Association between polygenic risk for tobacco or alcohol consumption and liability to licit and illicit substance use in young Australian adults

Lun-Hsien Chang, Baptiste Couvy-Duchesne, Mengzhen Liu, Sarah E. Medland, Brad Verhulst, Eric G. Benotsch, Ian B. Hickie, Nicholas G. Martin, Nathan A. Gillespie, GSCAN Consortium

ARTICLE INFO

Keywords:
Polygenic risk
Addiction
Genetics
Twins

ABSTRACT

Background: Co-morbid substance use is very common. Despite a historical focus using genetic epidemiology to investigate comorbid substance use and misuse, few studies have examined substance-substance associations using polygenic risk score (PRS) methods.

Methods: Using summary statistics from the largest substance use GWAS to date (258,797 - 632,802 subjects), GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN), we constructed PRSs for smoking initiation (PRS-SI), age of initiation of regular smoking (PRS-AI), cigarettes per day (PRS-CPD), smoking cessation (PRS-SC), and drinks per week (PRS-DPW). We then estimated the fixed effect of individual PRSs on 22 lifetimesubstance use and substance use disorder phenotypes collected in an independent sample of 2463 young Australian adults using genetic restricted maximum likelihood (GREML) in Genome-wide Complex Trait Analysis (GCTA), separately in females, males and both sexes together.

Results: After accounting for multiple testing, PRS-SI significantly explained variation in the risk of cocaine (0.67%), amphetamine (1.54%), hallucinogens (0.72%), ecstasy (1.66%) and cannabis initiation (0.97%), as well as DSM-5 alcohol use disorder (0.72%). PRS-DPW explained 0.75%, 0.59% and 0.90% of the variation of cocaine, amphetamine and ecstasy initiation respectively. None of the 22 phenotypes including emergent classes of substance use were significantly predicted by PRS-AI, PRS-CPD, and PRS-SC.

Conclusions: To our knowledge, this is the first study to report significant genetic overlap between the polygenic risks for smoking initiation and alcohol consumption and the risk of initiating major classes of illicit substances. PRSs constructed from large discovery GWASs allows the detection of novel genetic associations.

1. Introduction

1.1. Genetic Influence on Substance Use

Based on non-molecular but genetically informative family and twin studies, it is widely accepted that licit and illicit substance use (SU) and substance use disorders (SUDs) are polygenic with 30–80% heritability estimated across different substances as summarised by a comprehensive review (Kendler et al., 2012b). Twin studies have also shown that the genetic covariance between major licit and illicit substances use disorders can be explained by common genetic risks (Kendler et al., 2003, 2012a), with some evidence suggesting that highly correlated \( r_G = 0.82 \) genetic risks underpinning legal (nicotine, alcohol and caffeine) and illegal (cannabis and cocaine) SUDs (Kendler et al., 2007). Recent meta-analyses of genome-wide association studies (GWAS) provide evidence that SU and SUDs support the conclusion these are polygenic genetic behaviours (Minica et al., 2018, Pasman et al., 2018, Prom-Wormley et al., 2017, Walters et al., 2018) with multiple genes of small effect contributing to the variation in the risk. Despite their complex polygenicity (Pan et al., 2013, Robinson et al., 2014) the direct

https://doi.org/10.1016/j.drugalcdep.2019.01.015

Received 6 October 2018; Received in revised form 25 December 2018; Accepted 19 January 2019

Available online 16 February 2019

0376-8716/ © 2019 Elsevier B.V. All rights reserved.
effects of individual SU and SUD loci and genes are becoming increasingly clear. Multiple loci have been associated with liability to consuming nicotine (Stephens et al., 2013; Thorgeirsson et al., 2010; Tobacco and Genetics Consortium, 2010), alcohol (Agrawal, 1997; Agrawal et al., 2011; Bierut et al., 2012, 2010; Chen et al., 1999; Edenberg, 2007; Edenberg et al., 2010; Edenberg and Foroud, 2006; Ehlers et al., 2004; Frank et al., 2012; Guindalini et al., 2005; Hasin et al., 2002a, 2002b; Heath et al., 2011; Kendler et al., 2011; Lind et al., 2010, 2008; Luo et al., 2006; Macgregor et al., 2009; Moore et al., 2007; Schumann et al., 2011; Strat et al., 2008; Treutlein et al., 2009; Treutlein and Rietschel, 2011; Walters et al., 2018; Wang et al., 2012; Zuo et al., 2012, 2011), and cannabis (Minica et al., 2018; Pasman et al., 2018; Sherva et al., 2016; Stringer et al., 2016), and SNPs related to multi-substance use (Sherva et al., 2010) have now been identified. Given the availability of very large, informative GWAS summary statistics there now exists the opportunity to examine the degree to which the variation of SU and SUD risk can be explained by emerging allelic risks as opposed to statistical inference based on twin and family studies.

