis uses data up to the age 24  
114 assessment (details of all the data that is available through a searchable, web-based tool:  
115 <http://www.bristol.ac.uk/alspac/researchers/our-data/>).

116            *COGA* is a family-based sample consisting of alcohol dependent individuals (identified  
117 through treatment centers across the United States), their extended families, and community

## Running Header: Characterizing Risk for SUDs Across Cohorts

118 controls (N ~16,000)<sup>24,25</sup>. We use a prospective sample of offspring of the original COGA  
119 participants (baseline ages 12-22, N = 3,573) that have been assessed biennially since  
120 recruitment (2004-2019)<sup>26</sup>.

121 *FinnTwin12* is a population-based study of Finnish twins born 1983–1987 identified  
122 through Finland’s Central Population Registry. A total of 2,705 families (87% of all identified)  
123 returned the initial family questionnaire late in the year in which twins reached age 11<sup>27</sup>. Twins  
124 were invited to participate in follow-up surveys when they were ages 14, 17, and approximately  
125 22.

126 Each cohort includes a wide range of social, behavioral, and phenotypic data measured  
127 across the life course. The SUD measures were derived from the corresponding young adult  
128 phases of data collection in each cohort (mean ages ~ 22 - 28). A full description of each  
129 sample is presented in the supplementary information (section 2).

### 130 *Measures*

#### 131 *Lifetime Diagnosis of Substance Use Disorder*

132 We constructed measures of lifetime SUD diagnosis based on the data that were  
133 available in each of the samples, defined as meeting criteria for four, non-mutually exclusive  
134 categories of substance dependence: 1) alcohol dependence; 2) nicotine dependence; 3) drug  
135 dependence (inclusive of drugs such as cannabis, cocaine, opioids, sedatives, etc.); and 4) any  
136 substance dependence (alcohol, nicotine, or drug). Our analyses focused primarily on DSM-IV  
137 as this diagnostic system was most consistently used across all samples. There was one  
138 exception: in each of the samples, nicotine dependence was measured using a cutoff of 7 or  
139 higher on the Fagerstrom Test for Nicotine Dependence (FTND)<sup>28</sup>. Where possible, we drew  
140 measures of substance dependence from data collected during young adulthood to try and  
141 maintain temporal ordering between SUD diagnoses and measured risk factors.

## Running Header: Characterizing Risk for SUDs Across Cohorts

### 142 *Clinical/Environmental Risk Index*

143 We created a clinical/environmental risk index (CERI) considering a variety of  
144 established risk factors for SUD (Table 1). The CERI included ten validated early life risk factors  
145 associated with later development of SUDs, including: low childhood socioeconomic status  
146 (SES), family history of SUD, early initiation of substance use, childhood internalizing problems,  
147 childhood externalizing problems, frequent drinking in adolescence, frequent smoking in  
148 adolescence, frequent cannabis use in adolescence, peer substance use, and exposure to  
149 trauma/traumatic experiences<sup>11,29,30</sup>. We dichotomized each risk factor (present vs not present)  
150 and summed them into an index for each person ranging from 0 to 10, providing a single  
151 measure of aggregate risk. Dichotomizing these items allowed us to harmonize measures  
152 across each sample in an interpretable manner. A full list of how each measure is defined within  
153 each of the samples is available in the supplementary information (section 3).

### 154 *Polygenic Scores*

155 We constructed polygenic scores (PGS), which are aggregate measures of the number  
156 of risk alleles individuals carry weighted by effect sizes from GWAS summary statistics, from six  
157 recent GWAS of SUDs and comorbid conditions including: 1) externalizing problems (EXT)<sup>31</sup>; 2)  
158 major depressive disorder (MDD)<sup>32</sup>; 3) problematic alcohol use<sup>33</sup> (ALCP); 4) alcohol  
159 consumption (drinks per week, ALCC)<sup>34,35</sup>; 5) cigarettes per day/FTND (CPD)<sup>34,36</sup>; and 6)  
160 schizophrenia (SCZ)<sup>37,38</sup>. We focused on these PGS, specifically, because: 1) SUDs show  
161 strong genetic overlap with other externalizing<sup>39-41</sup>, internalizing<sup>32,42</sup>, and psychotic disorders  
162<sup>33,43,44</sup>; 2) both shared and substance-specific genetic risk are associated with later SUDs<sup>45-47</sup>;  
163 and 3) substance use and SUDs have only partial genetic overlap<sup>48,49</sup>. Therefore, our PGS  
164 cover a spectrum of genetic risk for SUDs, using the most current and well-powered results for  
165 each of the listed domains (see supplementary information section 4 for a detailed description).

## Running Header: Characterizing Risk for SUDs Across Cohorts

166 GWAS have been overwhelmingly limited to individuals of European ancestries<sup>50,51</sup>.  
167 Importantly, PGS derived from GWAS of one ancestry do not always transport into other  
168 ancestral populations<sup>52,53</sup>. We therefore used PRS-CSx<sup>54</sup>, a new method that combines  
169 information from well-powered GWAS (typically of European ancestries) and ancestrally  
170 matched GWAS to improve the predictive power of PGS in the African ancestry samples from  
171 Add Health and COGA. PRS-CSx integrates GWAS summary statistics across multiple input  
172 populations and employs a Bayesian approach to correct GWAS summary statistics for the non-  
173 independence of SNPs in linkage disequilibrium (LD) with one another<sup>54</sup>. For participants of  
174 European ancestries, we used the EUR derived PRS-CSx results, while we used the EUR+AFR  
175 meta-analyzed results for the African ancestry participants. See the supplementary information  
176 (section 5) for details.

### 177 *Analytic Strategy*

178 We pooled all the data for analysis using a fixed effects integrative data analytic (IDA)  
179 approach<sup>55</sup>. The IDA approach is more powerful than traditional meta-analyses when one has  
180 access to raw data for each of the contributing samples. Our approach to harmonization and  
181 pooling was as follows. First, we defined the measures and cutoffs to be used in each of the  
182 samples, creating the CERI, PGS, and SUD outcomes at the cohort level. Second, within each  
183 cohort, we regressed each PGS on age, age<sup>2</sup>, sex, sex\*age, sex\*age<sup>2</sup>, and the first 10 ancestral  
184 PCs (specific to each sample) to account for population stratification in the PGS. Next, we  
185 pooled all the data for analysis. We included cohort as a fixed effect for each of the six cohorts  
186 (4 samples, of which two were split by ancestry) in subsequent analyses. Additionally, we  
187 included age of last observation and sex as covariates.

188 We estimated a series of nested logistic regression models with the pooled data: 1) a  
189 baseline model (sex, age, and cohort), 2) a genetic risk model (baseline + PGS), 3) a

## Running Header: Characterizing Risk for SUDs Across Cohorts

190 clinical/environmental risk model (baseline + CERI), and 4) a combined risk model (baseline +  
191 PGS + CERI). Because COGA and FT12 included a large number of related individuals, we  
192 adjusted for familial clustering using cluster-robust standard errors<sup>56</sup>. To assess the predictive  
193 accuracy of each model, we took the difference in pseudo- $R^2$  ( $\Delta Pseudo-R^2$ )<sup>57</sup>, between the  
194 baseline and corresponding models. Finally, we calculated the discriminatory power of the  
195 combined model using the area under the curve (AUC) from a receiver operating characteristic  
196 (ROC) curve. We included a variety of robustness checks to ensure that no single cohort in the  
197 IDA was unduly influencing the results. Our analytic strategy was preregistered on the Open  
198 Science Framework (<https://osf.io/etbw8>). Deviations from the preregistration are described in  
199 the supplementary information (section 6).

