Supplemental Materials

Materials and Methods

Structural MRI High-resolution anatomical magnetic resonance images were acquired, including a 3D T1-weighted magnetization prepared gradient echo sequence based on the ADNI protocol (http://adni.loni.usc.edu). Structural MRI processing included data segmentation and normalization to the Montreal Neurological Institute template using the SPM optimized normalization routine. Gray matter images were modulated, thus facilitating comparisons of volumetric, rather than tissue concentration, differences (Ashburner & Friston, 2000). Whole-brain gray matter images were then used during ROI parcelation. At each gray matter ROI, the ratio of regional to total gray matter volume was calculated to account for anatomical variability across participants.

Functional MRI Full details of the magnetic resonance imaging (MRI) acquisition protocols and quality checks have been described previously, including an extensive period of standardization across MRI scanners (Schumann *et al.*, 2010). MRI Acquisition Scanning was performed at the eight IMAGEN assessment sites (London, Nottingham, Dublin, Mannheim, Dresden, Berlin, Hamburg, and Paris) with 3T whole body MRI systems made by several manufacturers (Siemens: 4 sites, Philips: 2 sites, General Electric: 1 site, and Bruker: 1 site). To ensure a comparison of MRI data acquired on these different scanners, we implemented image acquisition techniques using a set of parameters compatible with all scanners that were held constant across sites, for example, those directly affecting image contrast or fMRI preprocessing. Site was dummy-coded for use in the machine learning procedure.

Standardized hardware for visual and auditory stimulus presentation (NordicNeurolabs, Bergen Norway, http://www.nordicneurolab.com) was used at all sites. BOLD functional images were acquired with a gradient-echo echo planar imaging (EPI) sequence using a relatively short echo-time to optimize imaging of subcortical areas. Briefly, the functional imaging processing was as follows: Time series data were first corrected for slice-timing, then corrected for movement, non-linearly warped onto MNI space using a custom EPI template, and Gaussian-smoothed at 5mm-full width half maximum. Nuisance

variables were also added to the design matrix: estimated movement was added in the form of 12 additional regressors (3 translations, 3 rotations, 3 translations shifted 1 TR before and 3 translations shifted 1 TR later). Each individual fMRI time series underwent automatic spike detection, using a mean-squared based metric to identify unexpected values temporally and spatially slice per slice. Time-points with artifacts (if any) of each sequence were regressed out of each participant's data by adding a corresponding number of regressors with value 1 at the time- point of the artifact and 0 elsewhere to the design matrix.

Genotyping DNA purification and genotyping was performed by the Centre National de Génotypage in Paris. DNA was extracted from whole blood samples preserved in ethylene-eiamine-tetra-acetic acid vacutainer tubes (BD, Becton, Dickinson and Company, Oxford, United Kingdom) using Gentra Puregene Blood Kit (QIAGEN, Valencia, California) according to the manufacturer's instructions. Genotype information was collected at 582,892 markers using the Illumina HumanHap610 and HumanHap660 Genotyping BeadChips (San Diego, California). The SNPs with call rates of <95%, minor allele frequency < 1%, deviation from the Hardy-Weinberg equilibrium ("HWE", $p \leq 1 \times 10^{-6}$), and non-autosomal SNPs were excluded.

Markers data imputation and quality control for ambiguous SNPs, low MAF, missingness and HWE were done with MACH (Li *et al.*, 2010), following the ENIGMA2 guidelines. The 1000 Genomes project reference set of markers (http://www.internationalgenome.org) was used for the imputation after decreasing the markers from ~41 million to ~13 million relevant genetic variants observed more than once in the European populations. Four multidimensional scaling (MDS) components were calculated using a metric model in PLINK v1.9 (http://zzz.bwh.harvard.edu/plink/). MDS was then included as a covariate to account for population stratification as part of the cross-validation logistic regression, where the genotypes were coded following an additive model (as 0, 1, and 2 for the number of risk alleles).

Functional Tasks Descriptions

Stop Signal Task (SST) The SST required volunteers to respond to regularly presented visual go stimuli (arrows pointing left or right) but to withhold their motor response when the go stimulus was followed unpredictably by a stop-signal (an arrow pointing upwards). Stopping difficulty was manipulated across trials by varying the delay between the onset of the go arrow and the stop arrow (stop-signal delay, SSD) using a previously described tracking algorithm (Rubia *et al.*, 2005). A block contained 400 go trials and 80 variable delay stop trials with between 3 and 7 go trials between two stop trials. Stimulus duration in go trials was 1000 ms and in stop trials varied (0– 900ms in 50 ms steps) in accordance with the tracking algorithm (initial delay = 250 ms). We calculated contrast images for successful inhibitions ("stop success") and unsuccessful inhibitions ("stop fail"), both vs. an implicit baseline.

Monetary Incentive Delay The Monetary Incentive Delay (MID) task (adapted from a task described previously, Knutson *et al.*, 2001) required participants to respond to a briefly presented target by pressing either a left-hand or right-hand button as quickly as possible to indicate whether the target appeared on the left or the right side of the monitor display. If the participants responded while the target was on the screen, they scored points but if they responded before the target appeared or after the offset of the target they received no points. A cue preceded the onset of each trial, reliably indicating the position of the target and the number of points awarded for a successful response. A triangle indicated no points (No Win), a circle with one line 2 points (Small Win) and a circle with three lines 10 points (Large Win). Twenty-two trials of each type were presented in a pseudo-random order. The duration of the target was adjusted adaptively so that 66% of the trials produced a correct response. The participants were informed that at the end of the session they would receive one candy (M&M) for every five points won. We calculated contrast images for the anticipation period of Large Win minus No Win, and the outcome period for Large Win minus No Win.

