doi:10.1038/nature13402

METHODS AND MATERIALS

MRI Acquisition

Overview of IMAGEN protocols

Full details of the procedures employed by the IMAGEN study, including details on ethics, recruitment, standardized instructions for administration of the psychometric and cognitive behavioral measures, and for blood collection and storage are available to view in the Standard Operating Procedures for the IMAGEN project (http://www.imagen-europe.com/en/Publications_and_SOP.php).

FMRI acquisition and analysis

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[1]. MRI Acquisition Scanning was performed at the eight IMAGEN assessment sites (London, Nottingham, Dublin, Mannheim, Dresden, Berlin, Hamburg, and Paris) with a 3T whole body MRI system 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.

Structural MRI. High-resolution anatomical magnetic resonance images were acquired, including a 3D T1-weighted magnetization prepared gradient echo sequence (MPRAGE) based on the ADNI protocol (http://www.loni.ucla.edu/ADNI/Cores/index.shtml).

Structural MRI processing included data segmentation and normalization (to the Montreal Neurological Institute template) using the SPM[2] optimized normalization routine. Gray matter images were modulated, thus facilitating comparisons of volumetric, rather than tissue concentration differences[3].

Functional MRI. 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 echoplanar imaging (EPI) sequence using a relatively short echo-time to optimize imaging of subcortical areas (details of sequence parameters for structural and functional imaging, and the neuroimaging tasks are given in Ref 1). 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 timeseries 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 subject'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.

Functional Tasks description

Stop Signal Task. 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[4]. 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[5]) 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[6]. 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 (ambiguous, and angry faces), and nine blocks of the control condition (circles) were intermixed and presented to the subject in a 6-minute run. We calculated contrast images from angry faces minus ambiguous 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[7]. The Extraversion factor assesses preference for seeking and engaging in social interactions and may be linked to sensitivity to rewarding environmental cues[8]. 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 lability and a tendency to experience lowered mood and 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.

Substance Use Risk Profile Scale. The Substance Use Risk Profile Scale (SURPS[9]) assesses personality traits that confer risk for substance misuse and psychopathology. This scale measures four distinct and independent personality dimensions; anxiety sensitivity, hopelessness, sensation seeking, and impulsivity. The anxiety sensitivity dimension is characterized by the fear of symptoms of physical arousal. The hopelessness 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.

Temperament and Character Inventory. The novelty seeking scale of the Temperament and Character Inventory – Revised (TCI-R[10]) was administered. The Novelty seeking scale is composed of four sub-scales. Exploratory Excitability contrasts with 'stoic rigidity' and reflects sensation-seeking and novelty-seeking behaviors. Impulsiveness describes behavior on a dimension from impulsivity to reflection and captures elements of emotional reactivity, and unreflective, careless behavior. The Extravagance subscale assesses overspending behavior and poor planning and is believed to reflect a tendency to approach reward cues. Disorderliness reflects disorganized, uncontrolled, and antinormative behavior.

Cognition

Wechsler Intelligence Scale for Children. Participants completed a version of the Wechsler Intelligence Scale for Children WISC-IV[11], 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[12]) was administered to provide a measure of preference of 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.

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. Our 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. These data included the mean Go reaction time and the standard deviation of the Go reaction time from the Stop Signal Task (inhibitory control). 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.

History

Life-Events Questionnaire. The Life-Events Questionnaire (LEQ) is an adaptation of the Stressful Life-Event Questionnaire[13], which uses 39 items to measure the lifetime occurrence 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 were as follows: 'Very Unhappy', 'Unhappy' ,'Neutral' ,'Happy', 'Very Happy'.

Gestational cigarette and alcohol exposure. The Pregnancy and Birth Questionnaire (PBQ, adapted from[14]) assesses exposure of the child to potentially harmful conditions and

substances such as maternal alcohol and cigarette use before and during pregnancy, medical/physical conditions of child and mother as well as nutrition after birth. 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.