## 200 RESULTS

201 Table 2 contains the descriptive statistics for each of the cohorts and ancestries. Each  
202 cohort had similar proportions of females (~51% - 56%). The mean ages ranged from ~22 to  
203 ~29 years of age. The COGA cohorts (both European and African ancestries) reported the  
204 highest rates of SUD, an expected finding given the nature of the sample (highly selected for  
205 SUDs). Add Health participants generally had higher rates of SUD than ALSPAC or FinnTwin12,  
206 but lower than COGA. Finally, ALSPAC and FinnTwin12 reported similar levels of alcohol,  
207 nicotine, drug, and any substance dependence. COGA participants reported higher mean  
208 values on the CERI. The remaining cohorts report relatively similar rates of risk factor exposure.

209 Table 3 presents the results from the *PGS only*, *CERI only*, and *combined* models for  
210 each outcome. Three of the six PGS were associated with the SUD outcomes in the *PGS only*  
211 model. EXT was associated with each of the SUD outcomes (EXT OR = 1.18 – 1.50); ALCP  
212 was associated with alcohol dependence and any substance dependence (ALCP OR = 1.10 –  
213 1.13); and CPD was associated with nicotine dependence (CPD OR = 1.33). In the *CERI only*

## Running Header: Characterizing Risk for SUDs Across Cohorts

214 models, the CERI was consistently associated across each of the SUD categories (ORs = 1.37  
215 – 1.67). When we combined the PGS and CERI into the same model, the CERI remained  
216 significant across SUDs and was largely unchanged (ORs = 1.35 – 1.65). EXT remained  
217 associated with drug dependence (OR = 1.11) and nicotine dependence (OR = 1.33), ALCP  
218 remained associated alcohol dependence (OR = 1.12), and CPD remained associated with  
219 nicotine dependence (OR = 1.31). Both EXT and ALCP remained associated with any  
220 substance dependence diagnosis (ORs = 1.09 – 1.18). Overall, the combined model explained  
221 5.9%, 12.6%, 13.1%, and 12.8% of the variance in alcohol dependence, nicotine dependence,  
222 drug dependence, and any substance dependence, respectively.

223 Figure 1 (Panel A) presents the raw prevalence for each outcome across counts of the  
224 CERI. The proportion of those meeting criteria for SUDs among those reporting 3 or more, 5 or  
225 more, and 7 or more risk factors surpassed lifetime prevalence estimates from nationally  
226 representative samples for drug dependence, alcohol dependence, and nicotine dependence,  
227 respectively<sup>58</sup>. Panel B depicts the prevalence of each category of SUD across several mutually  
228 exclusive categories: 1) those in the bottom 90% of both the CERI and all PGS (averaged  
229 across the six scores); 2) those in the top 10% of the CERI but the bottom 90% of the PGS  
230 distribution; 3) those in the top 10% of the PGS distribution and the bottom 90% of the CERI;  
231 and 4) those in the top 10% of both PGS and the CERI. There is an increase in risk across  
232 those with elevated genetic risk, clinical/environmental risk, and both. Those in the top 10% of  
233 both PGS and CERI had the highest prevalence of each of the SUDs, though the error bars  
234 overlap with the estimates from those in the top 10% of the risk index, alone. Compared to  
235 those in the bottom 90% on both, those in the to the top 10% of both have a relative risk of 3.86  
236 (95% CI = 3.20, 4.65) for alcohol dependence, 6.11 (95% CI = 4.84, 7.72) for nicotine  
237 dependence, 8.04 (95% CI = 6.92, 9.36) for drug dependence, and 4.05 (95% CI = 3.64, 4.51)  
238 for any substance dependence.

## Running Header: Characterizing Risk for SUDs Across Cohorts

239 Finally, we considered the AUC for the combined model for each of the SUD categories.  
240 Figure 2 presents the ROC curves for the full (CERI and PGS) and baseline (covariates only)  
241 models for each SUD category. The AUC for each combined model was 0.74 for alcohol  
242 dependence, 0.82 for nicotine dependence, 0.86 for drug dependence, and 0.78 for any  
243 substance dependence. The overall change in AUC (from the baseline to the full model) that we  
244 achieve when adding the CERI and PGS was modest ( $\Delta\text{AUC} = 0.05 - 0.10$ ), and this  
245 improvement was due in large part to the explanatory power of the CERI. ROC curves for the  
246 CERI only and PGS only models are presented in Supplemental Figure 6.

### 247 *Sensitivity Analyses*

248 We performed a variety of sensitivity analyses. Results from leave-one-out (LOO) and  
249 sex-stratified analyses were largely similar to those from the main results. In ancestry stratified  
250 analyses, results in the cohorts of European ancestries largely mirrored the main results. None  
251 of the PGS were associated with SUDs in the cohorts of African ancestries. Effect sizes for the  
252 CERI were largely similar across European and African ancestries (see Supplemental Tables  
253 S1-S3) and were mostly stable when removing individual risk factors (supplemental information  
254 section 7).

255 We also tested for interactions between the PGS and CERI and cohort (Add Health  
256 EUR as the reference group). There were few significant interactions and no consistent patterns  
257 in variation for PGS, though the CERI did show considerable variation across cohort  
258 (Supplemental Table S4). Finally, we fit complimentary models using a random effects  
259 approach, allowing the slopes for the PGS and CERI to vary randomly across cohort. Random  
260 slopes for PGS did not consistently improve model fit, though a random slope for the CERI  
261 consistently improved model fit (Supplemental Table S5). We compared the parameter

## Running Header: Characterizing Risk for SUDs Across Cohorts

262 estimates from the random effect models to the main analyses and results were largely  
263 consistent (Supplemental Table S6).

## 264 **DISCUSSION**

265 Substance use disorders remain a serious threat to public health<sup>59</sup>. In the current  
266 analysis, we examined the combination of clinical, environmental, and genetic risk factors for  
267 determining who is more likely to develop a SUD in early adulthood. We used previously  
268 validated measures of environmental and clinical risk<sup>11,29,30</sup> and polygenic scores for  
269 externalizing problems<sup>31</sup>, major depressive disorder<sup>32</sup>, problematic alcohol use<sup>33,35</sup>, alcohol  
270 consumption<sup>34,35</sup>, cigarettes per day/nicotine dependence<sup>34,36</sup>, and schizophrenia<sup>37,38</sup>. The  
271 combination of genetic and social-environmental measures was significantly associated with the  
272 development of SUDs. The overall association was strongest for drug dependence, followed by  
273 any substance dependence, nicotine dependence, and alcohol dependence.

274 The CERI was the strongest association with each outcome. The proportion of those  
275 meeting criteria for each SUD surpassed lifetime estimates in persons with 3 or more, 5 or  
276 more, and 7 or more risk factors for drug dependence, alcohol dependence, and nicotine  
277 dependence, respectively. The discriminatory power of the combined model (AUC = .74 - .86)  
278 was similar to AUC estimates published in the original paper from which many of the risk index  
279 items were derived (AUC ~ 0.80)<sup>11</sup>. Interestingly, this risk index was originally developed for  
280 identifying persons with persistent SUD through early mid-life (~age 40). In the current analysis  
281 we demonstrated that the CERI in conjunction with demographic covariates and PGS does  
282 equally well for those who meet criteria for any SUD by young adulthood.