Face Task The Face task involved passive viewing of video clips that displayed ambiguous (emotionally ''neutral'') or angry face expressions or control (nonbiological motion) stimuli (Grosbras, 2005). Each trial consisted of short (2 to 5 s) black-and-white video clips depicting either a face in movement or the

control stimulus. The control stimuli consisted of black-and-white concentric circles of various contrasts, expanding and contracting at various speeds, roughly matching the contrast and motion characteristics of the face clips. The stimuli were presented through goggles (Nordic Neurolabs, Bergen, Norway) in the scanner and subtended a visual angle of 10^{o} by 7^{o}. The video clips were arranged into 18-s blocks; each block included seven to eight video clips. Five blocks of each biological-motion condition (neutral and angry faces), and nine blocks of the control condition (circles) were intermixed and presented to the participant in a 6-minute run. We calculated contrast images from angry faces minus control stimuli, neutral faces minus control stimuli, and angry faces minus neutral faces. After the scanning session, participants completed a recognition task in which they were presented with three of the faces previously presented in the scanning session and two novel faces.

Personality

NEO Broad dimensions of personality were assessed using the 60-item Neuroticism- Extraversion-Openness Five-Factor Inventory (NEO-FFI), which returns measures on the dimensions of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience as described in the Five-Factor Model of personality (Costa Jr. & McCrae, 1995). The Extraversion factor assesses preference for seeking and engaging in social interactions and may be linked to sensitivity to rewarding environmental cues(Watson & Clark, 1992). The Agreeableness factor assesses empathy and an individual's tendency towards compassion and co-operation rather than self-interest. Conscientiousness provides a measure of the degree to which a participant exercises self- discipline and expresses a preference for planned, rather than spontaneous, behavior. The Neuroticism factor captures emotional liability and a tendency to experience lowered mood and elevated anxiety. Openness to Experience measures intellectual curiosity and creativity; lower scores on 'openness' are associated with a reduced tolerance for change and a preference for familiarity over novelty. Hence, five mean scores for each personality dimension for the child, and five mean scores for the parent, were included in the analysis.

Substance Use Risk Profile Scale The Substance Use Risk Profile Scale (SURPS; Woicik *et al.*, 2009)

assesses personality traits that confer risk for substance misuse and psychopathology. This scale measures four distinct and independent personality dimensions; anxiety sensitivity, negative thinking, sensation seeking, and impulsivity. The anxiety sensitivity dimension is characterized by the fear of symptoms of physical arousal. The negative thinking dimension is identified as a risk factor for the development of depression and characterized by dismal feelings. The sensation seeking dimension is characterized by the desire for intense and novel experiences. The impulsivity dimension involves difficulties in the regulation (controlling) of behavioral responses. Hence, four mean scores for each personality dimension for the child, and for the parent, were included in the analysis.

Temperament and Character Inventory The novelty seeking scale of the Temperament and Character Inventory – Revised (TCI-R; Cloninger, 1999) was administered. The novelty seeking scale contains four subscales (impulsiveness, disorderliness, excitability, and extravagance). The impulsiveness subscale describes behavior on a dimension from impulsivity to reflection and captures elements of emotional reactivity, and unreflective, careless behavior. The disorderliness subscale reflects disorganized, uncontrolled, and anti-normative behavior. The excitability subscale contrasts with 'stoic rigidity' to convey novelty-seeking and sensation-seeking behaviors. The extravagant subscale assesses overspending behaviors and diminished planning, and conveys a tendency to approach reward cues. Novelty seeking personality is assessed as the sum across all four subscales. Hence, five mean scores for each personality dimension for the child, and five for the parent, were included in the analysis.

Cognition

Wechsler Intelligence Scale for Children. Participants completed a version of the Wechsler Intelligence Scale for Children WISC-IV(Wechsler, 2003), of which we included the following subscales. Perceptual Reasoning, consisting of Block Design (arranging bi-colored blocks to duplicate a printed image) and Matrix Reasoning (in which a series of colored matrices are presented and the child is asked to select the consistent pattern from a range of options). Verbal Comprehension consisting of Similarities (two similar but different objects or concepts are presented and the child is asked to explain how they are alike or

different) and Vocabulary (a picture is presented or a word is spoken aloud by the experimenter and the child is asked to provide the name of the depicted object or to define the word).

Delay Discounting The Monetary-Choice Questionnaire (MCQ; Kirby *et al.*, 1999) was administered to provide a measure of preference for immediate lower over delayed higher monetary rewards. The MCQ is a 27-item task in which the participant chooses between a smaller, immediate monetary reward and a larger, delayed monetary reward (e.g. €25 today or €60 in 14 days), with varying discrepancies and delays between the rewards. The task indexes impulsivity by providing a measure of the degree to which future rewards are diminished or discounted. The protocol is scored by calculating where the participant's answers place them in comparison to reference discounting curves, where placement amid steeper curves indicates higher levels of impulsivity. A single delay discounting measure, "k", was included in the analysis.

CANTAB Participants completed five of the CANTAB tests. The Affective Go/No-go task comprised of alternating blocks in which participants were presented with positively or negatively valenced target words embedded in a stream of neutral distracter words. Participants were instructed to respond to targets with a button press. Measures included in the analyses were the total number of omissions to positive and negative targets, and the average response latency to positive and negative target words.

