Addiction
RESEARCH REPORT
doi:10.1111/add.12491

Polygenic risk scores for smoking: predictors for alcohol and cannabis use?

Jacqueline M. Vink[1,2], Jouke Jan Hottenga[1], Eco J. C. de Geus[1], Gonneke Willemsen[1], Michael C. Neale[3], Helena Furberg[4] & Dorret I. Boomsma[1,2]

Department of Biological Psychology, VU University, Amsterdam, the Netherlands,[1] Neuroscience Campus Amsterdam, Amsterdam, the Netherlands,[2] Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA[3] and Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA[4]

ABSTRACT

Background and Aims A strong correlation exists between smoking and the use of alcohol and cannabis. This paper uses polygenic risk scores to explore the possibility of overlapping genetic factors. Those scores reflect a combined effect of selected risk alleles for smoking. **Methods** Summary-level *P*-values were available for smoking initiation, age at onset of smoking, cigarettes per day and smoking cessation from the Tobacco and Genetics Consortium (*n* between 22 000 and 70 000 subjects). Using different *P*-value thresholds (0.1, 0.2 and 0.5) from the meta-analysis, sets of 'risk alleles' were defined and used to generate a polygenic risk score (weighted sum of the alleles) for each subject in an independent target sample from the Netherlands Twin Register ($n = 1583$). The association between polygenic smoking scores and alcohol/cannabis use was investigated with regression analysis. **Results** The polygenic scores for 'cigarettes per day' were associated significantly with the number of glasses alcohol per week ($P = 0.005$, $R^2 = 0.4$–0.5%) and cannabis initiation ($P = 0.004$, $R^2 = 0.6$–0.9%). The polygenic scores for 'age at onset of smoking' were associated significantly with 'age at regular drinking' ($P = 0.001$, $R^2 = 1.1$–1.5%), while the scores for 'smoking initiation' and 'smoking cessation' did not significantly predict alcohol or cannabis use. **Conclusions** Smoking, alcohol and cannabis use are influenced by aggregated genetic risk factors shared between these substances. The many common genetic variants each have a very small individual effect size.

Keywords Alcohol, cannabis, genetic, heritability, polygenic risk score, smoking, SNP, substance use.

Correspondence to: Jacqueline M. Vink, Department of Biological Psychology, VU University, Amsterdam 1081 BT, the Netherlands. E-mail: jm.vink@vu.nl
Submitted 17 July 2013; initial review completed 9 October 2013; final version accepted 15 January 2014

INTRODUCTION

A strong correlation exists between smoking and the use of other substances such as alcohol and cannabis. Smoking is correlated positively with alcohol consumption, the severity of alcohol dependence [1] and the use of cannabis [2]. Twin and family studies have shown that smoking behaviour [3–5], alcohol consumption [6–8] and cannabis use [9–11] are influenced by genetic factors. Heritability estimates range from low to moderate for initiation of substance use to somewhat high for quantity and dependence [3,5,11–13]. The comorbidity of tobacco, alcohol and cannabis use is mediated by common genetic influences [14–16].

In past years, genome-wide association (GWA) studies of smoking behaviour revealed several regions and candidate genes [17–20]. However, none of these GWA studies reported genome-wide significant results, because of the limited sample sizes. It is now recognized that a well-powered GWA needs to include tens of thousands and possibly hundreds of thousands of subjects. In 2010, three large consortia, the Oxford-GlaxoSmithKline (Ox-GSK), Tobacco and Genetics Consortium (TAG) and the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium, each carried out a meta-analysis for smoking phenotypes. They also combined their analyses for smoking initiation and cigarettes per day (CPD) [21–24]. The most significant finding was the association between the number of CPD and a cluster of nicotinic receptor genes on chromosome 15 [21–24].

For cannabis use, several candidate genes are suggested based on linkage and association studies [25], but a GWA meta-analysis based on two samples (effective sample size 4312) [26] and a GWA analysis of cannabis

dependence did not reveal genome-wide significant results [27].

Rietschel & Treutlein [28] reviewed the current literature on alcohol GWAS and concluded that few genome-wide significant findings have been reported. Among the top findings are often alcohol dehydrogenase genes (ADH and ALDH2), although a variety of other genes is also reported.

Twin-family studies suggested a genetic overlap between use of different substances, but so far none of the top results in GWA studies for smoking, alcohol and cannabis overlapped.

Some examples exist of well-known substance-specific genes that are also associated with another substance. Mouse studies showed, for example, that polymorphisms located within the Chrna5–Chrna3–Chrnb4 cluster on mouse chromosome 9 (well-known smoking genes) co-segregate with alcohol preference in mice [29]. This suggests there is some overlap in risk genes for substance use or abuse.

The effect sizes of individual risk alleles underlying substance use are small, with most genotype relative risks in the range of 1.1–2.0. The joint effect of all measured DNA variants explained 19–28% of the variance in smoking initiation, 24–44% in current smoking [30] and 6% in cannabis use [26]. These findings suggest that individuals may be at risk for substance use through multiple genetic variants each with a small contribution.

Polygenic risk scores have been used to summarize genetic effects among a group of genetic variants that do not individually achieve significance in a large-scale association study. First, a meta-analysis on GWA results is conducted on an initial discovery sample, and the markers are ranked by their evidence for association, usually based on their *P*-values. An independent target sample is then analysed by constructing a polygenic score consisting of the weighted sum of the associated alleles within each subject. Association between a trait and this score implies a genetic effect of the trait in the discovery sample on the trait in the target sample. The first successful application of a polygenic risk score analysis was in a GWA study of schizophrenia [31]. A polygenic risk score based on the GWA for schizophrenia was associated with the risk of bipolar disorder, but not to several non-psychiatric diseases (which suggested disease specificity). The polygenic risk score method has been used in several studies, with mixed results. Some studies report positive associations (for example [32–35]), while others did not find evidence that common genetic risk variation is shared between two traits (for example [36,37]). This might be due to the size of the discovery sample (because the accuracy of the prediction score increases with the size of the discovery sample), or it may indicate a lack of genetic overlap.