283 The overall predictive power of the PGS alone was in the range of 1.1 – 3.7%. Only the  
284 PGS for externalizing problems, problematic alcohol use, and cigarettes per day were  
285 consistently associated with SUD outcomes. The PGS for externalizing problems was

## Running Header: Characterizing Risk for SUDs Across Cohorts

286 associated with drug dependence and nicotine dependence, the PGS for problematic alcohol  
287 use PGS was associated with alcohol dependence, and both were associated with any  
288 substance dependence. The PGS for cigarettes per day was only associated with nicotine  
289 dependence. Overall, these results support prior evidence that genetic risk for SUDs consists of  
290 a both shared and substance-specific variance<sup>31,41,47</sup>.

291 Interestingly, even though the effect sizes were attenuated in the model, the PGS for  
292 externalizing problems, problematic alcohol use, and cigarettes per day remained significantly  
293 associated when we included the CERI, though the additional information the PGS provided  
294 was minimal. Since the CERI also included many of the phenotypes each of the PGS measured  
295 (e.g., childhood conduct disorder for externalizing, childhood depression for major depressive  
296 disorder; and frequent alcohol use for alcohol consumption), part of this attenuation is likely due  
297 to the inclusion of the actual phenotypes through which risk for some of these disorders is  
298 expressed. PGS are also confounded by environmental variance<sup>59</sup> and the reduction in effect  
299 sizes could be accounting for some of that confounding. PGS may add information beyond well-  
300 known risk factors, which could prove useful when information on certain exposures or  
301 behaviors is unavailable.

302 Further refinement of risk measures may improve our ability to develop screening  
303 protocols for those at greater risk of developing substance-related problems. Early detection has  
304 the potential to improve prevention efforts, as prior work suggests that those at highest risk of  
305 substance misuse stand to benefit the most from prevention efforts<sup>60</sup>. Ideally, screening tools  
306 for SUD risk would include measures of social, clinical, and genetic risk factors, as each impacts  
307 the development of SUDs<sup>5-7,61,62</sup>. In the push for precision medicine, very often the focus is on  
308 biological information, but social determinants of health are also critically important.

309 Currently, these tools are not ready for clinical use. If we reach the point where social,  
310 clinical, and genetic information become sufficiently powerful, we must recognize that identifying  
311 persons for early intervention carries a significant risk. Screening for social determinants has the

## Running Header: Characterizing Risk for SUDs Across Cohorts

312 potential for unintended consequences, including further stigmatization<sup>63</sup>. Genetic information  
313 has even more potential for abuse and stigmatization. Policy makers must ensure that there is  
314 comprehensive legal protection against discrimination using any form of information.  
315 Additionally, any attempt to use social, clinical, or genetic information for targeted intervention or  
316 identification in a clinical setting must be done so in a patient-centered approach, rather than  
317 any “one-size fits all” that exclude patients from their own healthcare decisions<sup>64</sup>.

318 Our analysis has several important limitations. First, although we included individuals of  
319 diverse ancestries, the PGS for our samples of African ancestries were severely underpowered  
320 due to the small size of the discovery sample. Large-scale GWAS in diverse cohorts are vital to  
321 ensuring that any benefit of precision medicine is shared equitably across the population<sup>65</sup>.  
322 Second, while distinct, ancestry is related to race-ethnicity, and with it, racism and racial  
323 discrimination, some of the most profound social determinants of health<sup>66</sup>. Our measure of  
324 environmental risk was crude and may not fully capture risk factors that contribute to SUDs in  
325 populations beyond non-Hispanic Whites. Future studies should include racially relevant  
326 measures of risk (e.g., experiences of interpersonal racism/discrimination, racial residential  
327 segregation) as well as other social and environmental measures that are known risk factors for  
328 SUDs (e.g., neighborhood social conditions, alcohol outlet density). Further refinement of known  
329 risk factors may allow for better prediction of those at risk of developing an SUD. We did  
330 observe variation in the predictive ability of the CERI across cohorts, suggesting the observed  
331 effect may differ in magnitude across populations. We therefore urge caution in overinterpreting  
332 study results. Finally, while we tried to ensure time order between risk factors and onset of  
333 disorder, some risk factors (particularly adolescent substance use) could have occurred  
334 concurrently with diagnosis. Future work in samples with risk factors measured before the  
335 initiation of substance use (such as the Adolescent Brain Cognitive Development Study) will be  
336 important for replication efforts.

## Running Header: Characterizing Risk for SUDs Across Cohorts

337           Recognizing that multiple social, clinical, and genetic factors contribute to risk for SUDs  
338 is important as we move towards the goal precision medicine that benefits all segments of the  
339 population. There is still much work to be done before tools such as these are useful in a clinical  
340 setting. However, the results of this integrative data analysis provide initial evidence *all* of these  
341 risk factors contribute unique information to SUDs in early adulthood. Expanding our sources of  
342 information (such as electronic health records, census data from home of record) and making  
343 use of increasingly well-powered PGS will continue to improve our ability to understand how  
344 SUDs develop.

## 345 **ACKNOWLEDGEMENTS**

346 Research reported in this publication was supported by the National Institute on Alcohol Abuse  
347 and Alcoholism and the National Institute of Drug Abuse of the National Institutes of Health  
348 under award numbers R01AA015416, R01DA050721, R01DA042090, and K02AA018755; the  
349 Academy of Finland (grants 100499, 205585, 118555, 141054, 265240, 308248, 308698 and  
350 312073); and the Scientific and Technological Research Council of Turkey (TÜBİTAK) under  
351 award number 114C117 (FA); and the Sigrid Juselius Foundation. The content is solely the  
352 responsibility of the authors and does not necessarily represent the official views of any of the  
353 funding bodies. This research also used summary data from the Psychiatric Genomics  
354 Consortium (PGC), the Million Veterans Program (MVP), the GWAS and Sequencing  
355 Consortium for Alcohol and Nicotine (GSCAN), UK Biobank, the Genomic Psychiatry Cohort  
356 (GPC) and 23andMe, Inc. We would like to thank the many studies that made these consortia  
357 possible, the researchers involved, and the participants in those studies, without whom this  
358 effort would not be possible. We would also like to thank the research participants and  
359 employees of 23andMe.

## Running Header: Characterizing Risk for SUDs Across Cohorts

360 **The Externalizing Consortium:** Principal Investigators: Danielle M. Dick, Philipp Koellinger, K.  
361 Paige Harden, Abraham A. Palmer. Lead Analysts: Richard Karlsson Linnér, Travis T. Mallard,  
362 Peter B. Barr, Sandra Sanchez-Roige. Significant Contributors: Irwin D. Waldman. The  
363 Externalizing Consortium has been supported by the National Institute on Alcohol Abuse and  
364 Alcoholism (R01AA015416 - administrative supplement), and the National Institute on Drug  
365 Abuse (R01DA050721). Additional funding for investigator effort has been provided by  
366 K02AA018755, U10AA008401, P50AA022537, as well as a European Research Council  
367 Consolidator Grant (647648 EdGe to Koellinger). The content is solely the responsibility of the  
368 authors and does not necessarily represent the official views of the above funding bodies. **Add**  
369 **Health:** Add Health is directed by Robert A. Hummer and funded by the National Institute on  
370 Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and  
371 Hummer) at the University of North Carolina at Chapel Hill. Waves I-V data are from the Add  
372 Health Program Project, grant P01 HD31921 (Harris) from *Eunice Kennedy Shriver* National  
373 Institute of Child Health and Human Development (NICHD), with cooperative funding from 23  
374 other federal agencies and foundations. Add Health was designed by J. Richard Udry, Peter S.  
375 Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.  
376 **ALSPAC:** We are extremely grateful to all the families who took part in this study, the midwives  
377 for their help in recruiting them, and the whole ALSPAC team, which includes interviewers,  
378 computer and laboratory technicians, clerical workers, research scientists, volunteers,  
379 managers, receptionists, and nurses. The UK Medical Research Council and Wellcome (Grant  
380 ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This  
381 publication is the work of the authors, and Peter Barr and Danielle Dick will serve as guarantors  
382 for the contents of this paper. A comprehensive list of grants funding is available on the  
383 ALSPAC website ([http://www.bristol.ac.uk/alspac/external/documents/grant-](http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf)  
384 [acknowledgements.pdf](http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf)); This research was specifically funded by the Medical Research  
385 Council (MRC) under grants MR/L022206/1, MR/M006727/1, and G0800612/86812; the