In the Pattern Recognition Memory task participants were required to remember 12 abstract patterns; the percentage of patterns correctly recognized on a two alternative forced choice task completed immediately after encoding was included in the analyses.

The Spatial Working Memory Task required participants to "search" for a token hidden by one of a number of boxes on the monitor by selecting the boxes in sequence. Once the token is uncovered, participants must search again with the condition that the token will not be hidden in the same location more than once. The number of times participants returned to search a box that had already contained the token was entered into the analyses as an error measure. We also included a strategy score (ranging from 1-37, with lower scores indicating a more strategic approach), which reflects how often a search sequence was initiated from a novel position.

The Rapid Visual Information Processing task comprised of a stream of digits presented at 1.67Hz and participants were required to monitor the stream for target sequence of three digits. We included a signal detection measure of sensitivity to the target sequence in the analyses.

The Cambridge Guessing Task (CGT) was a modified version of the Cambridge Gambling Task, renamed in order to make it appropriate to administer to adolescents. On each trial of the CGT the participant was presented with 10 boxes, some of which are blue, some of which are red, and must "guess" which color box conceals a hidden yellow token. Participants start the task with 100 points and lose or acquire points by wagering on their guess. The options the participant can choose to wager are determined by the program as a proportion of their total number of points, presented in either increasing or decreasing amounts. The analyses included measures of the time taken to select the option on which to bet, an average of the proportion of the total number of points wagered on each trial, the proportion of trials on which the more likely outcome was selected (quality of decision making), an average of the proportion wagered on trials when the participant selected the more likely result (rational bets), and an index of delay aversion reflected in making higher bets when the amount to bet is presented in descending order rather than in ascending order.

Behavioral data from functional imaging tasks. Behavioral data from the Monetary Incentive Delay (reward) task were as follows: the number of Big Win trials on which the target was not hit, the number of Big Win trials on which the target was hit, the number of Small Win trials on which the target was not hit, the number of Small Win trials on which the target was hit, the number of No Win trials on which the target was not hit, and the number of No Win trials on which the target was hit. Behavioral data from the Faces (emotional reactivity) task included the number of targets and the number of foils correctly categorized. Participants were not informed prior to the scanning session about the subsequent recall task. Behavioral data from the stop signal task was incomplete due to technical errors, therefore this data was omitted from the modeling procedures, however, the stop signal task had an adaptive performance algorithm to account for individual differences in reaction time.

History

Life-Events Questionnaire The Life-Events Questionnaire (LEQ) is an adaptation of the Stressful Life-Event Questionnaire(Newcomb *et al.*, 1981), which uses 39 items to measure the lifetime occurrence (frequency) and the perceived desirability of stressful events covering the following domains: Family/Parents, Accident/Illness, Sexuality, Autonomy, Deviance, Relocation, and Distress. The life-events valence labels measured on an ordinal scale from -2 to +2 as follows: -2='Very Unhappy', -1='Unhappy', 0='Neutral', +1='Happy', +2='Very Happy'. Hence, six measures related to the frequency, and six measures related to the valence, for each domain were included in the analysis.

Gestational cigarette and alcohol exposure. The Pregnancy and Birth Questionnaire (PBQ, adapted from Pausova *et al.*, 2007) assesses exposure of the child to potentially harmful conditions and substances such as maternal alcohol, cigarette, and cannabis use before and during pregnancy. The questionnaire was completed by each participant's parent or guardian and parental cigarette and alcohol use during pregnancy were recorded, then recoded as binary variables.

Alcohol Misuse. Michigan Alcohol Screening Test questions (MAST; Selzer, 1971), such as 'have you ever been in a hospital because of drinking', was used to assess alcohol misuse in the parent. A signle measure based off the summation of dependency items and a single binary measure for alcoholism was included in the analysis.

Puberty Development Scale. The Puberty Development Scale (PDS, Carskadon & Acebo, 1993) was used to assess the pubertal status of each participant. This scale provides an eight-item self-report measure of physical development based on the Tanner stages with separate forms for males and females. For this scale, there are five categories of pubertal status: 1= prepubertal, 2=beginning pubertal, 3=midpubertal, 4=advanced pubertal, 5=postpubertal. Participants answered questions about their growth in stature and pubic hair, as well as menarche in females and voice changes in males.

Socioeconomic Status. The socioeconomic status score was comprised of the sum of the following variables: Mother's Education Score, Father's Education Score, Family Stress Unemployment Score,

Financial Difficulties Score, Home Inadequacy Score, Neighborhood Score, Financial Crisis Score, Mother Employed Score, Father Employed Score.

ESPAD Quality Assurance As the Psytools program was run at the participant's home without direct supervision by the research team, the reliability of the data were checked in a two-stage procedure. Automated flags highlighted potentially problematic testing situations and were followed-up by research assistants face-to-face with the volunteer in a confidential setting. Final reliability ratings were assigned which led to exclusion of the data. Exclusion criteria for substance use measures included an indication that the participant was in a hurry, somebody was watching, or an indication to have known or taken the sham drug "Relevin". Inconsistency between baseline (age 14) and follow up (age 16) for all drugs was also an exclusion criterion (e.g., scoring 1 for cannabis at age 14 years, but 0 at age 16 years).

The specific item used to assign group membership reads "On how many occasions IN YOUR WHOLE LIFETIME have you used marijuana (grass, pot) or hashish (hash, hash oil)?"