In the present study, polygenic risk scores for smoking were identified based on the large meta-analysis of the Tobacco and Genetics (TAG) Consortium including 20 000–70 000 subjects. Four phenotypes were included in the TAG GWA meta-analysis: ever versus never regular smoking (ever), age at onset of smoking (AOS), CPD and smoking cessation (former). The risk alleles from TAG were used to calculate a polygenic risk scores in an unrelated sample of the Netherlands Twin Register (NTR) ($n = 1583$) and the association between this risk score for smoking and alcohol/cannabis use was explored.

METHODS

Discovery sample from the TAG consortium

The TAG consortium reported summary-level *P*-values of the GWA meta-analysis of four smoking phenotypes based on 20 000–70 000 subjects [38]. Sixteen studies contributed to the meta-analysis and performed their own genotyping, quality control and imputation. Studies ranged in size from 585 to 22 037 and were genotyped on six different GWAS platforms.

Four dimensions of smoking behaviour were analysed [38]:

1. Ever versus never regular smokers (ever): regular smokers (1) were defined as those who reported having smoked ≥100 cigarettes during their life-time, and never regular smokers (0) were defined as those who reported having smoked between 0 and 99 cigarettes during their life-time; total sample size $n = 69.409$.
2. Age at onset of smoking (AOS): age of smoking initiation was the reported age the participant started smoking cigarettes; total sample size $n = 22.438$.
3. Cigarettes per day (CPD): the average or maximum (depending on study) number of cigarettes smoked per day; total sample size $n = 38.181$.
4. Smoking cessation (former): smoking cessation contrasted former (= 0) versus current (= 1) smokers.

Each study conducted uniform cross-sectional analyses for each smoking phenotype using an additive genetic model. Linear regression was used for quantitative traits (CPD and AOS), and logistic regression was used for dichotomous traits (ever and former); total sample size $n = 35.845$.

The analyses were run separately for males and females. Because the TAG consortium did not detect significant interactions by sex, data were analysed together. Age was not included as a covariate. Case–control studies included case/control status as a covariate, while cohort studies did not include an additional covariate.

Target sample from the NTR

The target sample consisted of subjects from the NTR who were not part of the TAG meta-analysis. The NTR collects longitudinal data in twin-families [39,40]. In total, eight waves of survey data on personality, health and life-style were collected in 1991, 1993, 1995, 1997, 2000, 2002, 2004 and 2009.

1 Age at regular alcohol use; answer options: < 11 years, 12, 13, 14, 15, 16, 17, 18 years or older, never (surveys 2, 3, 4 and 8). When longitudinal data were available (also for age at first cannabis use): with a discrepancy of 1 or 2 years, the youngest age is selected; with a discrepancy of more than 2 years, the variable is set to missing 0–5% of cases).
2 Glasses of alcohol per week; answer options: less than one glass, one to five glasses, six to 10 glasses, 11–15 glasses, 21–40 glasses, more than 40 glasses. When longitudinal data were available we used the highest number of glasses reported in all available data (surveys 2–8). No survey data on alcohol use were available for 203 subjects.
3 Age at first time cannabis use; answer options: < 11 years, 12, 13, 14, 15, 16, 17, 18 years or older, never (surveys 2, 3, and 8).
4 Ever cannabis use: the question age at first cannabis use is collapsed into ever (1) and never (0).

We have chosen alcohol and cannabis phenotypes as similar as possible to smoking phenotypes from the TAG study (CPD → glasses alcohol per week, AOS → age at regular alcohol use/age at cannabis initiation, ever smoked → ever used cannabis).

DNA samples [41] were genotyped in different projects and genotyping was performed on Affymetrix 6.0 ($n = 298$), Affymetrix Perlegen 5.0 ($n = 3697$), Illumina 370 ($n = 290$), Illumina 660 ($n = 1439$) and Illumina Omni Express 1 M ($n = 455$) platforms. Calls were made with platform-specific software (Genotyper; Beadstudio, San Diego, CA, USA). The quality control thresholds for single nucleotide polymorphisms (SNPs) were minor allele frequency (MAF) > 1%, Hardy–Weinberg equilibrium (HWE) > 0.00001, call rate >95% and 0.30 < heterozygosity <0.35. Samples were excluded from the data if their expected sex and IBD status did not match, or if the genotype missing rate was >10%. SNPs were aligned to the positive strand of the Hapmap-2-Build 36-release-24 CEU reference set. Alignment was checked using individuals and family members tested on multiple platforms. SNPs were excluded if allele frequencies differed more than 15% from the reference set and/or the other platforms. The data of the platforms were merged into a single data set ($n = 5856$). This merged set was imputed against the reference set using IMPUTE version 2. After imputation, genotype dosage was calculated if the highest genotype probability was above 90%. Badly imputed SNPs were removed based on HWE < 0.00001, proper-info < 0.40, MAF < 1%, allele frequency difference >0.15 against reference.

NTR subjects who participated in the Genetic Association Information Network (GAIN)–NTR study were excluded because those subjects were included in the original TAG meta-analysis. Family members of subjects in the GAIN–NTR study were also excluded (except non-biological members, such as spouses of twins). This resulted in a sample of 1583 subjects with genotype data, and 72% of the sample was female. The year of birth ranged between 1915 and 1994 (median 1958). Subjects were of European descent.

Polygenic risk scores and statistical analysis

The polygenic risk scores reflect a combined effect of a number of selected SNPs [38]. Different *P*-value thresholds (Pt) of 0.1, 0.2 and 0.5 were used to define large sets of 'risk alleles' in the discovery sample (from the TAG meta-analysis summary-level data). Those sets of risk alleles are used to generate a polygenic risk score for individuals in an independent target sample from the NTR. The individual risk score is calculated by multiplying the number of risk alleles per SNP (0, 1, 2) with the regression coefficient from the GWA meta-analysis, summed over all SNPs in the considered set of SNPs [42]. The individual polygenic risk scores for the NTR participants were calculated using PLINK, with commands: –bfile NTRfile –maf 0.01 –mind 0.1 –geno 0.1 –hwe 0.000001 –score TAG_AOS_P5.dat –out TAG_AOSp5. Only SNPs that overlapped between the TAG sample and the NTR sample were included (Table 1).