1.2. Polygenic Risk of Substance Use

A key implication of the recent molecular findings is that a complete list of true replicable signals is not required for GWAS to demonstrate significant concurrent or predictive criterion validity. The upper tails of well-powered GWAS summary test distributions are expected to be highly enriched with many true signals (The International Schizophrenia Consortium, 2009) that can be used to estimate individualized polygenic risk score for complex behaviours including SU and SUDs.

Polygenic risk score (PRS) analysis aggregates the effects of thousands of genetic variants that are associated with a trait using a spectrum of significance levels. A PRS is calculated as a weighted sum of the number of risk alleles at the selected SNPs that are carried by an individual. The weight is obtained from the effect size (e.g. beta for a continuous trait; log transformation of the odds ratio for a binary trait) associated with the SNPs. These scores allow us to compare and correlate PRSs between different individuals or PRSs for different phenotypes.

A common problem in PRS methodology is that a PRS typically explains a small proportion, in comparison to twin estimates (Kendler et al., 2012b) of the variation of a target phenotype, which may be improved with larger-sized discovery sample from which the PRS is calculated. The criterion validity of PRS to predict a variety of important behavioural outcomes has clearly been demonstrated. For example, PRS based on much smaller discovery samples already predict complex traits (Agerbo et al., 2015; Clarke et al., 2015; Jervis et al., 2015; Moor et al., 2015) including cannabis use and cannabis use frequency (Power et al., 2014). A number of reports have also linked PRS for psychiatric disorders to SU and SUDs outcomes (Carey et al., 2016; Du Rietz et al., 2017; Hartz et al., 2017; Reginsson et al., 2017; Verweij et al., 2017). However, very few studies (Verweij et al., 2017; Vink et al., 2014) have investigated the associations between PRSs for SU and SUDs, and none of these studies have leveraged the most recent and largest SU GWAS to date. With PRSs derived from a very large discovery sample GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN), the current study is likely to maximise the power to detect an association between PRSs and SU and SUDs.

1.3. Aims

The aim of this study was to establish the association between polygenic risks for licit substance use and the risk of self-report illicit substance use, alcohol use disorders (AUD) as well as measures of cannabis misuse in an Australian sample of young adults.

2. Materials and Methods

Our overall aim was to calculate the PRSs for five measures of tobacco and alcohol consumption, and then test their associations with self-reported illicit SU phenotypes while appropriately adjusting for multiple testing. PRSs are commonly calculated using independent SNPs that meet different association p-value thresholds, which allows us calculating PRS from SNPs that did not reach genome-wide significance.

2.1. Target Sample and Measures

Self-report measures of SU and SUDs were obtained from a sample of young adult Australian twins and their families who participated in the Brisbane Longitudinal Twin Study 19Up Project between 2009 and 2016 (Couvy-Duchesne et al., 2018; Gillespie et al., 2013) in Brisbane, Australia. Twins and their families were invited to participate in this project when they turned 19 or older. The 19Up study was approved by the QIMR Human Research Ethics Committee. Data were stored in compliance with national regulations regarding personal data protection. Informed consent was obtained from all the participants.