Analytic Methods

Sex-specific prediction analyses were run 100 times to account for the subtle differences in results incurred due to the random assignment of participants to folds. During *k*-fold cross-validation, the full sample of data is partitioned into subsamples of data, where *k* equals the number of partitions (or "folds") of the original starting sample. *k*-fold cross-validation then becomes an iterative process whereby a single fold is set aside as the test sample ("test fold"), and a "training model" is estimated on the observations in the remaining *k*-1 folds ("training folds"). The training model is then used to predict the observations in the set aside test fold, thereby ensuring the independence of the test fold sample. This procedure returns *k* final models. These analyses were implemented using the "glmnet" function in MATLAB (v. R2014a, Natick, MA). Results were thresholded to identify only the predictors that were present in at least six final models (from k=10) across all 100 runs within a use level analysis. Predictors passing this threshold were selected for use in *post-hoc* analyses. See supplemental tables 6 and 7 for each predictors count of runs passing this threshold for each use level.

Imputation of Missing Data Missing data for all three domains were replaced (where possible) by imputation. Continuous variables were replaced with the 95% trimmed mean derived according to the participant's site and sex taken from the whole IMAGEN database (N=2,462). Ordinal data were similarly replaced with the mode of that variable for the participant's site and sex.

Elastic-net regularization and feature selection. Regression with elastic-net regularization is an example of a sparse regression method, which imposes a hybrid of both L1- and L2-norm penalties (i.e., penalties on the absolute (L1-norm) and squared values of the β weights (L2-norm)). As such, the elastic-net penalizes both the sum of the squared and absolute values of the regression coefficients, effectively setting some coefficients to zero, thereby performing feature selection during model estimation. This allows relevant but correlated coefficients to coexist in a sparse model fit, by doing automatic variable selection and continuous shrinkage simultaneously, and selects or rejects groups of correlated variables.

The elastic-net estimation procedure is tuned using two parameters (α, λ). The α balances the contribution of the LASSO (L1-norm) to ridge (L2-norm) estimation methods. The second parameter, λ controls the magnitude of the shrinkage applied to the coefficients. The α and the λ values are tuned within a nested cross-validation procedure in order to identify the optimal set of parameter values that minimize the test error returned from evaluating model fit on an independent sample of observations. These tuning parameters are always non-negative values, such that $0 \leq \alpha \leq 1$ and $0 \leq \lambda$. For each sex-specific prediction analysis, a reduced design matrix with significant predictors [p x 10 **(**β-per-fold)] x 100 runs was returned.

Gene-only Analysis A gene-specific analysis was conducted collapsed across sex (N=1,581) and included only SNPs and nuisance covariates (age, sex, handedness, puberty status, and four MDS factors) as independent variables. The same threshold as explained in *Analytic Methods* above was applied to identify predictive SNPs. As the AUC was non-significant at the maximal use level (ESPAD ≥6), results from these analyses were not probed.

Head Motion As head motion has been shown to confound structural and functional MRI findings (Pardoe *et al.*, 2016), *a-priori* 2-sample *t*-tests confirmed that head motion (mean framewise

displacement, "FD") within each task did not differ between age 16 users and controls for either sex. The framewise displacement (FD) for each participant for each fMRI task was calculated using the six displacement parameters estimated during image realignment preprocessing procedures (see supplemental table 8 for statistics).

Binge Drinking Sample Participants in the binge drinking sample were identified as having no binge drinking episodes at age 14 (and a maximum of two lifetime drinks), and then going on to have any level of binge drinking episodes by age 16. Overlapping participants identified as cannabis users *and* binge drinkers by age 16 were excluded from the binge drinking sample (N=400; n=208 of whom transitioned to binge drinking by age 16).

***Post-hoc* Regression Modeling** Comparative model fit statistics were generated by computing the relative change in model fit between a model estimated using only the base rate (which corresponds to a threshold at which the exact baseline rate would be classified as a cannabis user or binge drinker by age 16) and a model estimated using only the predictors identified from the preceding prediction analyses ("model with predictors"). For *post-hoc* scenarios, independent variables were standardized to z-scores by sex before model estimation. Thereafter, coefficients and model fit statistics were freely estimated from their respective samples. *Post-hoc* regressions were executed in SPSS v. 24 (IBM Corp. Arnonk, NY).

Resampling of Males to Female Distribution The superior prediction for females is notable given that they were fewer in number and lighter in use compared to males (see supplemental table 1). To assess the impact that the differences in sample size and use severity might have on modeling, an analysis conducted on a smaller number of males, matched on sample size and use levels (from ESPAD ≥3 only) to the female cohort, was shown to worsen the prediction for males (mean ΔAUC= -.08 relative to full male sample).

Permutation Analyses To check for spurious findings and to confirm the independence of the 10-fold model training and testing procedure, random permutation analyses were conducted. Sex-specific prediction analyses (ESPAD ≥3 only) was repeated 100 times while randomly assigning group membership to each participant while keeping original group sample sizes consistent. Essentially, these

analyses tested for significant findings on a sample-generated null model. These permutation models performed no better than chance (Males: mean ROC AUC=0.53, σ=.02, p>.05; Females: mean ROC AUC=0.51, σ=.03, p>.05). In addition to these models failing to predict randomized outcomes, the predictors selected for each final model did not mirror the predictors selected from the true analyses.

Supplemental References

Ashburner, J. & Friston, K.J. (2000) Voxel-Based Morphometry—The Methods. *NeuroImage*., **11**, 805–821.

Carskadon, M.A. & Acebo, C. (1993) A self-administered rating scale for pubertal development. *J. Adolesc. Health*., **14**, 190–195.