## Running Header: Characterizing Risk for SUDs Across Cohorts

386 Wellcome Trust under grant 086684; and the National Institute on Alcohol Abuse and  
387 Alcoholism under 5R01AA018333-05. GWAS data was generated by Sample Logistics and  
388 Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of  
389 America) using support from 23andMe. **COGA:** We thank The Collaborative Study on the  
390 Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, T. Foroud;  
391 Scientific Director, A. Agrawal; Translational Director, D. Dick, includes eleven different centers:  
392 University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, T. Foroud, Y. Liu,  
393 M.H. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman, J. Kramer); SUNY  
394 Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan, A. Pandey);  
395 Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of  
396 California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, R. Hart, J. Salvatore);  
397 The Children's Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Virginia  
398 Commonwealth University (D. Dick); Icahn School of Medicine at Mount Sinai (A. Goate, P.  
399 Slesinger); and Howard University (D. Scott). Other COGA collaborators include: L. Bauer  
400 (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X., Xuei, D. Lai, S. O'Connor,  
401 (Indiana University); G. Chan (University of Iowa; University of Connecticut); D.B. Chorlian, J.  
402 Zhang, P. Barr, S. Kinreich, G. Pandey (SUNY Downstate); N. Mullins (Icahn School of  
403 Medicine at Mount Sinai); A. Anokhin, S. Hartz, E. Johnson, V. McCutcheon, S. Saccone  
404 (Washington University); J. Moore, Z. Pang, S. Kuo (Rutgers University); A. Merikangas (The  
405 Children's Hospital of Philadelphia and University of Pennsylvania); F. Aliev (Virginia  
406 Commonwealth University); H. Chin and A. Parsian are the NIAAA Staff Collaborators. We  
407 continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI  
408 and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA,  
409 including Ting- Kai Li, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their  
410 critical contributions. This national collaborative study is supported by NIH Grant U10AA008401

Running Header: Characterizing Risk for SUDs Across Cohorts

411 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute  
412 on Drug Abuse (NIDA). All code necessary to replicate this study is available upon request.

### 413 **ETHICS DECLARATIONS**

414 The authors have no conflicts of interest to declare.

415

## 416 REFERENCES

- 417 1. U.S. Overdose Deaths In 2021 Increased Half as Much as in 2020 - But Are Still Up 15%.  
418 [https://www.cdc.gov/nchs/pressroom/nchs\\_press\\_releases/2022/202205.htm](https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm).
- 419 2. Degenhardt, L. *et al.* The global burden of disease attributable to alcohol and drug use in  
420 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of  
421 Disease Study 2016. *The Lancet Psychiatry* **5**, 987–1012 (2018).
- 422 3. Sacks, J. J., Gonzales, K. R., Bouchery, E. E., Tomedi, L. E. & Brewer, R. D. 2010  
423 National and State Costs of Excessive Alcohol Consumption. *American Journal of*  
424 *Preventive Medicine* **49**, e73–e79 (2015).
- 425 4. National Drug Intelligence Center. *National drug threat assessment*. vol. 2019 (2011).
- 426 5. Verhulst, B., Neale, M. C. & Kendler, K. S. The heritability of alcohol use disorders: a  
427 meta-analysis of twin and adoption studies. *Psychol Med* **45**, 1061–1072 (2015).
- 428 6. Verweij, K. J. H. *et al.* Genetic and environmental influences on cannabis use initiation  
429 and problematic use: A meta-analysis of twin studies. *Addiction* **105**, 417–430 (2010).
- 430 7. Kendler, K. S., Jacobson, K. C., Prescott, C. A. & Neale, M. C. Specificity of Genetic and  
431 Environmental Risk Factors for Use and Abuse/Dependence of Cannabis, Cocaine,  
432 Hallucinogens, Sedatives, Stimulants, and Opiates in Male Twins. *American Journal of*  
433 *Psychiatry* **160**, 687–695 (2003).
- 434 8. Galea, S., Nandi, A. & Vlahov, D. The Social Epidemiology of Substance Use.  
435 *Epidemiologic Reviews* **26**, 36–52 (2004).
- 436 9. Barr, P. B. Neighborhood conditions and trajectories of alcohol use and misuse across  
437 the early life course. *Health and Place* **51**, 36–44 (2018).
- 438 10. Barr, P. B., Silberg, J., Dick, D. M. & Maes, H. H. Childhood socioeconomic status and  
439 longitudinal patterns of alcohol problems: Variation across etiological pathways in genetic  
440 risk. *Social Science and Medicine* **209**, 51–58 (2018).

## Running Header: Characterizing Risk for SUDs Across Cohorts

- 441 11. Meier, M. H. *et al.* Which adolescents develop persistent substance dependence in  
442 adulthood? Using population-representative longitudinal data to inform universal risk  
443 assessment. *Psychol Med* **46**, 877–889 (2016).
- 444 12. Schaefer, J. D. *et al.* Associations between polygenic risk of substance use and use  
445 disorder and alcohol, cannabis, and nicotine use in adolescence and young adulthood in  
446 a longitudinal twin study. *Psychological Medicine* 1–11 (2021)  
447 doi:10.1017/S0033291721004116.
- 448 13. Deak, J. D. *et al.* Alcohol and nicotine polygenic scores are associated with the  
449 development of alcohol and nicotine use problems from adolescence to young adulthood.  
450 *Addiction* **117**, 1117–1127 (2022).
- 451 14. Barr, P. B. *et al.* Using polygenic scores for identifying individuals at increased risk of  
452 substance use disorders in clinical and population samples. *Translational Psychiatry* **10**,  
453 196 (2020).
- 454 15. Kinreich, S. *et al.* Predicting risk for Alcohol Use Disorder using longitudinal data with  
455 multimodal biomarkers and family history: a machine learning study. *Molecular Psychiatry*  
456 **26**, 1133–1141 (2021).
- 457 16. Gu, F. *et al.* Combining common genetic variants and non-genetic risk factors to predict  
458 risk of cutaneous melanoma. *Human Molecular Genetics* (2018)  
459 doi:10.1093/hmg/ddy282.
- 460 17. O'Sullivan, J. W. *et al.* Combining clinical and polygenic risk improves stroke prediction  
461 among individuals with atrial fibrillation. *medRxiv* 2020.06.17.20134163 (2020)  
462 doi:10.1101/2020.06.17.20134163.
- 463 18. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV  
464 disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*  
465 **62**, 593 (2005).