Regression models were used to test the association with the polygenic risk scores based on smoking (predictor variable) and alcohol and cannabis variables (independent variables). Linear regression models were used for continuous variables and logistic regression models for the dichotomous outcome variables. Regression analyses were carried out in Stata (version 9.0) and corrected for family clustering by employing the robust cluster option. Sex and birth cohort were added as covariates. To clarify how much variance is explained by the risk score itself and how much by the covariates, the R^2 of the regression models will be presented including only the polygenic risk score (model 1), the regression model with risk score and sex (model 2) and the regression model with risk score, sex and age.

An association between a polygenic risk score and an outcome variable was considered significant if $P < 0.005$ (we used a more stringent *P*-value than 0.05 to correct for multiple testing). We considered the results with $0.05 < P < 0.005$ as marginally significant, and discuss the results in this context.

Table 1 Overview of the number of available single nucleotide polymorphisms (SNPs) in the Tobacco and Genetics Consortium (TAG) sample (all SNPs TAG), the number of overlapping SNPs between the TAG sample and the Netherlands Twin Register (NTR) sample, and the number of SNPs selected with the different *P*-value selection criteria.

P-value thresholds (Pt)	*n SNPs CPD*	*n SNPs AOS*	*n SNPs ever*	*n SNPs former*
All SNPs TAG	2 502 107	2 500 547	2 498 833	2 499 522
SNPs TAG and NTR	2 123 025	2 122 544	2 121 558	2 121 558
Pt = 0.5	1 088 808	1 079 361	1 103 228	1 085 301
Pt = 0.2	450 210	442 816	474 407	449 091
Pt = 0.1	230 447	224 460	252 924	233 788

AOS = age at onset of smoking; CPD = cigarettes per day.

Because correlations between the four different risk scores were (relatively) low, we also analysed the four risk scores simultaneously in a regression analysis to explore whether the risk scores have an independent effect when corrected for the other risk scores.

RESULTS

Table 2 shows the distribution of the alcohol and cannabis variables for the NTR target sample. Approximately 3.5% of the sample never initiated alcohol use, and those subjects were excluded for age at regular drinking. From the subjects who ever tried alcohol, 22.6% never started to drink regularly and more than half of the subjects (58.7%) started regular drinking after the age of 17. Almost 9% reported drinking more than 20 glasses alcohol per week. In the total sample, 85% had never tried cannabis. Most of the subjects who tried cannabis started at 18 years or older.

All polygenic risk scores for smoking showed a marginally significant association with one or more smoking variables ($0.005 < P < 0.05$) in our independent target sample, except the polygenic risk score for age at smoking onset (Supporting information, Table S1).

The polygenic risk score based on age at onset of smoking was associated significantly with age at which regular drinking started. This risk score was not associated with any of the other alcohol or cannabis phenotypes (Table 3a). The polygenic risk scores ever and former smoking did not significantly predict alcohol or cannabis use (Table 3b,c). The risk scores based on CPD were associated significantly with the number of glasses of alcohol per week and cannabis initiation, but not with age at regular drinking or age at first cannabis use (Table 3d).

Figures 1–3 show the regression coefficients or odds ratios of the significant associations, as well as the proportion of explained variance. The polygenic risk score for age at smoking onset explained 1.1–1.5% of the variance in age at regular drinking. When sex and birth cohort were included in the model, the explained variance was higher (around 20% for model 3; see Table 3a).

Table 2 Distribution of the alcohol and cannabis variables in the Netherlands Twin Register (NTR) target sample ($n = 1583$).

Variable	*Categories*	*n (%)*
Age regular drinking (among subjects who ever tried alcohol)	11 or younger	1 (0.1%)
	12 years	1 (0.1%)
	13 years	1 (0.1%)
	14 years	14 (1.5%)
	15 years	41 (4.4%)
	16 years	97 (10.4%)
	17 years	69 (7.4%)
	18 years or older	497 (51.3%)
	Never	210 (22.6%)
	Missing	396
Glasses alcohol per week (among subjects who ever tried alcohol)	<weekly	290 (21.2%)
	1–5 glasses	271 (19.8%)
	6–10 glasses	276 (20.2%)
	11–15 glasses	243 (17.8%)
	16–20 glasses	166 (12.2%)
	21–40 glasses	104 (7.6%)
	>40 glasses	15 (1.1%)
	Missing	13
Age at first cannabis[a]	11 or younger	0 (0%)
	12 years	1 (0.1%)
	13 years	0 (0%)
	14 years	10 (0.9%)
	15 years	15 (1.4%)
	16 years	28 (2.6%)
	17 years	16 (1.5%)
	18 years or older	93 (8.5%)
	Never	925 (85.0%)
	Missing	495

[a]This variable is also collapsed into ever/never cannabis use.

The polygenic risk score for CPD predicted 0.4–0.5% of the variance in the number of glasses alcohol per week (see Table 3d) in the target sample. The polygenic risk score for CPD explained 0.6–0.9% of the variance in cannabis use.

The correlation between the four different risk scores is moderate to low (Table 4). We compared the risk scores based on Pt = 0.2. The score for ever/never smoking was not associated significantly with the scores for former smoking or CPD. The highest correlation was found

Table 3 (a,b,c,d) Overview of results from linear (continuous variables) and logistic (dichotomous variables) regression analyses to predict alcohol and cannabis use with polygenic risk scores for smoking. Polygenic scores are based on meta-analyses summary data of the Tobacco and Genetics consortium. Regression analyses are carried out on independent sample from the Netherlands Twin Register (NTR). Sex was used as covariate in the regression analyses. P = P-value from the regression analyses; R^2 = the explained variance; β = regression coefficient; standardized β/odds ratio (OR) = standardized beta from linear regression analyses or OR from logistic regression. The standardized beta coefficients are the estimates resulting from the analysis carried out on the independent variables that have been standardized so that their variances are 1. Therefore, standardized coefficients refer to how many standard deviations a dependent variable will change, per standard deviation increase in the predictor variable. A positive beta reflects a positive association (for example, the higher the polygenic risk score for age at smoking onset, the higher the age at regular drinking).