Cloninger, C.R. (1999) The temperament and character inventory-revised. *St Louis MO Cent. Psychobiol. Personal. Wash. Univ.*

Costa Jr., P.T. & McCrae, R.R. (1995) Domains and Facets: Hierarchical Personality Assessment Using the Revised NEO Personality Inventory. *J. Pers. Assess*., **64**, 21–50.

Grosbras, M.-H. (2005) Brain Networks Involved in Viewing Angry Hands or Faces. *Cereb. Cortex*., **16**, 1087–1096.

Kirby, K.N., Petry, N.M., & Bickel, W.K. (1999) Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *J. Exp. Psychol. Gen*., **128**, 78–87.

Knutson, B., Fong, G.W., Adams, C.M., Varner, J.L., & Hommer, D. (2001) Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport*., **12**, 3683–3687.

Li, Y., Willer, C.J., Ding, J., Scheet, P., & Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol*., **34**, 816–834.

Newcomb, M.D., Huba, G.J., & Bentler, P.M. (1981) A Multidimensional Assessment of Stressful Life Events among Adolescents: Derivation and Correlates. *J. Health Soc. Behav*., **22**, 400.

Pardoe, H.R., Kucharsky Hiess, R., & Kuzniecky, R. (2016) Motion and morphometry in clinical and nonclinical populations. *NeuroImage*., **135**, 177–185.

Pausova, Z. et al. (2007) Genes, maternal smoking, and the offspring brain and body during adolescence: Design of the Saguenay Youth Study. *Hum. Brain Mapp*., **28**, 502–518.

Rubia, K., Smith, A.B., Brammer, M.J., Toone, B., & Taylor, E. (2005) Abnormal Brain Activation During Inhibition and Error Detection in Medication-Naive Adolescents With ADHD. *Am. J. Psychiatry*., **162**, 1067–1075.

Schumann, G. et al. (2010) The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry*., **15**, 1128–1139.

Selzer, M.L. (1971) The Michigan Alcoholism Screening Test: The Quest for a New Diagnostic Instrument. *Am. J. Psychiatry*., **127**, 1653–1658.

Watson, D. & Clark, L.A. (1992) On Traits and Temperament: General and Specific Factors of Emotional Experience and Their Relation to the Five-Factor Model. *J. Pers*., **60**, 441–476.

Wechsler, D. (2003) Wechsler intelligence scale for children–Fourth Edition (WISC-IV). *San Antonio TX Psychol. Corp.*

Woicik, P.A., Stewart, S.H., Pihl, R.O., & Conrod, P.J. (2009) The substance use risk profile scale: A scale measuring traits linked to reinforcement-specific substance use profiles. *Addict. Behav.*, **34**, 1042–1055.

Supplemental Tables: Titles and Legends

Supplemental Table 1: Comparison of Age 16 Dropouts vs. Retained Sample.
Participants who completed the baseline ESPAD assessment and reported no lifetime cannabis use but then were unavailable for follow up assessment two years later were assigned to the dropout sample. Compared to the retained sample, the dropout sample had significantly higher age, and lower IQs and SES.

Supplemental Table 2: Summary of data used as independent variables in predictive modeling.
A related analysis including psychopathology measures was conducted but did not improve predictive performance. Site was also modeled in the analysis and yielded Paris (data not shown) as a significant predictor due to the higher prevalence of cannabis use at age 16 for both sexes.

Supplemental Table 3: Binge Drinking Sample Demographics.
All participants at baseline reported no lifetime binge drinking episodes and a maximum of 2 lifetime alcoholic drinks. Participants who then went on to report any level of binge drinking by age 16 were included in the binge drinking at age 16 sample, compared to participants who endorsed a maximum of 2 lifetime drinks.

Supplemental Table 4: Post-hoc Regression Model Summaries.
Features identified from each cannabis predictive modeling scenario were used to probe sex- and drug-specific effects. Male & Female shared psychosocial predictors of cannabis use also predicted binge drinking by age 16. Male brain predictors and female brain predictors failed to model cannabis use in the opposite sex, or, binge drinking in the same sex. *ΔAIC always in reference to the better fitting model. $\Delta AIC = AIC_{model\_i} - AIC_{min}$ and reflects the relative increase in information gained from the AIC_{min} (better) model. Values >=2 favor the AIC_{min} model.

Supplemental Table 5: Statistics and Frequencies for Cannabis Predictive SNPs.
Measures of Hardy-Weinberg Equilibrium (HW), Minor Allele Frequency (MAF). Association with cannabis use by age 16 calculated using Spearman's rank correlation between SNP and the outcome measure collapsed across sex. H_{minor}: Homozygote minor (high-risk genotype), HT: heterozygote (intermediate-risk genotype), H_{major}: Homozygote major (low-risk genotype).

Supplemental Table 6: Frequency of Selected Male Features.
Count of the number of runs (out of 100) that a predictor was selected in at least 6 of 10 final models.

Supplemental Table 7: Frequency of Selected Female Features.
Count of the number of runs (out of 100) that a predictor was selected in at least 6 of 10 final models.

Supplemental Table 8: Analysis of Head Motion.
Framewise displacement was calculated from the six-directional head motion parameters estimated during image realignment. 2-sample *t*-tests on the participants endorsing any cannabis at age 16 vs. their non-using peers failed to detect significant differences in head motion (mean FD) for any of the tasks for either sex, with the exception of the faces task for females. The modest motion effect detected for the faces task in females is driven by outliers in the comparison sample. Exclusion of these participants does not affect predictive model performance. Furthermore, the faces task predictors were lower activity (with one exception) in the cannabis use sample therefore, any motion effects are likely non-influential.