Running Header: Characterizing Risk for SUDs Across Cohorts

- 466 19. Harris, K. M., Halpern, C. T., Haberstick, B. C. & Smolen, A. The National Longitudinal  
467 Study of Adolescent Health (Add Health) sibling pairs data. *Twin Research and Human*  
468 *Genetics* **16**, 391–398 (2013).
- 469 20. Boyd, A. *et al.* Cohort profile: The 'Children of the 90s'-The index offspring of the avon  
470 longitudinal study of parents and children. *International Journal of Epidemiology* (2013)  
471 doi:10.1093/ije/dys064.
- 472 21. Fraser, A. *et al.* Cohort profile: The avon longitudinal study of parents and children:  
473 ALSPAC mothers cohort. *International Journal of Epidemiology* **42**, 97–110 (2013).
- 474 22. Harris, P. A. *et al.* Research electronic data capture (REDCap)-A metadata-driven  
475 methodology and workflow process for providing translational research informatics  
476 support. *Journal of Biomedical Informatics* **42**, 377–381 (2009).
- 477 23. Northstone, K. *et al.* The Avon Longitudinal Study of Parents and Children (ALSPAC): an  
478 update on the enrolled sample of index children in 2019 [version 1; peer review: 2  
479 approved]. *Wellcome Open Research* **4**, (2019).
- 480 24. Edenberg, H. J. The collaborative study on the genetics of alcoholism: An update. *Alcohol*  
481 *Research and Health* **26**, 214–218 (2002).
- 482 25. Begleiter, H. The Collaborative Study on the Genetics of Alcoholism. *Alcohol Health and*  
483 *Research World* **19**, 228 (1995).
- 484 26. Bucholz, K. K. *et al.* Comparison of Parent, Peer, Psychiatric, and Cannabis Use  
485 Influences Across Stages of Offspring Alcohol Involvement: Evidence from the COGA  
486 Prospective Study. *Alcoholism: Clinical and Experimental Research* **41**, 359–368 (2017).
- 487 27. Rose, R. J. R. J. *et al.* *FinnTwin12 Cohort: An Updated Review*. *Twin Research and*  
488 *Human Genetics* vol. 22 (2019).
- 489 28. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. & Fagerstrom, K. O. The Fagerstrom  
490 Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br*  
491 *J Addict* **86**, 1119–1127 (1991).

Running Header: Characterizing Risk for SUDs Across Cohorts

- 492 29. Hughes, K. *et al.* The effect of multiple adverse childhood experiences on health: a  
493 systematic review and meta-analysis. *The Lancet Public Health* (2017)  
494 doi:10.1016/S2468-2667(17)30118-4.
- 495 30. Sher, K. J., Grekin, E. R. & Williams, N. A. The development of alcohol use disorders.  
496 *Annual Review of Clinical Psychology* **1**, 493–523 (2005).
- 497 31. Karlsson Linner, R. *et al.* Multivariate genomic analysis of 1.5 million people identifies  
498 genes related to addiction, antisocial behavior, and health. *Nature Neuroscience*.
- 499 32. Levey, D. F. *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and  
500 meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nature*  
501 *Neuroscience* (2021) doi:10.1038/s41593-021-00860-2.
- 502 33. Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563  
503 individuals yields insights into biology and relationships with other traits. *Nature*  
504 *Neuroscience* (2020) doi:10.1038/s41593-020-0643-5.
- 505 34. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the  
506 genetic etiology of tobacco and alcohol use. *Nature Genetics* **51**, 237–244 (2019).
- 507 35. Kranzler, H. R. *et al.* Genome-wide association study of alcohol consumption and use  
508 disorder in 274,424 individuals from multiple populations. *Nature Communications* **10**,  
509 1499 (2019).
- 510 36. Quach, B. C. *et al.* Expanding the genetic architecture of nicotine dependence and its  
511 shared genetics with multiple traits. *Nature Communications* **11**, (2020).
- 512 37. Trubetsky, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in  
513 schizophrenia. *Nature* **2022** 1–13 (2022) doi:10.1038/s41586-022-04434-5.
- 514 38. Bigdeli, T. B. *et al.* Genome-Wide Association Studies of Schizophrenia and Bipolar  
515 Disorder in a Diverse Cohort of US Veterans. *Schizophrenia Bulletin* (2020)  
516 doi:10.1093/schbul/sbaa133.

## Running Header: Characterizing Risk for SUDs Across Cohorts

- 517 39. Barr, P. B. & Dick, D. M. The Genetics of Externalizing Problems. *Current Topics in*  
518 *Behavioral Neurosciences* **47**, 93–112 (2020).
- 519 40. Krueger, R. F. *et al.* Etiological connections among substance dependence, antisocial  
520 behavior and personality: Modeling the externalizing spectrum. *J Abnorm Psychol* **111**,  
521 411–424 (2002).
- 522 41. Kendler, K. S. & Myers, J. The boundaries of the internalizing and externalizing genetic  
523 spectra in men and women. *Psychological Medicine* **44**, 647–655 (2014).
- 524 42. Polimanti, R. *et al.* Evidence of causal effect of major depression on alcohol dependence:  
525 Findings from the psychiatric genomics consortium. *Psychological Medicine* (2019)  
526 doi:10.1017/S0033291719000667.
- 527 43. Johnson, E. C. *et al.* A large-scale genome-wide association study meta-analysis of  
528 cannabis use disorder. *The Lancet Psychiatry* (2020) doi:10.1016/S2215-0366(20)30339-  
529 4.
- 530 44. Zhou, H. *et al.* Association of OPRM1 Functional Coding Variant With Opioid Use  
531 Disorder: A Genome-Wide Association Study. *JAMA Psychiatry* (2020)  
532 doi:10.1001/jamapsychiatry.2020.1206.
- 533 45. Kendler, K. S., Gardner, C. & Dick, D. M. Predicting alcohol consumption in adolescence  
534 from alcohol-specific and general externalizing genetic risk factors, key environmental  
535 exposures and their interaction. *Psychol Med* **41**, 1507–1516 (2011).
- 536 46. Meyers, J. L. *et al.* Genetic Influences on Alcohol Use Behaviors Have Diverging  
537 Developmental Trajectories: A Prospective Study Among Male and Female Twins.  
538 *Alcoholism: Clinical and Experimental Research* **38**, 2869–2877 (2014).
- 539 47. Barr, P. B. *et al.* Parsing Genetically Influenced Risk Pathways: Genetic Loci Impact  
540 Problematic Alcohol Use Via Externalizing and Specific Risk. *medRxiv*  
541 2021.07.20.21260861 (2021) doi:10.1101/2021.07.20.21260861.

## Running Header: Characterizing Risk for SUDs Across Cohorts

- 542 48. Sanchez-Roige, S., Palmer, A. A. & Clarke, T. K. Recent Efforts to Dissect the Genetic  
543 Basis of Alcohol Use and Abuse. *Biological Psychiatry* (2020)  
544 doi:10.1016/j.biopsych.2019.09.011.
- 545 49. Walters, R. K. *et al.* Trans-ancestral GWAS of alcohol dependence reveals common  
546 genetic underpinnings with psychiatric disorders. *Nature Neuroscience* **21**, 1656–1669  
547 (2018).
- 548 50. Dick, D. M., Barr, P., Guy, M., Nasim, A. & Scott, D. Review: Genetic research on alcohol  
549 use outcomes in African American populations: A review of the literature, associated  
550 challenges, and implications. *American Journal on Addictions* vol. 26 486–493 (2017).
- 551 51. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies.  
552 *Communications Biology* **2**, 9 (2019).
- 553 52. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across  
554 Diverse Populations. *Am J Hum Genet* **100**, 635–649 (2017).
- 555 53. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse  
556 human populations. *Nature Communications* (2019) doi:10.1038/s41467-019-11112-0.
- 557 54. Ruan, Y. *et al.* Improving Polygenic Prediction in Ancestrally Diverse Populations. *Nature*  
558 *Genetics* (2022) doi:10.1038/s41588-022-01054-7.
- 559 55. Curran, P. J. & Hussong, A. M. Integrative Data Analysis: The Simultaneous Analysis of  
560 Multiple Data Sets. *Psychological Methods* **14**, 81–100 (2009).
- 561 56. Cameron, C. A., Gelbach, J. B. & Miller, D. L. Robust inference with multiway clustering.  
562 *Journal of Business and Economic Statistics* **29**, 238–249 (2011).
- 563 57. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination.  
564 *Biometrika* **78**, 691–692 (1991).
- 565 58. Hasin, D. S. & Grant, B. F. The National Epidemiologic Survey on Alcohol and Related  
566 Conditions (NESARC) Waves 1 and 2: review and summary of findings. *Social Psychiatry*  
567 *and Psychiatric Epidemiology* **50**, 1609–1640 (2015).