Our target sample comprised 2463 individuals who were genotyped (mean age: 26.1 years; age range: 18.7–38.6) at QIMR Berghofer. This included 1977 twin individuals (835 twin pairs), and 486 siblings from 1163 families. The Composite International Diagnostic Interview (CIDI; Kessler and Üstün, 2004) was used to identify Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition, DSM-IV) diagnoses of lifetime use of illicit substance (cannabis, amphetamine, inhalants, sedatives, hallucinogens, opioids, ecstasy, prescription painkillers, prescription stimulants, and cannabis), alcohol abuse, alcohol dependence, cannabis abuse, and cannabis dependence. The prevalence of the ten binary initiation phenotypes ranged from 5% for lifetime inhalants to 47% for lifetime cannabis use (Table 1). For the four abuse and dependence phenotypes, the prevalence was 7% to 33%. Significantly different prevalence was seen between males and females in these 14 phenotypes, except for sedative, opioid and painkiller initiation (Table 1). The DSM-5 AUD and CUD outcomes (0 = absent, 1 = mild, 2 = moderate, and 3 = severe) were constructed based on the number of symptoms identified, with two or three symptoms indicating a mild SUD, four or five symptoms indicating a moderate SUD, and six or more symptoms indicating a severe SUD. In addition to these two four-point scales, we derived binary phenotypes by dichotomising those who scored 0 or 1 and those who score 2 or 3.

2.2. Discovery Sample

We used GWAS summary statistics from the global ‘GSCAN’ consortium (Liu et al., 2019) with the QIMR sample removed prior to constructing PRSs for four smoking and one alcohol-related discovery phenotypes: smoking initiation (SI) (N = 631,564, 52% smokers, 53.6% females); age of initiation of regular smoking (AI) (N = 258,251, 50.0% females); cigarettes per day (CPD) (N = 258,999, 55.1% females); smoking cessation (SC) (N = 312,273, 40% current smokers, 50.6% females); and drinks per week (DPW) (N = 527,402, 53.2% females).

2.3. Genetic Correlations Between Discovery Phenotypes

Genetic correlations (\(r_G\)) between discovery phenotypes were estimated using cross-trait LD score correlation (LDSC; Bulik-Sullivan et al., 2015). This method is suitable for our study as it requires only GWAS summary statistics, is not biased by sample overlap and estimates of \(r_G\) between a binary and continuous trait without having to specify a scale. We merged each of the five GWAS files with the HapMap3 SNP list, converted the summary statistics into the LDSC format, and calculated...
and denote the random genetic effect and error term respectively. Next, we ensured QC for SNPs in the target BLTS cohort sample by retaining SNPs with minor allele frequencies between 0.01 and 0.99 (an indication of imputation by retaining SNPs with more than one occurrence, keeping SNPs that begin with “rs” in their variant IDs, removing SNPs from ambiguous genotypes across all the chromosomal blocks of 22 autosomes, removing SNPs with minor allele frequencies between 0.01 and 0.99 and with genotype imputation \( R^2 \geq 0.6 \) (an indication of imputation quality across different platforms), and applying the same criteria as the QC used for the discovery sample SNPs. Thirdly, an inner join was then performed to obtain SNPs that were common to the discovery and target SNPs using chromosome number and base pair position (CHR: BP) as the merging key. Fourthly, we accounted for linkage disequilibrium (LD) by selecting one representative SNP per haplotype, a process known as LD-based SNP clumping. Specifically, we set the R-squared threshold (plink–clump-r2) to 0.1, and the distance threshold (plink–clump-kb) to 10,000 kb. The number of SNPs that met our eight p-value thresholds of 5e-08, 1e-05, 1e-03, 1e-02, 5e-02, 0.1, 0.5 and 1 was compared before (see QC in Table S1) and after (see LD in Table S1) clumping based on their linkage disequilibrium (LD) in individual autosomes. Across all of these autosomes (see Table in Table S1), less than 1% of SNPs were selected through LD-based clumping. The difference in the number of clumped SNPs tended to be small between the five discovery phenotypes at higher p-value thresholds, with 400,877–403,341 clumped SNPs at p-value < 1e-03, 280,074–281,387 clumped SNPs at p < 0.5, 86,672–91,667 clumped SNPs at p < 0.1, 49,797–55,505 clumped SNPs at p < 5e-02, 13,235–18,428 clumped SNPs at p < 1e-02. The PRS-SI at a p-value threshold lower than 1e-03 had nearly twice the number of clumped SNPs compared to the four other discovery phenotypes constructed with the same p-value. Fifthly, we constructed PRSs from imputed genotype dosage by summing up allelic scores of the clumped SNPs across all the chromosomal blocks of 22 autosomes, resulting in one PRS per p-value threshold and individual in the target sample:

\[
\text{PRS}_i = \sum_{j=1}^{n} \beta_j \times G_{ij}
\]

where \( i \) is an individual, \( \beta \) is the effect size of an independent SNP, \( G_{ij} \) is the number of risk alleles at the SNP, in the individual \( i \). The number of independent SNPs, denoted as \( n \), varies with the discovery phenotypes and p-value thresholds. Prior to the PRS calculation, the genotypes had been imputed on the Michigan Imputation Server (Das et al., 2016) using the Haplotype Reference Consortium (McCarthy et al., 2016) version r1.1 as the reference panel and then converted to plink dosage format at QIMR. Lastly, we standardised the PRSs and merged them back into the QIMR cohort who were assessed with SU and SUDs.

### 2.5. Univariate Mixed Model That Models Familial and Cryptic Relatedness

To estimate the proportion of variance in each target phenotype that was explained by the PRS, we performed linear mixed modeling using genetic restricted maximum likelihood (GREML) by the software GCTA (Yang et al., 2011):

\[
Y = X \beta + g + \varepsilon
\]

where \( Y \) is a n x 1 vector of either a binary phenotype or continuous phenotype with \( n \) being the number of individuals in the input data, \( X \) is a vector of fixed effect covariates, and \( \beta \) is the effect estimate of the \( X \). Considering the predictive power of our PRSs may be confounded by sex, we conducted the mixed modelling separately for males (M), females (F) and both sexes combined (F + M). The covariates used in the both-sex model included categorical covariates of study wave, sex, GWAS array, and quantitative covariates, such as a single PRS, age, age², age x sex, age² x sex, and the first ten principal components derived from the SNP data. The two single-sex models used the same covariates except for sex, age x sex, and age² x sex.

The \( g \) and \( \varepsilon \) denote the random genetic effect and error term respectively. The genetic effect has a known variance-covariance structure that is defined by the genetic relationship matrix (GRM), which estimates the genetic relatedness between individuals using SNP data. Although the GCTA-GREML is developed mainly for estimating SNP heritability, this method provides additional functionality of estimating fixed effect parameters in a sample of related individuals. Our fixed effects were estimated using the GCTA –reml-est-fix option. We then derived an estimate of the association expressed as a proportion of the target phenotype (both binary and continuous) variance explained by

---

**Table 1**

<table>
<thead>
<tr>
<th>Target phenotype</th>
<th>Females</th>
<th>Males</th>
<th>Assoc: prevalence and sex (DF = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)</td>
<td>Cases (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Ever used cocaine</td>
<td>1336 (58%)</td>
<td>184 (14%)</td>
<td>976 (42%)</td>
</tr>
<tr>
<td>Ever used amphetamine</td>
<td>1335 (58%)</td>
<td>235 (18%)</td>
<td>975 (42%)</td>
</tr>
<tr>
<td>Ever used inhalants</td>
<td>1331 (58%)</td>
<td>45 (3%)</td>
<td>972 (42%)</td>
</tr>
<tr>
<td>Ever used sedatives</td>
<td>1334 (58%)</td>
<td>177 (13%)</td>
<td>975 (42%)</td>
</tr>
<tr>
<td>Ever used hallucinogens</td>
<td>1334 (58%)</td>
<td>139 (10%)</td>
<td>974 (42%)</td>
</tr>
<tr>
<td>Ever used opioids</td>
<td>1335 (58%)</td>
<td>72 (5%)</td>
<td>974 (42%)</td>
</tr>
<tr>
<td>Ever used ecstasy</td>
<td>1337 (58%)</td>
<td>318 (24%)</td>
<td>975 (42%)</td>
</tr>
<tr>
<td>Ever used prescription pain killers</td>
<td>1335 (58%)</td>
<td>246 (18%)</td>
<td>974 (42%)</td>
</tr>
<tr>
<td>Ever used prescription stimulants</td>
<td>1334 (58%)</td>
<td>151 (11%)</td>
<td>973 (42%)</td>
</tr>
<tr>
<td>Ever used alcohol abuse</td>
<td>1106 (59%)</td>
<td>450 (41%)</td>
<td>784 (41%)</td>
</tr>
<tr>
<td>Alcohol dependence</td>
<td>1368 (59%)</td>
<td>386 (28%)</td>
<td>959 (41%)</td>
</tr>
<tr>
<td>Alcohol abuse</td>
<td>1368 (59%)</td>
<td>205 (22%)</td>
<td>959 (41%)</td>
</tr>
<tr>
<td>DSM-5 AUD (ctrl mild vs moderate severe)</td>
<td>1368 (59%)</td>
<td>310 (23%)</td>
<td>959 (41%)</td>
</tr>
<tr>
<td>Cannabis abuse</td>
<td>1368 (59%)</td>
<td>105 (8%)</td>
<td>959 (41%)</td>
</tr>
<tr>
<td>Cannabis dependence</td>
<td>1368 (59%)</td>
<td>61 (4%)</td>
<td>959 (41%)</td>
</tr>
<tr>
<td>DSM-5 CUD (ctrl mild vs moderate severe)</td>
<td>1368 (59%)</td>
<td>75 (5%)</td>
<td>959 (41%)</td>
</tr>
</tbody>
</table>
PRS, $R^2$ as