Table 3 (a) Age at smoking initiation (AOS).

	n = 721, 561 families (ever alcohol = yes)					
Age at regular drinking	*β*	*β standardized*	*P*	*R^2 model 1*	*R^2 model 2*	*R^2 model 3*
Pt = 0.1	4.923	0.103	**0.001**	0.015	0.029	0.202
Pt = 0.2	6.957	0.098	**0.001**	0.011	0.028	0.201
Pt = 0.5	11.610	0.087	**0.002**	0.011	0.024	0.200
	n = 1069, 819 families (ever alcohol = yes)					
Glasses alcohol per week	*β*	*β standardized*	*P*	*R^2*		
Pt = 0.1	1.791	0.011	0.717	0.015	0.108	0.117
Pt = 0.2	0.165	0.002	0.950	0.000	0.108	0.117
Pt = 0.5	−0.982	−0.006	0.843	0.000	0.108	0.117
	n = 163, 154 families (ever cannabis = yes)					
Age at first cannabis	*β*	*β standardized*	*P*	*R^2*		
Pt = 0.1	−1.975	−0.032	0.655	0.000	0.002	0.101
Pt = 0.2	−1.581	−0.017	0.814	0.001	0.003	0.100
Pt = 0.5	0.621	0.004	0.961	0.002	0.004	0.100
	n = 1088, 820 families					
Ever cannabis	*β*	*OR*	*P*	*R^2*		
Pt = 0.1	−3.275	0.038	0.382	0.001	0.004	0.116
Pt = 0.2	−4.090	0.017	0.442	0.000	0.004	0.116
Pt = 0.5	−3.625	0.0266	0.716	0.000	0.004	0.116

Pt = *P*-value thresholds.

between the polygenic scores for CPD and former smoking (−0.20).

Because the correlations between the different risk scores were low, we also analysed the four polygenic risk scores simultaneously in a regression analysis. The risk score for CPD still predicted smoking initiation, even when correcting for the other risk scores, while both the CPD risk score and the ever risk score predicted CPD. (Supporting information, Table S2).

DISCUSSION

The aim of this study was to investigate the overlap in polygenic risk factors between smoking behaviour and alcohol and cannabis use. Using polygenic risk scores derived from the GWA meta-analysis results of the TAG Consortium we predicted alcohol and cannabis use in an independent sample from the NTR. The risk scores for CPD explained 0.4–0.5% of the variance in glasses alcohol per week and 0.6–0.9% of the variance in cannabis initiation. The polygenic risk scores for age at onset of smoking predicted about 1.1–1.5% of the variance in age at regular alcohol use. The risk scores for smoking initiation and smoking cessation were not associated significantly with alcohol and cannabis use.

When complex phenotypes, such as addiction phenotypes, display a polygenic genetic architecture it is unlikely that GWA studies lead to straightforward results

Table 3 (b) Ever/never smoked.

	n = 721, 561 families (ever alcohol = yes)					
Age regular drinking	*β*	*β standardized*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
Pt = 0.1	0.026	0.004	0.911	0.000	0.014	0.192
Pt = 0.2	–0.098	–0.010	0.785	0.001	0.014	0.192
Pt = 0.5	–0.288	–0.015	0.661	0.001	0.014	0.192
	n = 1069, 819 families (ever alcohol = yes)					
Glasses alcohol per week	*β*	*β standardized*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
Pt = 0.1	0.108	0.014	0.638	0.000	0.108	0.118
Pt = 0.2	–0.038	–0.003	0.910	0.000	0.108	0.117
Pt = 0.5	–0.216	–0.009	0.737	0.000	0.108	0.118
	n = 163, 154 families (ever cannabis = yes)					
Age at first cannabis	*β*	*β standardized*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
Pt = 0.1	0.755	0.090	0.196	0.009	0.010	0.108
Pt = 0.2	1.103	0.086	0.208	0.010	0.011	0.107
Pt = 0.5	2.503	0.101	0.125	0.012	0.013	0.110
	n = 1088, 820 families					
Ever cannabis	*β*	*OR*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
Pt = 0.1	0.427	1.533	0.606	0.001	0.005	0.116
Pt = 0.2	0.488	1.628	0.689	0.001	0.004	0.116
Pt = 0.5	0.461	1.586	0.836	0.000	0.004	0.116

Pt = *P*-value thresholds.

that can be replicated in independent samples. The TAG meta-analysis showed that even with large sample sizes no genome-wide significant results were obtained for smoking initiation and age at smoking initiation. For CPD, a very strong association was observed for the SNPs in the cluster of nicotinic receptor genes on chromosome 15 (15q25.1) [38]. This CPD phenotype was also responsible for significant results in the present study. Interestingly, the significant associations we observed were not driven by the top SNPs on chromosome 15. The association between the polygenic risk score for CPD and glasses of alcohol or cannabis initiation was not significant when a smaller number of SNPs was selected; for example: Pt = 0.01 (data not shown). A recent study composed a polygenic risk score based on four of the top SNPs from the 15q25.1 region and two SNPs from another region (19q13.2). This score was unrelated to smoking initiation, but the individuals with a high score were more likely to convert to heavy, persistent smoking [43]. Another study incorporated a SNP score of 92 top SNPs (based on meta-analysis [23]) in a developmental model of CPD. The SNP score was associated with CPD, but not with the frequency of alcohol use at different ages [44]. Our results suggested that, besides the top SNPs from the meta-analysis, a large number of SNPs with all small individual effect sizes contribute to substance (ab)use.

The correlations between the four polygenic risk scores for smoking were moderate to low to non-significant. The highest correlation was found between the polygenic scores for CPD and former smoking and it was negative, suggesting that being an ex-smoker is associated with a high number of cigarettes per day. This can be explained by the fact that former smokers reported on the maximum number of CPD while smokers report on their current number of CPD. The moderate correlations between the four polygenic risk scores might be the result of much error variance resulting from random, non-generalizable, non-linear and/or interactive genetic effects. Previous twin studies have suggested some overlap between smoking-related variables, varying from only a small proportion of shared genetic variance [45,46] between age at first cigarette and smoking variables to a higher genetic overlap between smoking persistence and initiation [47]. A study of the NTR showed two separate dimensions for smoking initiation and nicotine dependence, but those dimensions were not independent [5].