## Running Header: Characterizing Risk for SUDs Across Cohorts

- 568 59. Conrod, P. J. *et al.* Effectiveness of a Selective, Personality-Targeted Prevention  
569 Program for Adolescent Alcohol Use and Misuse: A Cluster Randomized Controlled Trial.  
570 *JAMA Psychiatry* **70**, 334–342 (2013).
- 571 60. Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* (1979) **359**,  
572 424–428 (2018).
- 573 61. Burt, S. A. Are there meaningful etiological differences within antisocial behavior? Results  
574 of a meta-analysis. *Clin Psychol Rev* **29**, 163–178 (2009).
- 575 62. Rhee, S. H. & Waldman, I. D. Genetic and environmental influences on antisocial  
576 behavior: A meta-analysis of twin and adoption studies. *Psychol Bull* **128**, 490–529  
577 (2002).
- 578 63. Garg, A., Boynton-Jarrett, R. & Dworkin, P. H. Avoiding the Unintended Consequences of  
579 Screening for Social Determinants of Health. *JAMA* **316**, 813–814 (2016).
- 580 64. Davidson, K. W. & McGinn, T. Screening for Social Determinants of Health: The Known  
581 and Unknown. *JAMA* **322**, 1037–1038 (2019).
- 582 65. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health  
583 disparities. *Nature Genetics* **51**, 584–591 (2019).
- 584 66. Williams, D. R., Mohammed, S. A., Leavell, J. & Collins, C. Race, socioeconomic status,  
585 and health: complexities, ongoing challenges, and research opportunities. *Ann N Y Acad*  
586 *Sci* **1186**, 69–101 (2010).

587

## 588 DATA AVAILABILITY

589 All data sources are described in the manuscript and supplemental information. No new  
590 data were collected. Only data from existing studies or study cohorts were analyzed, some of  
591 which have restricted access to protect the privacy of the study participants. Add Health genetic

## Running Header: Characterizing Risk for SUDs Across Cohorts

592 data obtained through dbGaP (Study Accession: phs001367.v1.p1). Instructions on gaining  
593 access to Add Health restricted use data can be found at:  
594 <https://data.cpc.unc.edu/projects/2/view>. COGA genetic data available through dbGaP (Study  
595 Accession: phs000763.v1.p1). Instructions for access to ALSPAC data available at:  
596 <http://www.bristol.ac.uk/alspac/researchers/access/>. The process for obtaining the GWAS  
597 summary statistics used in these analyses are described in the corresponding original GWAS  
598 publications.

## 599 **CODE AVAILABILITY**

600 No custom algorithms or software was developed in this study. All code is available by  
601 request from the corresponding author. Polygenic scores generated using PRS-CSx  
602 (<https://github.com/getian107/PRScsx>). All primary analyses completed in R 4.1.0 using the  
603 *data.table* (1.14.0), *pROC* (1.18.0), *lme4* (1.1-27.1), and base packages.

## Running Header: Characterizing Risk for SUDs Across Cohorts

604

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

Measure	Definition
1) Low childhood SES	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.
2) Family history of SUD	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.
3) Childhood externalizing problems	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.
4) Childhood internalizing problems	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.
5) Early initiation of substance use	Respondent reports age of first whole alcoholic drink, smoked whole cigarette, or tried cannabis before the age of 15.
6) Adolescent alcohol use	Frequency of self-reported use 5 or more days per week at age 18 and below.
7) Adolescent tobacco use	Frequency of self-reported use at daily use at age 18 and below.
8) Adolescent cannabis use	Frequency of self-reported use 5 or more days per week at age 18 and below.
9) Peer substance use	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.
10) Traumatic events	Respondent reports exposure to any traumatic event.

605 Full description of sample specific definitions available in the supplementary information.

**Table 2: Prevalence of SUDs and CERI by Cohort**

	<i>Add Health</i> AFR (N = 1,605)*		<i>Add Health</i> EUR (N = 4,855)*		<i>ALSPAC</i> EUR (N = 4,733)*		<i>COGA</i> AFR (N = 870)*		<i>COGA</i> EUR (N = 1,878)*		<i>FinnTwin12</i> EUR (N = 1,193)*	
	<u>Mean (SD)/%</u>		<u>Mean (SD)/%</u>		<u>Mean (SD)/%</u>		<u>Mean (SD)/%</u>		<u>Mean (SD)/%</u>		<u>Mean (SD)/%</u>	
Female	55.26%	-	53.59%	-	56.71%	-	51.38%	-	51.33%	-	53.73%	-
Age (at last observation)	28.89	(1.69)	28.84	(1.70)	22.47	(2.20)	24.13	(5.12)	24.24	(5.26)	22.44	(0.72)
Alcohol dependence	3.93%	-	12.75%	-	5.92%	-	11.49%	-	21.14%	-	8.55%	-
Nicotine dependence	2.74%	-	10.28%	-	1.54%	-	3.91%	-	7.83%	-	2.26%	-
Drug dependence	6.73%	-	10.79%	-	0.78%	-	26.44%	-	23.59%	-	1.34%	-
Any substance dependence <sup>†</sup>	11.21%	-	25.81%	-	8.87%	-	30.69%	-	34.66%	-	10.98%	-
CERI	1.95	(1.48)	2.07	(1.65)	2.08	(1.19)	3.98	(2.24)	3.65	(2.38)	2.62	(1.27)

\* Available samples with genotypic, phenotypic, and environmental risk data

<sup>†</sup> Any substance dependence includes those who meet criteria for alcohol, nicotine, or drug dependence.

AFR = African ancestries; EUR = European ancestries; CERI = clinical/environmental risk index