$$R^2 = \frac{\hat{\beta}_{\text{PRS}}}{\text{SD}(Y)}^2 \times (\text{SD}(\hat{\beta}_{\text{PRS}}))^2$$

where $\hat{\beta}_{\text{PRS}}$ is the fixed effect estimate of a PRS, SD$_Y$ is the standard deviation of a binary or continuous target phenotype Y, and SD($\hat{\beta}_{\text{PRS}}$) is the standard deviation of the $\hat{\beta}_{\text{PRS}}$. Here, the $R^2$ is defined as the square of the correlation between Y and PRS. We calculated two-sided $p$-values from at distribution in R.

2.6. Multiple Testing

We tested the association between each of the 22 target phenotypes and each of the 40 PRSs, giving a total of 880 association tests. To account for multiple testing (Table S2), we presented $p$-values adjusted for the effective number of independent target and discovery phenotypes (threshold T4) and Bonferroni-corrected $p$-values (threshold T5). Bonferroni correction is known to be extremely conservative and increase the likelihood of a type II error. We, therefore, interpreted our results based on the associations that remained significant after accounting for the T4 threshold, which was calculated as

$$p_{\text{nom}} = \frac{p_{\text{adj}}} {M_{\text{eff}-t} \times M_{\text{eff}-d}}$$

Where $p_{\text{nom}}$ is the nominal $p$-value, $M_{\text{eff}-t}$ is the effective number of independent target phenotypes, and $M_{\text{eff}-d}$ is the effective number of independent discovery phenotypes (Li and Ji, 2005; Nyholt, 2004). We estimated the $M_{\text{eff}-t}$ as 14 and $M_{\text{eff}-d}$ as five in females, males, and both sexes.

2.7. Statistical Software

The pipeline of PRS calculation was coded in BASH (Free Software Foundation, 2007) and R (R Core Team, 2017). LD-based SNP clumping and PRS calculation were performed using Plink (Chang et al., 2015). Univariate GREML was performed using GCTA (Yang et al., 2011). We used R and Base SAS 9.4 (SAS Institute Inc., 2017) for data cleaning, statistical analyses, graph and table generation.

3. Results

3.1. Genetic Correlations Between Discovery Phenotypes

Shown in Table 2, the five discovery phenotypes were significantly correlated with each other, with positive genetic correlation

<table>
<thead>
<tr>
<th>Trait1</th>
<th>Trait2</th>
<th>$r_G$</th>
<th>SE</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>CPD</td>
<td>-0.381</td>
<td>0.037</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AI</td>
<td>DPW</td>
<td>-0.15</td>
<td>0.031</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AI</td>
<td>SC</td>
<td>-0.294</td>
<td>0.041</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AI</td>
<td>SI</td>
<td>-0.685</td>
<td>0.024</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CPD</td>
<td>DPW</td>
<td>0.084</td>
<td>0.028</td>
<td>0.0030</td>
</tr>
<tr>
<td>CPD</td>
<td>SC</td>
<td>0.426</td>
<td>0.033</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CPD</td>
<td>SI</td>
<td>0.28</td>
<td>0.033</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>DPW</td>
<td>SC</td>
<td>0.104</td>
<td>0.033</td>
<td>0.0016</td>
</tr>
<tr>
<td>DPW</td>
<td>SI</td>
<td>0.403</td>
<td>0.018</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SC</td>
<td>SI</td>
<td>0.39</td>
<td>0.029</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

The $r_G$, standard errors (SE) and $p$ values were estimated by cross-trait LD score regression. These five phenotypes are smoking initiation (SI, binary), age at starting regular smoking (AI, continuous), cigarettes per day (CPD, continuous), smoking cessation (SC, binary), and drinks per week (DPW, continuous).