Table 3 (c) Smoking cessation (former).

	n = 931, 712 families (ever alcohol = yes)					
Age regular drinking	*β*	*β standardized*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
P = 0.1	0.026	0.004	0.911	0.000	0.014	0.192
P = 0.2	–0.098	–0.010	0.785	0.000	0.014	0.192
P = 0.5	–0.288	–0.015	0.66	0.001	0.014	0.192
	n = 1313, 1004 families (ever alcohol = yes)					
Glasses alcohol per week	*β*	*β standardized*	*P*	R^2		
P = 0.1	0.108	0.014	0.638	0.000	0.108	0.118
P = 0.2	–0.038	–0.003	0.910	0.000	0.108	0.118
P = 0.5	–0.216	–0.009	0.737	0.000	0.108	0.118
	n = 163, 154 families (ever cannabis = yes)					
Age at first cannabis	*β*	*β standardized*	*P*	R^2		
P = 0.1	0.755	0.090	0.169	0.009	0.010	0.108
P = 0.2	1.103	0.087	0.204	0.010	0.011	0.107
P = 0.5	2.503	0.101	0.125	0.012	0.013	0.110
	n = 1087, 829 families					
Ever cannabis	*β*	*OR*	*P*	R^2 *model 1*	R^2 *model 2*	R^2 *model 3*
P = 0.1	–0.015	0.985	0.979	0.001	0.004	0.116
P = 0.2	–0.509	0.601	0.573	0.000	0.003	0.116
P = 0.5	–0.990	0.371	0.526	0.000	0.003	0.116

Pt = *P*-value thresholds.

In the present study, the explained variance in the regression analysis varied from 0.4% for glasses of alcohol per week up to 1.5% for age at regular drinking. Other studies reported explained variances varying from 0.1 to 3% [32,48–50]. Even when taking all available SNPs into account, the explained variance is lower than the heritability estimates from twin studies [26,30]. An explanation for this 'missing heritability' problem is that the mutations causing variation in a trait are not in perfect linkage disequilibrium with any of the measured SNPs, and therefore part of the genetic variance is undetected by the SNPs. The causal variants are expected to have lower minor allele frequencies (<0.1) because they are more likely to be subject to some form of natural selection that leads to variants associated negatively with reproductive fitness [51].

Low reported values of R^2 might not directly reflect the degree of missing heritability, but could also reflect the effect of sampling variation on the variance explained by an estimated score [50]. Because the individual SNP effects are very small, they are estimated with a great deal of error. Although we can obtain an unbiased estimate of a SNP effect, a prediction of a phenotype using the estimated SNP effect suffers from the sampling variance with which the effect is estimated. The crudeness of the measures of substance use in the present study might have limited the explained variance. The worse the estimate of the effect size of the variant in the discovery sample, the worse the variance will be explained by the predictor in the validation sample [49,50].

Simulations showed that large sample sizes of the discovery sample are necessary [50]. A strength of the present study is that summary-level results of the TAG meta-analysis were used as discovery sample. The TAG meta-analysis is currently the largest GWA meta-analysis for smoking behavior. The chances of success of polygenic risk score analysis depend primarily on the size of the discovery set. If the sample size is too small, the risk profiles will be based on random noise and are not expected to explain variance in the target set [31,50,52].

For traits with a moderate heritability (h^2 0.40) the required sample sizes of the discovery samples are about twice as high as for a trait with high heritability (h^2 0.80) [50]. Simulations showed that even with high heritability traits the sample size of, for example, TAG is still somewhat low. Besides sample size other factors, such as the

Table 3 (d) Cigarettes per day (CPD)

	n = 721, 561 families (ever alcohol = yes)					
Age at regular drinking	*β*	*β standardized*	*P*	*R^2 model 1*	*R^2 model 2*	*R^2 model 3*
Pt = 0.1	0.007	0.005	0.886	0.000	0.014	0.192
Pt = 0.2	0.004	0.002	0.958	0.000	0.014	0.192
Pt = 0.5	–0.035	–0.009	0.783	0.000	0.014	0.192
	n = 1069, 819 families (ever alcohol = yes)					
Glasses alcohol per week	*β*	*β standardized*	*P*	*R^2 model 1*	*R^2 model 2*	*R^2 model 3*
Pt = 0.1	0.111	0.068	*0.016*	0.004	0.113	0.122
Pt = 0.2	0.191	0.078	**0.005**	0.005	0.114	0.124
Pt = 0.5	0.342	0.075	**0.007**	0.004	0.114	0.123
	n = 163, 154 families (ever cannabis = yes)					
Age at first cannabis	*β*	*β standardized*	*P*	*R^2 model 1*	*R^2 model 2*	*R^2 model 3*
Pt = 0.1	0.039	0.026	0.758	0.000	0.002	0.101
Pt = 0.2	0.029	0.013	0.874	0.000	0.002	0.100
Pt = 0.5	–0.144	–0.033	0.690	0.002	0.003	0.101
	n = 1088, 820 families					
Ever cannabis (yes/no)	*β*	*OR*	*P*	*R^2 model 1*	*R^2 model 2*	*R^2 model 3*
Pt = 0.1	0.278	1.321	*0.015*	0.006	0.010	0.123
Pt = 0.2	0.481	1.617	**0.004**	0.009	0.013	0.125
Pt = 0.5	0.897	2.453	**0.004**	0.007	0.011	0.125

Pt = *P*-value thresholds.

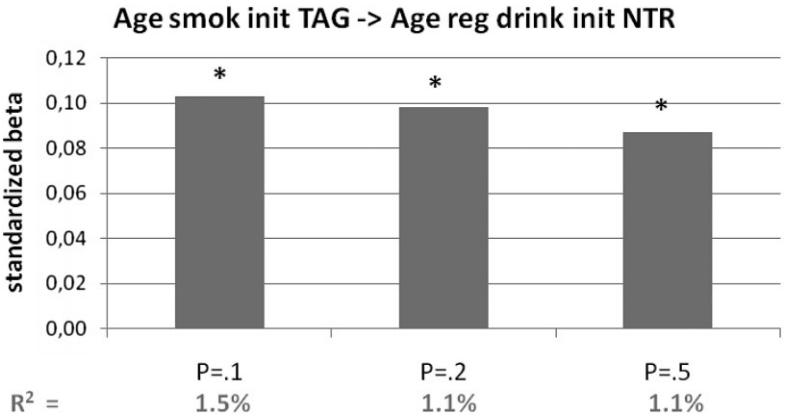

Figure 1 Polygenic risk score of age at smoking initiation (with different *P*-value thresholds) predicting age at regular drinking in target sample from the Netherlands Twin Register (NTR). The vertical axis shows the standardized regression coefficients (beta) from the regression analyses. The bottom row shows the explained variance (R^2). Bars marked with $*P < 0.05$; TAG: Tobacco and Genetics Consortium