606  
607  
608  
609

**Table 3: Estimates for PGS Only, CERI Only, and Combined Models**

		Alcohol Dependence			Nicotine Dependence			Drug Dependence			Any substance dependence		
		OR	95% CI	CI	OR	95% CI	CI	OR	95% CI	CI	OR	95% CI	CI
PGS Only Model*	ALCC PGS	1.05	(0.99, 1.11)		0.96	(0.89, 1.04)		1.05	(0.98, 1.12)		1.00	(0.96, 1.05)	
	ALCP PGS	<b>1.13</b>	<b>(1.06, 1.20)</b>		1.01	(0.93, 1.10)		1.07	(1.00, 1.15)		<b>1.10</b>	<b>(1.05, 1.16)</b>	
	EXT PGS	<b>1.18</b>	<b>(1.11, 1.26)</b>		<b>1.50</b>	<b>(1.38, 1.63)</b>		<b>1.27</b>	<b>(1.19, 1.36)</b>		<b>1.31</b>	<b>(1.25, 1.38)</b>	
	MDD PGS	1.00	(0.94, 1.06)		1.06	(0.98, 1.15)		1.08	(1.02, 1.15)		1.02	(0.98, 1.07)	
	SCZ PGS	1.04	(0.97, 1.10)		0.98	(0.90, 1.06)		1.03	(0.96, 1.11)		1.00	(0.96, 1.05)	
	CPD PGS	1.00	(0.94, 1.06)		<b>1.33</b>	<b>(1.24, 1.43)</b>		1.01	(0.95, 1.08)		1.08	(1.03, 1.13)	
	$\Delta Pseudo-R^2$			0.011			0.037			0.014			0.022
CERI Only Model*	CERI	<b>1.37</b>	<b>(1.33, 1.41)</b>		<b>1.63</b>	<b>(1.57, 1.70)</b>		<b>1.67</b>	<b>(1.61, 1.72)</b>		<b>1.58</b>	<b>(1.54, 1.63)</b>	
$\Delta Pseudo-R^2$			0.054			0.107			0.129			0.120	
Combined Model*	CERI	<b>1.35</b>	<b>(1.31, 1.40)</b>		<b>1.58</b>	<b>(1.52, 1.65)</b>		<b>1.65</b>	<b>(1.59, 1.70)</b>		<b>1.55</b>	<b>(1.51, 1.60)</b>	
	ALCC PGS	1.04	(0.97, 1.10)		0.94	(0.87, 1.03)		1.03	(0.96, 1.11)		0.99	(0.94, 1.04)	
	ALCP PGS	<b>1.12</b>	<b>(1.05, 1.19)</b>		0.99	(0.91, 1.08)		1.06	(0.98, 1.14)		<b>1.09</b>	<b>(1.04, 1.15)</b>	
	EXT PGS	1.08	(1.01, 1.15)		<b>1.33</b>	<b>(1.22, 1.45)</b>		<b>1.11</b>	<b>(1.03, 1.20)</b>		<b>1.18</b>	<b>(1.12, 1.24)</b>	
	MDD PGS	0.97	(0.91, 1.03)		1.02	(0.94, 1.10)		1.03	(0.96, 1.10)		0.98	(0.93, 1.03)	
	SCZ PGS	1.03	(0.97, 1.10)		0.96	(0.88, 1.05)		1.01	(0.94, 1.08)		1.00	(0.95, 1.05)	
	CPD PGS	0.98	(0.92, 1.04)		<b>1.31</b>	<b>(1.22, 1.42)</b>		0.98	(0.92, 1.04)		1.06	(1.01, 1.11)	
$\Delta Pseudo-R^2$			0.059			0.126			0.131			0.128	

\* All models included age, sex, and cohort as covariates. See Supplementary Table 7 for all parameter estimates. PGS residualized on age, sex, and first 10 ancestral principal components.

Bolded estimates =  $p < .05$  after correction for multiple testing ( $p < .05/4 = 0.0125$ )

$\Delta Pseudo-R^2$  denotes pseudo- $R^2$  above model including age, sex, and cohort. CI = confidence interval; PGS = polygenic score; CERI = clinical/environmental risk index

## 610 **FIGURE CAPTIONS**

### 611 *Figure 1: SUD Prevalence Across Genetic and Environmental Risk Factors*

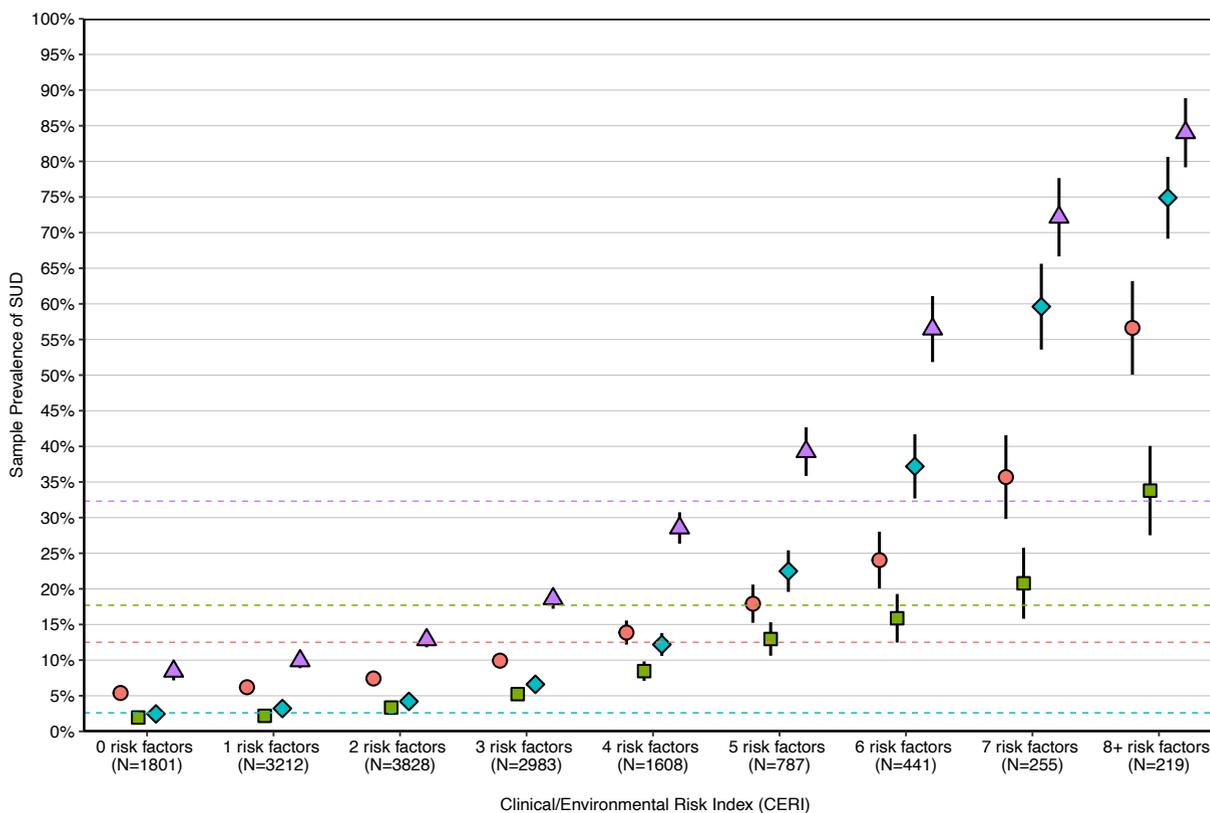
612 Panel A: Prevalence (and 95% confidence intervals) of those who meet criteria for alcohol,  
613 nicotine, drug, or any substance dependence across counts for items in the risk index. Panel B:  
614 Prevalence (and 95% confidence intervals) of those who meet criteria for alcohol, nicotine, drug,  
615 or any substance dependence across four categories: 1) those below the 90<sup>th</sup> percentile for all  
616 PGS and the CERI; 2) those at or above the 90<sup>th</sup> percentile for the CERI; 3) those at or above  
617 the 90<sup>th</sup> percentile for all PGS; and 4) those at or above the 90<sup>th</sup> percentile for both the CERI  
618 and PGS. PGS and risk index were first residualized on sex, age, age<sup>2</sup>, cohort, sex\*age,  
619 sex\*age<sup>2</sup>, sex\*cohort, cohort\*age, cohort\*age<sup>2</sup>, sex\*cohort\*age, and sex\*cohort\*age<sup>2</sup>. Dotted  
620 colored lines represent corresponding lifetime prevalence estimates for alcohol dependence  
621 (red), nicotine dependence (green), drug dependence (blue), and any substance use disorder  
622 (purple) from nationally representative data <sup>58</sup>.