The $r_G$ range is visualised for both sexes (Fig. S1), males (Fig. S2) and females (Fig. S3). Effect sizes of PRSs were all positive in these associations (estimates highlighted in blue and bold-face, Table S3’). The highest correlation found between individual target phenotypes and PRSs were generally very small, ranging between -0.3 and 0.37 (Table S4’), with the highest correlation found between age at onset of cannabis dependence and PRS-DPW calculated at $p$ value < 5e-08.

4. Discussion

The aim of this study was to examine the association between molecular-based estimates of polygenic risk for alcohol and nicotine use and self-report measures of common and emergent types of illicit substance use and misuse. The pattern of genetic correlations (positive $r_G$ range: 0.08–0.43, negative $r_G$ range: -0.68~−0.15, see Table 2), as estimated by cross-trait LD score regression, between the five discovery PRS was very similar to those observed in the full GSCAN full sample ($r_G$ range: -0.71−0.42, Liu et al., 2019). The highest $r_G$ was between AI and SI, along with moderate to weak correlations in nine other pairwise correlations. Among the SU phenotypes predicted by PRS-SI or PRS-DPW were lifetime cocaine, amphetamine, hallucinogen, ecstasy, and cannabis use (Fig. 1). The DSM-5 AUD was the only clinical SU diagnosis that was predicted by PRS-SI. Importantly, none of the PRSs that were based only on SNPs that reached genome-wide significant ($p$-value < 5e-08) significantly predicted any of the SU or SUD outcomes. However, significant associations were found when PRSs were based on less stringent $p$-values (5e-08 < $p$-value < 1). Despite the small-sized target sample, the use of PRSs based on a large discovery GWAS samples has the power to predict, to some extent, individuals at risk of comorbid SU and progression to AUDs.

4.1. Prediction by PRSs

Although the range in variances explained by our PRSs was small (0.67%–1.66% in both sexes, 0.00%~2.21% in females, 0.58%~1.80% in males, see Table S3’), these estimates are commensurate with other studies that have used aggregated genetic risk based on well-powered GWAS discoveries to predict substance use and misuse. For example, GSCAN PRS for AI, CPD, SI and DPW explained between 1% and 4% of the variance in similar measures assessed by the Add Health and Health and Retirement Study datasets (Liu et al., 2019). Likewise, PRS based on a recent genome-wide meta-analysis of alcohol dependence explained between 0.3% and 1.7% of the variance in alcohol use and misuse phenotypes in the large Avon Longitudinal Study of Parents and Children (ALSPAC), Generation Scotland (GS), and Collaborative Study...
on the Genetics of Alcoholism (COGA) samples (Walters et al., 2018).

We found that cannabis initiation was significantly predicted by the PRS-SI, but not by PRS-CPD or PRS-SC. In contrast, Vink et al. (2014) reported an association between cannabis initiation and a PRS for CPD, based on the Tobacco and Genetics Consortium (TAG), but not with a TAG-based PRS for SI (N = 69,409), or PRS for SC. Apart from the obvious discrepancy in the two discovery sample sizes, the differences in the prediction by the smoking-related PRSs predicts may be attributable to the heterogeneity in terms of how the discovery GWAS consortia defined their phenotypes. For example, the GSCAN SI was a binary phenotype based on whether subjects had ever been a regular smoker, whereas the TAG-based SI used by Vink et al. was more strictly defined. The GSCAN CPD used the average number of cigarettes smoked per day scored on an ordinal scale with five response categories (1–5, 6–15, 16–25, 26–35, or 36 or more cigarettes smoked per day) whereas the CPD used by Vink et al. was the average or maximum number of cigarettes smoked per day from contributing studies. We cannot rule out the possibility that our study was underpowered to detect the effect of PRS-CPD on the initiation of cannabis, and even the other substances.