proportion of SNPs having effect on the trait, are of importance [50]. We have used four smoking dimensions from the TAG meta-analysis, and the heritability of these phenotypes varied. The fact that a significant association with alcohol and cannabis use was found for the polygenic risk scores of CPD and AOS but not for smoking initiation and smoking cessation might be (partly) explained by differences in heritability. In samples of the NTR the heritability was 75% for nicotine dependence, 51% for CPD, 60% (males) and 39% (females) for age at first cigarette and 36%–44% for smoking initiation [5,53,54]. CPD might mirror more 'severe' phenotypes that reflect addictive behaviour (such as nicotine dependence).

The present results support the idea of a shared genetic background between smoking and use of alcohol and cannabis. In conclusion, our data point to a genetic architecture of many common variants with very small

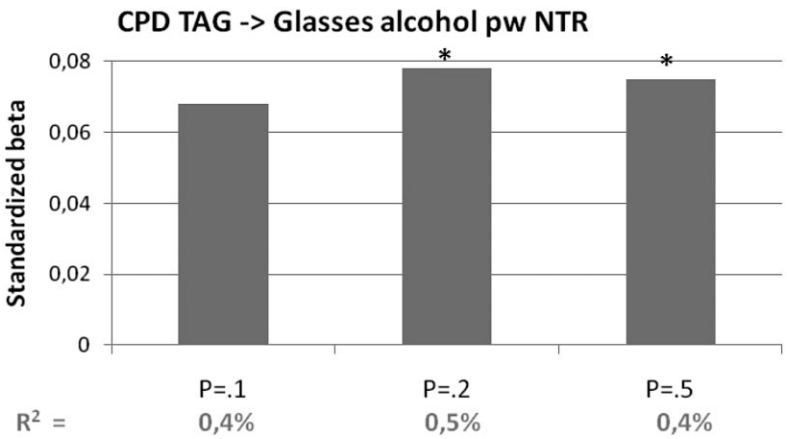

Figure 2 Polygenic risk score of cigarettes per day (with different *P*-value thresholds) predicting daily drinking in target sample from the Netherlands Twin Register (NTR). The vertical axis shows the odds ratios from the regression analyses. The bottom row shows the explained variance (R^2). Bars marked with * $P < 0.05$; TAG: Tobacco and Genetics Consortium; CPD: cigarettes per day

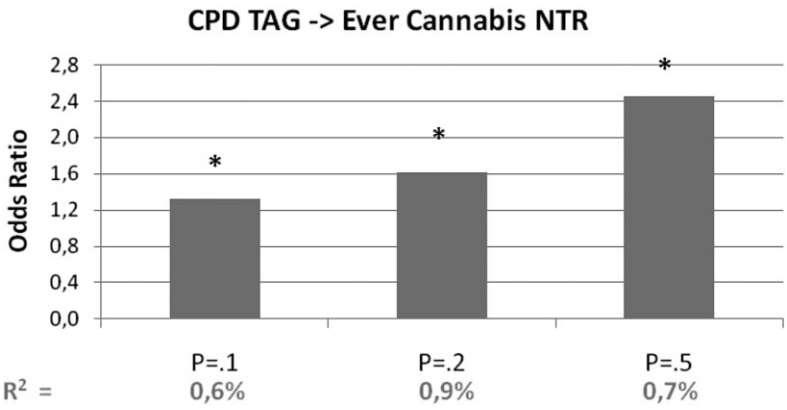

Figure 3 Polygenic risk score of cigarettes per day (with different *P*-value thresholds) predicting cannabis initiation in target sample from the Netherlands Twin Register (NTR). The vertical axis shows the odds ratios from the regression analyses. The bottom row shows the explained variance (R^2). Bars marked with *$P < 0.05$; TAG: Tobacco and Genetics Consortium; CPD: cigarettes per day

Table 4 Correlation between polygenic risk scores at Pt = 0.2

Risk profile at Pt = 0.2	*AOS*	*Ever*	*Former*	*CPD*
AOS	1			
Ever	−0.1183*	1		
Former	−0.1562*	−0.0346	1	
CPD	0.0831*	0.0200	−0.2044*	1

AOS = age at onset of smoking; CPD = cigarettes per day; ever = ever/never regular smoker; former = smoking cessation yes/no. *Significant correlations ($P < 0.05$). Pt = *P*-value thresholds.

individual effect sizes, influencing both smoking behaviour and alcohol and cannabis use. This analysis provides the first evidence that aggregated genetic risk factors are shared between substances. The finding that genetic variants have cross-substance effects is an important step towards understanding the common co-occurrence of the use of different substances. Our findings suggest that besides 'substance-specific' genes, we will also have to search for 'general substance-use' genes.

Acknowledgements

This study was supported by the European Research Council (ERC Starting Grant 284167 Principal Investigator Vink, ERC Advanced Grant 230374 Principal Investigator Boomsma), Netherlands Organization for Scientific Research (NWO: MagW/ZonMW grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717, Addiction-31160008 Middelgroot-911-09-032, Spinozapremie 56-464-14192), Center for Medical Systems Biology (CSMB, NWO Genomics), NBIC/BioAssist/RK (2008.024), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL, 184.021.007), the VU University's Institute for Health and Care Research (EMGO+), Neuroscience Campus Amsterdam (NCA), the European Science Foundation (ESF, EU/QLRT-2001-01254), the European Community's Seventh Framework Program (FP7/2007–2013), ENGAGE (HEALTH-F4-2007-201413), Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the Avera Institute, Sioux Falls, South Dakota (USA), the National Institutes of Health (NIH, R01D0042157-01A, NIH DA-18673, DA-026119, Principal Investigator Neale), the Genetic Association Information Network (GAIN) of the Foundation for the US National Institutes of Health, and the US National Institutes of Mental Health (NIMH, MH081802, 1RC2MH089951-01 Principal Investigator Sullivan, 1RC2 MH089995-01 Principal Investigator Hudizak).