Supplemental Figures: Titles and Legends

Supplemental Figure 1: Schematic of Analytic Method
First, data are divided into *k*(10) outer-folds. *k*-1 outer-folds are then divided into *k*(10) nested subfolds. Elastic-net regularized logistic regression applied to *k*-1 subfolds, during which the α, λ parameters are tuned by finding the optimal pair returning the highest AUC when it's model is tested on the *kth* subfold. The iterative process is completed for the *k*(10) subfolds, generating 10 final nested models. The 10 nested models are ranked by their AUC returned when tested on each respective test-fold. The highest-ranking model is then tested on the outer fold, and used to generate the reported test AUC. This process is repeated *k*-times, and the entire procedure executed 100 times.

Supplemental Figure 2: Receiver-operating characteristic (ROC) mean AUC for Gene-specific Analysis. ROC AUC indicates the performance of the predictive models on independent samples. This plot visualizes the mean AUC across 100 runs for each use level collapsed across sex.

Supplemental Figure 3: Correlations Between Identified SNPs and Outcome Measure by Sex.
Pearson's point-biserial correlation (*r*) between SNP and outcome. Error bars represent 95% confidence intervals generated from 5000 bootstrap samples.

Supplemental Table 1: Comparison of Age 16 Dropouts vs. Retained Sample

Measure	Groups		*p*
	Age 16 Dropouts (n=437)	**Retained Sample (n=1581)**	
Age *(M,SD)*	14.6, 0.41	14.5, 0.42	.002
Sex (Male, Female)	229, 208	745, 836	.051
Handedness (L,R)	37, 400	169, 1412	.174
PDS *(M,SD)*	3.6, 0.7	3.5, 0.8	.573
Perceptual IQ *(M,SD)*	104.5, 13.24	108.11, 13.8	.000
Verbal IQ *(M,SD)*	106.8, 14.8	111.2, 13.5	.000
SES *(M,SD)*	17.01, 4.5	18.00, 3.8	.000

Supplemental Table 2: Summary of data used as independent variables in predictive modeling.

Domain	Measures	Data points
Psychosocial	Demographics Cognitive assessments Personality assessment Life-events questionnaires Baseline cigarette & alcohol use Parent personality and drug use	• 80 measures
Genetic	A-priori SNPs • Cannabinoid Receptor • Catecholamine Receptors • Opioid Receptors	• 108 SNPs
Structural Neuroimaging	Total GMV Gray-Matter Volume ROIs	• 1 total GMV • 278 GMV ROIs
Functional Neuroimaging	Reward Processing Task • (2 Contrasts) Stop Signal Task • (2 Contrasts) Face Processing Task • (3 Contrasts)	• 1946 ROIs • 278 per contrast
Total predictors per subject		**2413**

Supplemental Table 3: Binge Drinking Sample Demographics.

Measure	Groups		*p*
	Binge Drinkers by age 16 (n=208)	**Comparison Group (n=192)**	
Age *(M,SD)*	14.5, 0.41	14.5, 0.39	.706
Sex (Male, Female)	103, 105	77, 115	.060
Handedness (L,R)	21, 171	20, 188	.663
PDS *(M,SD)*	2.8, 0.6	2.9, 0.6	.610
Perceptual IQ *(M,SD)*	106.2, 13.5	105.8, 14.3	.773
Verbal IQ *(M,SD)*	109.5, 13.1	108.5, 14.5	.505
SES *(M,SD)*	18.1, 3.7	17.8, 3.8	.785

Supplemental Table 4: Post-hoc Regression Model Summaries

Cannabis Predictive Features	**Test Sample**		**Model Fit**	
	Sex-Specificity	**Drug-Specificity**	**χ^2, p**	**ΔAIC***
Shared Psychosocial Features		Binge Drinking	29.6, p<.01	19.6 (base rate model –model with predictors)
Male Brain Features	Females: Cannabis Use		9.9, p >.05	6.1 (model with predictors –base rate model)
		Males: Binge Drinking	8.3, p >.05	7.6 (model with predictors –base rate model)
Female Brain Features	Males: Cannabis Use		18.8, p >.05	15.2 (model with predictors –base rate model)
		Females: Binge Drinking	16.6, p >.05	17.4 (model with predictors –base rate model)
Shared Genetic Features		Binge Drinking	9.03, p >.05	9 (model with predictors –nuisance model)

Supplemental Table 5: Statistics and Frequencies for Cannabis Predictive SNPs.

Locus	Gene	HW P value	MAF	Major: Minor Alleles	Imputation Quality (R^2)	Association with age 16 Cannabis Use		Genotype (% H_{minor} : HT : H_{major})		Minor Allele Effect On Cannabis Use
						r	*p*	Cannabis Use by age 16	Comparison Group	
rs1042711	ADRB2	.86	.122	T:C	.97	.06	.02	12:54:34	16:55:30	Protection
rs1801704	ADRB2	.86	.122	T:C	.97	.06	.02	12:54:34	16:55:30	Protection
rs6888306	ADRA1b	.92	.099	C:T	.89	.03	.25	3:33:64	4:32:64	Protection
rs686	DRD1	.85	.135	A:G	.85	-.03	.23	12:51:37	12:44:44	Risk
rs11746641	DRD1	.84	.060	T:G	.64	-.05	.05	4:25:71	2:22:76	Risk
rs2281617	OPRM1	.88	.098	G:T	.86	.01	.72	2:23:76	1:23:76	Risk
rs563649	OPRM1	.91	.158	G:A	.89	.03	.27	0:14:86	1:13:86	Protection
rs10485057	OPRM1	.89	.094	A:G	.87	.02	.41	1:13:86	1:14:85	Protection
rs1074287	OPRM1	.90	.256	A:G	.99	-.04	.15	9:34:57	8:29:63	Risk
rs511420	OPRM1	.87	.097	T:C	.99	-.04	.09	2:18:80	1:17:83	Risk