623

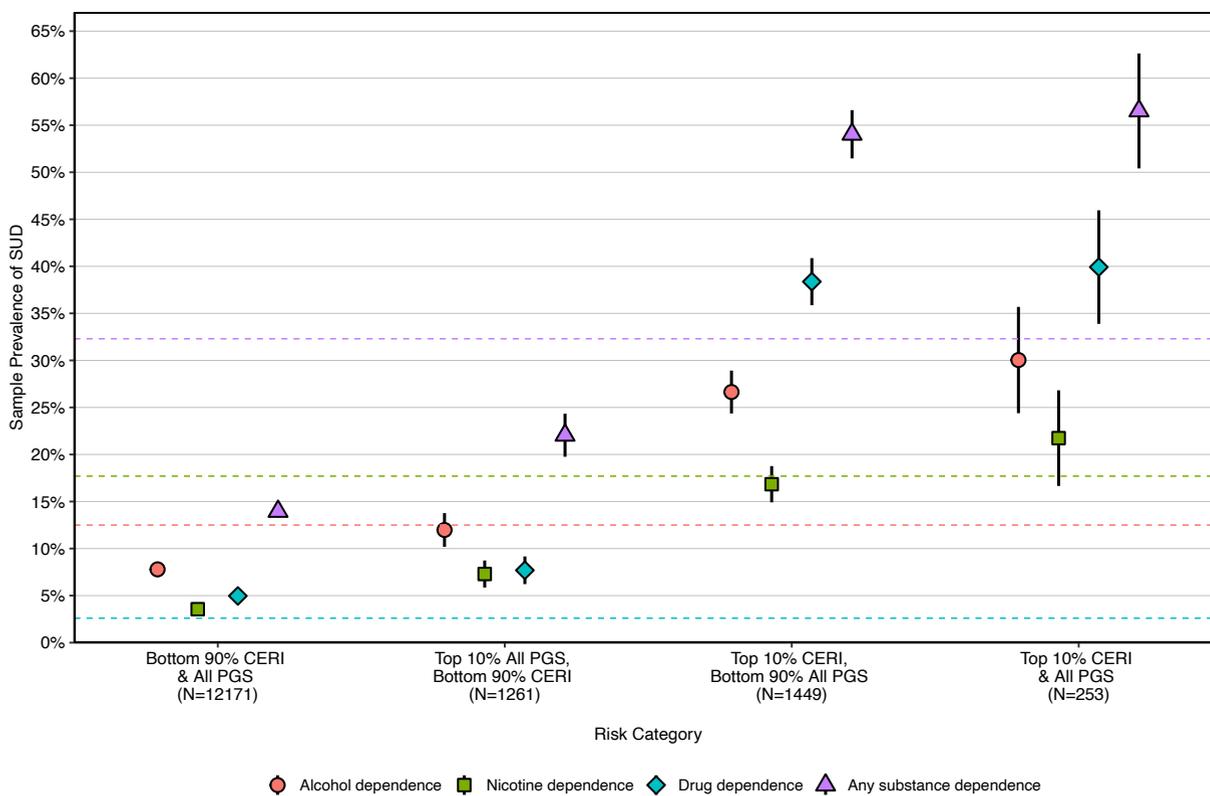
### 624 *Figure 2: ROC Curves for Combined and Baseline Models*

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

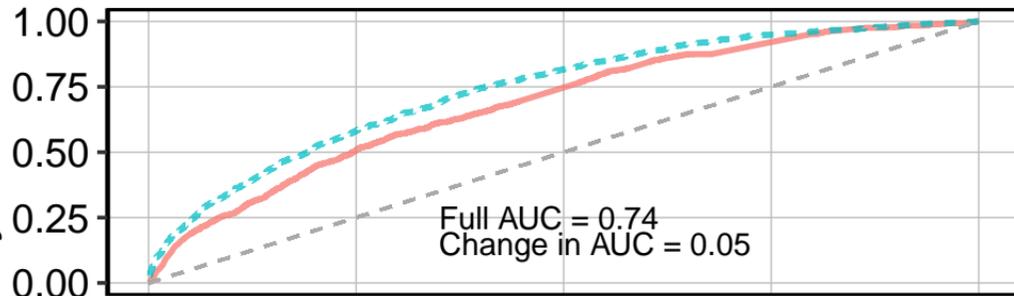
A



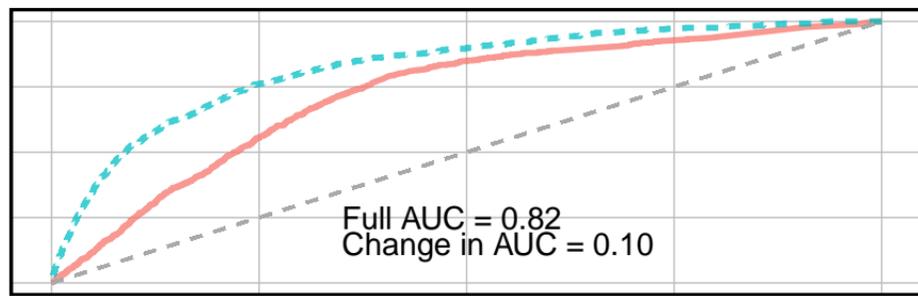
B



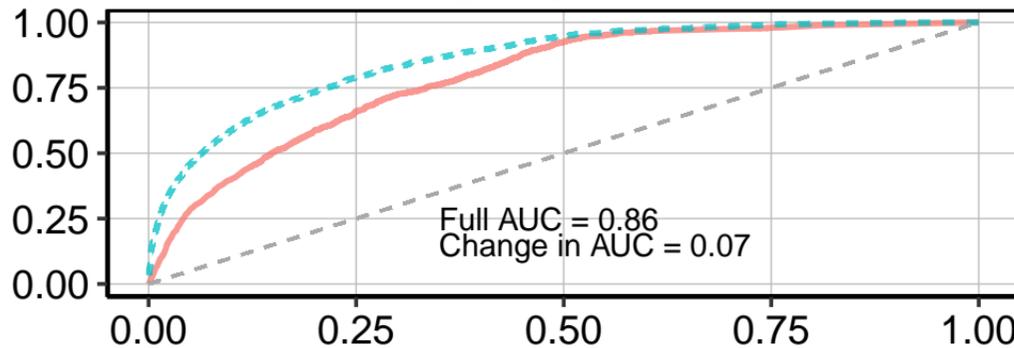
### Alcohol Dependence



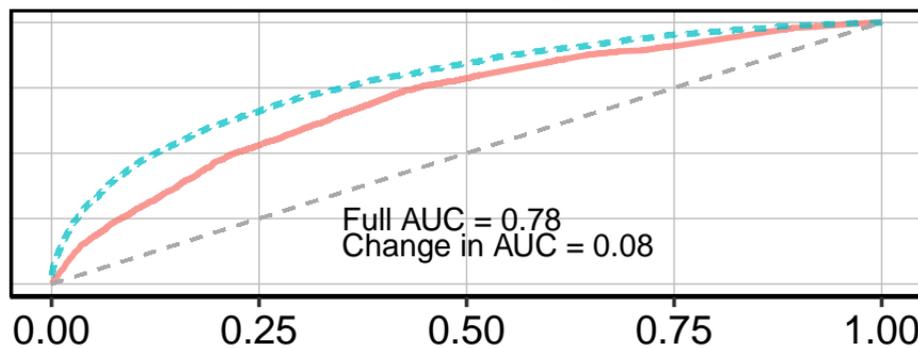
### Nicotine Dependence



### Drug Dependence



### Any Substance Dependence



1 - Specificity

Model — Base - - Full