Declaration of interests

None.

References

1. John U., Meyer C., Rumpf H.-J., Schumann A., Thyrian J. R., Hapke U. Strength of the relationship between tobacco smoking, nicotine dependence and the severity of alcohol dependence syndrome criteria in a population-based sample. *Alcohol Alcohol* 2003; **38**: 606–12.
2. Agrawal A., Budney A. J., Lynskey M. T. The co-occurring use and misuse of cannabis and tobacco: a review. *Addiction* 2012; **107**: 1221–33.
3. Li M. D., Cheng R., Ma J. Z., Swan G. E. A meta-analysis of estimated and environmental effects on smoking behavior in male and female adult twins. *Addiction* 2003; **98**: 23–31.
4. Sullivan P. F., Kendler K. S. The genetic epidemiology of smoking. *Nicotine Tob Res* 1999; **1**: S51–S57.
5. Vink J. M., Willemsen G., Boomsma D. I. Heritability of smoking initiation and nicotine dependence. *Behav Genet* 2005; **35**: 397–406.
6. van Beek J. H., Kendler K. S., de Moor M. H., Geels L. M., Bartels M., Vink J. M. *et al*. Stable genetic effects on symptoms of alcohol abuse and dependence from adolescence into early adulthood. *Behav Genet* 2012; **42**: 40–56.
7. de Moor M. H., Vink J. M., van Beek J. H., Geels L. M., Bartels M., de Geus E. J. *et al*. Heritability of problem drinking and the genetic overlap with personality in a general population sample. *Front Genet* 2011; **2**: 76.
8. Heath A. C., Martin N., Lynskey M. T., Todorov A. A., Madden P. A. F. Estimating two-stage models for genetic influences on alcohol, tobacco or drug use initiation and dependence vulnerability in twin and family data. *Twin Res* 2002; **5**: 113–24.
9. Vink J. M., Wolters L. M., Neale M. C., Boomsma D. I. Heritability of cannabis initiation in Dutch adult twins. *Addict Behav* 2010; **35**: 172–4.
10. Distel M. A., Vink J. M., Bartels M., van Beijsterveldt C. E., Neale M. C., Boomsma D. I. Age moderates non-genetic influences on the initiation of cannabis use: a twin-sibling study in Dutch adolescents and young adults. *Addiction* 2011; **106**: 1658–66.
11. Verweij K. J., Zietsch B. P., Lynskey M. T., Medland S. E., Neale M. C., Martin N. G. *et al*. Genetic and environmental influences on cannabis use initiation and problematic use: a meta-analysis of twin studies. *Addiction* 2010; **105**: 417–30.
12. Kendler K. S., Neale M. C., Sullivan P., Corey L. A., Gardner C. O., Prescott C. A. A population-based twin study in women of smoking initiation and nicotine dependence. *Psychol Med* 1999; **29**: 299–308.
13. Kendler K. S., Schmitt E., Aggen S. H., Prescott C. A. Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. *Arch Gen Psychiatry* 2008; **65**: 674–82.
14. Koopmans J. R., Doornen Van L. J. P., Boomsma D. I. Association between alcohol use and smoking in adolescent and young adult twins: a bivariate genetic analysis. *Alcohol Clin Exp Res* 1997; **21**: 537–46.
15. Young S., Rhee S. H., Stallings M., Corley R., Hewitt J. Genetic and environmental vulnerabilities underlying adolescent substance use and problem use: general or specific? *Behav Genet* 2006; **36**: 603–15.
16. Madden P. A. F., Heath A. C. Shared genetic vulnerability in alcohol and cigarette use and dependence. *Alcohol Clin Exp Res* 2002; **26**: 1919–21.
17. Uhl G. R., Liu Q. R., Drgon T., Johnson C., Walther D., Rose J. E. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. *BMC Genet* 2007; **8**: 10.
18. Liu Q. R., Drgon T., Johnson C., Walther D., Hess J., Uhl G. R. Addiction Molecular genetics: 639,401 SNP whole genome association identifies many 'cell adhesion' genes. *Am J Med Genet B Neuropsychiatr Genet* 2006; **141B**: 918–25.
19. Bierut L. J., Madden P. A. F., Breslau N., Johnson E. O., Hatsukami D., Pomerleau O. *et al*. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007; **16**: 24–35.
20. Vink J. M., Smit A. B., de Geus E. J., Sullivan P., Willemsen G., Hottenga J. J. *et al*. Genome-wide association study of smoking initiation and current smoking. *Am J Hum Genet* 2009; **84**: 367–79.
21. Amos C. I., Spitz M. R., Cinciripini P. Chipping away at the genetics of smoking behavior. *Nat Genet* 2010; **42**: 366–8.
22. Liu J. Z., Tozzi F., Waterworth D. M., Pillai S. G., Muglia P., Middleton L. *et al*. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42**: 436–40.
23. Thorgeirsson T. E., Gudbjartsson D. F., Surakka I., Vink J. M., Amin N., Geller F. *et al*. Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010; **42**: 448–53.
24. Consortium T. A. G. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–7.
25. Agrawal A., Lynskey M. T. Candidate genes for cannabis use disorders: findings, challenges and directions. *Addiction* 2009; **104**: 518–32.
26. Verweij K. J. H., Vinkhuyzen A. A. E., Benyamin B., Lynskey M. T., Quaye L., Agrawal A. *et al*. The genetic aetiology of cannabis use initiation: a meta-analysis of genome-wide association studies and a SNP-based heritability estimation. *Addict Biol* 2013 Sep; **18**: 846–50.
27. Agrawal A., Lynskey M. T., Hinrichs A., Grucza R., Saccone S. F., Krueger R. *et al*. A genome-wide association study of DSM-IV cannabis dependence. *Addict Biol* 2011; **16**: 514–8.
28. Rietschel M., Treutlein J. The genetics of alcohol dependence. *Ann NY Acad Sci* 2013; **1282**: 39–70.
29. Symons M., Weng J., Diehl E., Heo E., Kleiber M., Singh S. Delineation of the role of nicotinic acetylcholine receptor genes in alcohol preference in mice. *Behav Genet* 2010; **40**: 660–71.
30. Lubke G. H., Hottenga J. J., Walters R., Laurin C., de Geus E. J., Willemsen G. *et al*. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol Psychiatry* 2012; **72**: 707–9.
31. Shaun M. P., Naomi R. W., Jennifer L. S., Peter M. V., Michael C. O. D., Patrick F. S. *et al*. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–52.
32. Demirkan A., Penninx B. W. J. H., Hek K., Wray N. R., Amin N., Aulchenko Y. S. *et al*. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. *Mol Psychiatry* 2011 Jul; **16**: 773–83.