Supplemental Table 6: Frequency of Selected Male Features

		Analysis Levels					
Domain	**Feature**	**≥1x**	**≥3x**	**≥6x**	**≥10x**	**≥20x**	**≥40x**
Psychosocial	Lifetime Cigarette Use	100	100	100	100	100	100
	Parental Cannabis Use	100	100	100	98	100	100
	Feelings of Deviance	100	100	100	100	0	5
	Lifetime Alcohol Use	100	100	90	0	0	0
	Sensation Seeking Personality (Parent)	100	100	33	0	0	0
	Disorderly Personality	100	96	97	100	100	100
	Novelty Seeking Personality	40	100	100	100	69	0
	Novelty Seeking Personality (Parent)	28	88	100	20	0	0
Structural MRI	L. Mid-Cingulate Cortex	24	0	0	0	0	100
	R. Medial Prefrontal Cortex	0	0	0	0	0	100
Functional MRI	Stop Success: R.Midbrain-Thalamus	100	12	0	0	0	0
	Stop Success: L. Inferior Temporal Gyrus	84	9	2	15	100	100
	Stop Success: L. Post-Lateral Hemisphere	0	0	100	100	100	100
	Stop Success: L. Anterior Cerebellum	0	0	0	0	0	100
	Stop Success: L. Paravermis	0	1	97	0	100	98
	Neutral Faces: R.Midbrain-Thalamus	0	0	0	0	0	100

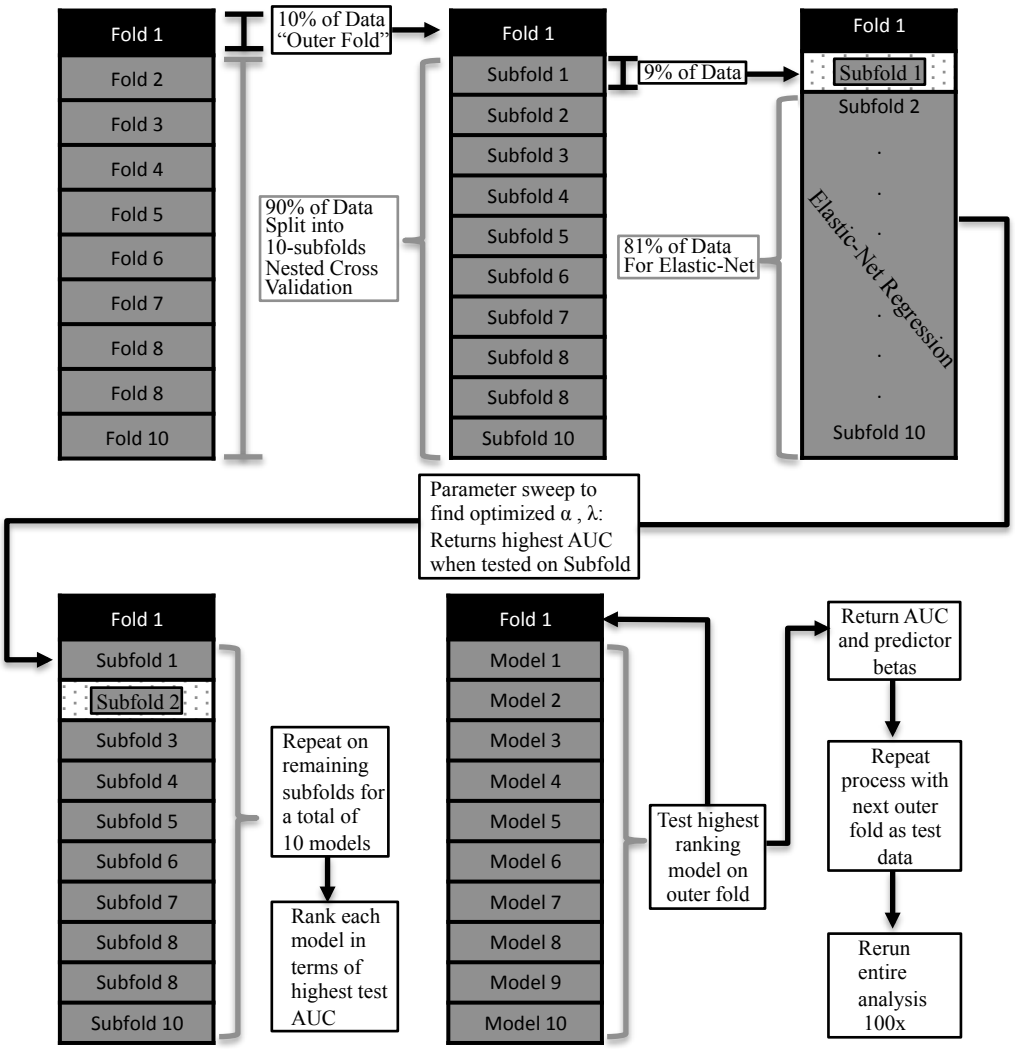
Supplemental Table 7: Frequency of Selected Female Features