Given the sample sizes and success of recent GWAS meta-analyses investigating alcohol use and misuse (Liu et al., 2019; Walters et al., 2018) our *a priori* expectation was that the PRS-DPW should have predicted the self-reported measures of alcohol misuse. Instead, the PRS-DPW did not predict any alcohol-related disorders. In addition to the lack of power related to the BLTS samples size, this null finding could be attributable to the possibility that the PRS-DPW was assessing an unrelated genetic liability. Note that GSCAN DPW phenotype was based on either current or former drinkers, combined all types of liquor, or averaged across all assessments when based on longitudinal data (Liu et al., 2019). Kendler et al. (2012a) found that the DSM-IV syndrome of alcohol dependence does not reflect a single dimension but instead is best characterized by three distinct genetic factors: tolerance and heavy use; loss of control; and withdrawal symptoms (Kendler et al., 2012b). Therefore, it is plausible that GSCAN-based PRS for DPW may be indexing the genetics of tolerance and heavy use, which is in contrast to...
our measure of alcohol misuse that is more likely to be indexing genetic risks related to loss of control and withdrawal.

The pattern of non-significant associations between the alcohol and nicotine-based PRS and the non-medical use of prescribed or over the counter analgesics is commensurate with our recent twin modeling exploring the etiology of this emergent class of SU. Briefly, we have found that apart from opioids, broadly defined non-medical use of analgesics is genetically correlated \((r_G = 0.44)\) but mostly distinct from the genetic risks underpinning lifetime nicotine use (Gillespie et al., 2019). In terms of diagnostic outcomes, non-medical use of analgesics was only partially genetically correlated with either alcohol use disorder \((r_G = 0.24)\) and nicotine dependence \((r_G = 0.32)\) (Gillespie et al., 2019), which suggests that most of the genetic risks in non-medical use of analgesics are variable specific.

4.2. Limitations

Our results must be interpreted in the context of the following two limitations. First, the discovery sample was based predominately on Caucasian ancestral groups. Consequently, the significant PRS-phenotype correlations observed here may not apply to other ethnic groups especially those with different haplotype structures. The need to collect genetic data from non-European-ancestry populations has been recently discussed (Bentley et al., 2017; Kessler et al., 2016; Lewis and Vassos, 2017).

Second, the significant associations between genetic risks for alcohol- and nicotine-related phenotypes and self-report measures of SU and SUDs do not imply causation. Third, our PRSs explained a relatively small proportion of the genetic variance, a common problem in addiction studies (Miles et al., 2017; Vink et al., 2014) that uses the PRS approach. Although it is premature to discuss the clinical utility of PRSs in genetic risk prediction for substance use, the PRS method has been shown to improve risk prediction in other diseases, such as prostate cancer (Helfand et al., 2016), when it is combined with a conventional risk factor family history. As discovery sample sizes increase and the PRS approach continues to refine, the PRS predictor can prove useful in discriminating patients in the top and bottom risk deciles (Lewis and Vassos, 2017).

4.3. Significance of this Study

The current results lend support for there being shared genetic risks between initiation of tobacco and various addictive substances (e.g. cocaine, amphetamine, hallucinogens, ecstasy, and cannabis), as well as between alcohol consumption and several of these substances. Twin studies show that common genetic factors tend to exert a varying degree of influence on the covariation of tobacco and cannabis use and misuse at different stages of the involvement of these two substances (Huizink et al., 2010; Neale et al., 2006), with the stronger influence at the earlier stages (e.g. initiation) than the later stages (e.g. progression) of the use. Taken together with our finding, this stage-dependent genetic relationship may be generalised to tobacco and other addictive substances (i.e. cocaine, amphetamine, hallucinogens, ecstasy).

Role of Funding Source

Research in the current study was funded by National Health and Medical Research Council [APP10499110] in Australia and in part by National Institutes of Health [K99RO0, R00DA23549, R21DA038852, R01DA37904]. LHC is supported by a QIMR International PhD scholarship. SEM is supported by an NHMRC fellowship [APP1103623].

References


Dar, S., Forrer, L., Schroierz, S., Schidore, C., Locke, A.E., Kwong, A., Vrieze, S.L., Chew,