33. Luciano M., Huffman J. E., Arias-Vásquez A., Vinkhuyzen A. A. E., Middeldorp C. M., Giegling I. *et al*. Genome-wide association uncovers shared genetic effects among personality traits and mood states. *Am J Med Genet B Neuropsychiatr Genet* 2012; **159B**: 684–95.
34. Middeldorp C. M., Moor M. H. M. D., McGrath L. M., Gordon S. D., Blackwood D. H., Costa P. T. *et al*. The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data. *Transl Psychiatry* 2011; **1**: e50. doi: 10.1038/tp.2011.45
35. Cross-Disorder Group of the Psychiatric Genomics Consortium; Genetic Risk Outcome of Psychosis (GROUP) Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013; **381**: 1371–9.
36. Vorstman J. A. S., Anney R. J. L., Derks E. M., Gallagher L., Gill M., de Jonge M. V. *et al*. No evidence that common genetic risk variation is shared between schizophrenia and autism. *Am J Med Genet B Neuropsychiatr Genet* 2013; **162**: 55–60.
37. van Scheltinga A. F. T., Bakker S. C., van Haren N. E. M., Derks E. M., Buizer-Voskamp J. E., Cahn W. *et al*. Schizophrenia genetic variants are not associated with intelligence. *Psychol Med* 2013; **43**: 2563–2570.
38. Tobacco and Genetic Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–7.
39. Boomsma D. I., de Geus E. J., Vink J. M., Stubbe J. H., Distel M. A., Hottenga J. J. *et al*. Netherlands Twin Register: from twins to twin families. *Twin Res Hum Genet* 2006; **9**: 849–57.
40. Willemsen G., Vink J. M., Abdellaoui A., den Braber A., van Beek J. H. D. A., Draisma H. H. M. *et al*. The adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res Hum Genet* 2013; **16**: 271–81.
41. Willemsen G., de Geus E. J. C., Bartels M., Van Beijsterveldt C. E. M., Brooks A. I., Estourgie-Van Burk G. F. *et al*. The Netherlands Twin Register Biobank: a resource for genetic epidemiological studies. *Twin Res Hum Genet* 2010; **13**: 231–45.
42. Evans D. M., Visscher P. M., Wray N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 2009; **18**: 3525–31.
43. Belsky D. W., Moffitt, T. E., Baker, T. B., Biddle, A. K., Evans, J. P., Harrington, H. *et al*. Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA Psychiatry* 2013; **70**: 534–42.
44. Vrieze S., McGue M., Iacono W. The interplay of genes and adolescent development in substance use disorders: leveraging findings from GWAS meta-analyses to test developmental hypotheses about nicotine consumption. *Hum Genet* 2012; **131**: 791–801.
45. Hardie T. L., Moss H. B., Lynch K. G. Genetic correlations between smoking initiation and smoking behaviors in a twin sample. *Addict Behav* 2006; **31**: 2030–7.
46. Broms U., Silventoinen K., Madden P. A. F., Heath A. C., Kaprio J. Genetic architecture of smoking behavior: a study of Finnish adult twins. *Hum Genet* 2006; **9**: 64–72.
47. Madden P. A. F., Heath A. C., Pedersen N. L., Kaprio J., Koskenvuo M. J., Martin N. G. The genetics of smoking persistence in men and women: a multicultural study. *Behav Genet* 1999; **29**: 423–44.
48. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–52.
49. Davies G., Tenesa A., Payton A., Yang J., Harris S. E., Liewald D. *et al*. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol Psychiatry* 2011; **16**: 996–1005.
50. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLOS Genet* 2013; **9**: e1003348.
51. Visscher P. M., Yang J., Goddard M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang *et al*. (2010). *Twin Res Hum Genet* 2010; **13**: 517–23.
52. Wray N. R., Goddard M. E., Visscher P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* 2007; **17**: 1520–8.
53. Vink J. M., Beem A. L., Posthuma D., Neale M. C., Willemsen G., Kendler K. S. *et al*. Linkage analysis of smoking initiation and quantity in Dutch sibling pairs. *Pharmacogenomics J* 2004; **4**: 274–82.
54. Vink J. M., Posthuma D., Neale M. C., Eline Slagboom P., Boomsma D. I. Genome-wide linkage scan to identify loci for age at first cigarette in Dutch sibling pairs. *Behav Genet* 2006; **36**: 100–11.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. (a,b,c,d) Overview of results from linear (continuous variables) and logistic (dichotomous variables) regression analyses to predict smoking (ever smoked, cigarettes per day and age at first cigarette) in the target sample with polygenic risk scores for smoking. Polygenic scores are based on meta-analyses summary data of the Tobacco and Genetics consortium. Regression analyses are carried out on independent samples from the Netherlands Twin Register (NTR). Sex was used as covariate in the regression analyses. *P* = *P*-value from the regression analyses; R^2 = the explained variance; β = regression coefficient; standardized β/OR = standardized beta from linear regression analyses or odds ratio from logistic regression. (a) Polygenic risk score for age at onset of smoking. (b) Polygenic risk score for ever smoking. (c) Polygenic risk score for former smoking. (d) Polygenic risk scores for CPD.

Table S2 Best-fitting models when predicting smoking variables [dependent variables: ever smoked, cigarettes per day (CPD), age at regular smoking] with all four polygenic risk scores at Pt = 0.2 [age at onset of smoking (AOS), ever, former, CPD] simultaneously and sex and birth cohort as predictors in a regression analysis. A backward method is used, with *P*-value threshold of 0.05. Only the variables with at least one significant risk score are shown: (a) ever smoked, (b) CPD.