		Analysis Levels					
Domain	**Feature**	**≥1x**	**≥3x**	**≥6x**	**≥10x**	**≥20x**	**≥40x**
Psychosocial	Lifetime Cigarette Use	100	100	100	100	100	100
	Lifetime Alcohol Use	100	100	100	100	100	50
	Novelty Seeking Personality	100	100	100	100	42	0
	Parental Cannabis Use	100	100	100	100	0	0
	Extravagant Personality (Parent)	100	100	87	97	0	0
	Feelings of Deviance	100	89	23	100	0	0
	Disorderly Personality	100	46	100	100	100	81
	Verbal IQ	100	17	0	0	0	0
	Impulsive Personality	100	0	0	0	0	0
	Frequency of Sexual Life Events	97	44	100	100	0	0
	Extravagant Personality	33	89	86	100	1	0
Structural MRI	R.Pre-Supplementary Motor Area	100	35	95	0	0	0
	R.Middle Frontal Gyrus	0	0	70	100	60	58
Functional MRI	Stop Success: L. Orbital Frontal Cortex	100	0	0	0	0	0
	Stop Success: R. Orbital Frontal Cortex	100	0	0	0	0	0
	Stop Success: R.Middle Temporal Gyrus	100	0	0	0	0	0
	Stop Success: R.Middle Temporal Gyrus	100	0	0	0	0	0
	Stop Failure: L.Midbrain	100	0	0	0	0	0
	Stop Failure: R.Post-Central Gyrus	100	0	0	0	0	0
	Stop Failure: R.Inferior Frontal Gyrus	100	4	0	0	0	0
	Stop Failure: R.Pre-Supplementary Motor Area	87	62	100	100	100	4
	Stop Failure: L.Lateral Paravermis	0	0	0	100	9	34
	Stop Failure: L.Pre-Post Central Gyrus	100	100	53	0	0	0
	Angry Faces: R.Anterior Cerebellum	100	0	0	1	0	5
	Angry Faces: L.Ventromedial Prefrontal Cortex	100	0	0	0	0	0
	Neutral Faces: R. Superior Frontal Gyrus	100	0	0	0	0	0
	Reward Anticipation: L.Middle Frontal Gyrus	100	0	0	0	0	0

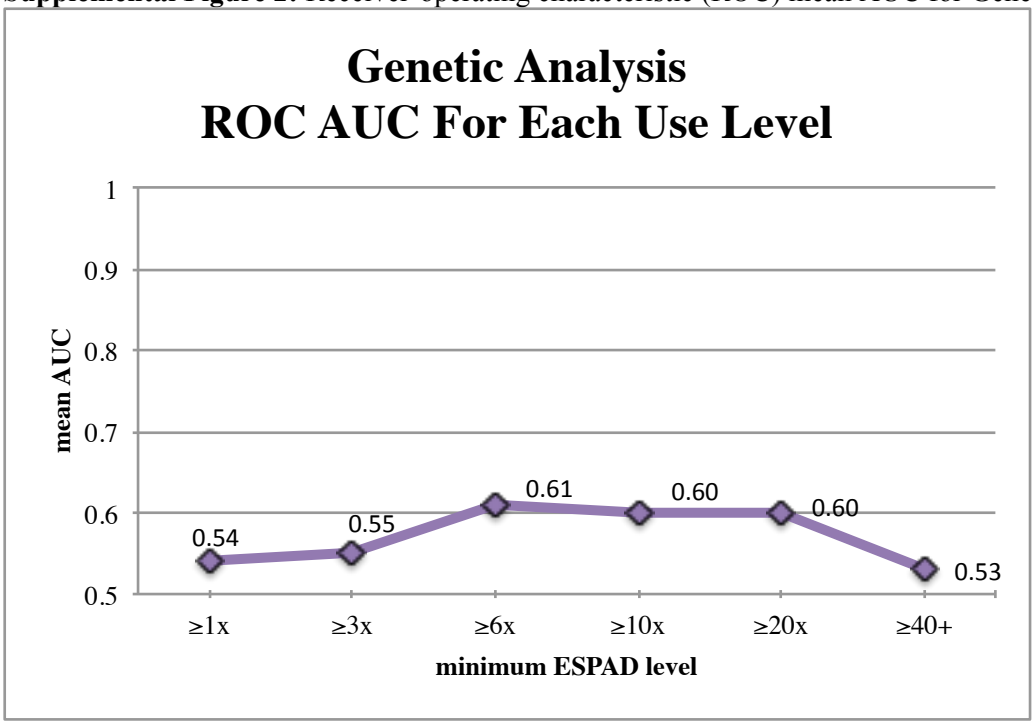
Supplemental Table 8: Analysis of Head Motion.

Sex	Task	Mean Framewise Displacement: Age 16 Users vs. Comparison Group
Males	Faces	t_{720}= -0.73, p > .05
	MID	t_{684}= -0.85, p > .05
	Stop Signal	t_{669}= -1.69, p > .05
Females	Faces	t_{806}= -2.09, p =.04
	MID	t_{772}= -0.22, p > .05
	Stop Signal	t_{765}= -1.00, p > .05

Supplemental Figure 1: Schematic of Analytic Method



Supplemental Figure 2: Receiver-operating characteristic (ROC) mean AUC for Gene-specific analysis



Supplemental Figure 3: Correlations Between Identified SNPs and Outcome Measure by Sex.