Family History classification. Subjects were classified into one of three categories: family history negative (a score of 0), neither positive nor negative (a score of 1), and family history positive (a score of 2). Subjects were classified as family history positive if at least one first-degree relative was positive for alcohol abuse, derived from information from the Family History questionnaire (general) or from diagnostic Michigan Alcohol Screening Test questions (MAST[15]), such as 'have you ever been in a hospital because of drinking'. Subjects were classified as family history negative if there was no alcoholism in 1st or 2nd degree relatives (Family History questionnaire), the Parent-AUDIT[16] was less than or equal to 3 and the Parent-MAST was less than or equal to 2 and there was no drug abuse in any relative.

Demographics

Puberty Development Scale. The Puberty Development Scale (PDS[17]) was used to assess the pubertal status of our adolescent sample. 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.

Substance misuse measures. The European School Survey Project on Alcohol and Drugs (ESPAD[18]) was administered using the software program Psytools[19] which is a computerized assessment platform. Psytools presented questionnaire items and response alternatives on a computer screen. Jump rules were implemented where applicable to skip irrelevant questions (e.g., drinking-related questions in self-reported non-drinkers) for the sake of brevity. As the Psytools program was run at the participant's home without direct supervision by the research team, the reliability of individual data was checked in a two-stage procedure. Before every task, adolescents were asked to report on the current testing context including questions about their attentional focus and the confidentiality of the setting. 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 years) and follow up (age 16 years) for all drugs was also an exclusion criterion (e.g., scoring 1 for cannabis at age 14 years, but 0 at age 16 years).

Stratified sampling. In order to ensure adequate representation of each group (control or binge drinker) in each fold, each outer and middle fold was balanced for the following variables: site, sex, handedness, pubertal development, performance and verbal IQ and age. Subjects were assigned randomly to folds 100,000 times and a cost function seeking to minimize the

differences in representation among the folds was calculated. The requirement to have equal numbers from either group in each fold was weighted by a factor of 10, requirement for equal sites was weighted by a factor of 4, and sex ratio by a factor of 3. The assignment of subjects to folds with the lowest cost function was retained for model generation and validation. Data for the inner folds were assigned randomly, as these data would later be collapsed across inner folds. Analysis 1 contained data from an additional control participant (i.e., n = 151). This participant's data were inadvertently removed from subsequent analyses.

Imputation of missing data. Missing personality, family history, life events, and IQ were replaced 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). Nominal data were replaced with the mode of that variable for the participant's site and sex. Extended Data Table 2 displays the amount of imputed data for each group and measure. Missing genetic data were imputed. First, PLINK[20] was used to extract the allelic values from the dataset, which was then examined for missing data. Most SNPs were missing from between 1-4 participants. 20 datasets were imputed, with the exception of (rs717207) that was missing data from 347 participants. To replace these data, multiple imputation was used within IBM SPSS Statistics 19.0 with fully conditional specification using all SNPs and sex as predictors and imputing the few missing data for each SNP from the others. All datasets replaced the missing SNPs in the same manner - essentially by replacing with the major allele. The imputed dataset was then used in subsequent genetic analyses.

Elastic Net regularization. Regression with Elastic Net[21] regularization is an example of a sparse regression method, which imposes a hybrid of both L_1- and L_2-norm penalties (i.e., penalties on the absolute (L_1 norm) and squared values of the β weights (L_2 norm)). 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. Least absolute shrinkage and selection operator (LASSO[22]) and ridge regression[23] are special cases of the Elastic Net.

References

1. Schumann, G. *et al.* The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Molecular Psychiatry* **15**, 1128–39 (2010).

2. Wellcome Department of Neuroimaging, London, United Kingdom; http://www.fil.ion.ucl.ac.uk/spm

3. Ashburner, J. & Friston, K. Voxel-based morphometry--the methods. *NeuroImage.* **11**, 805–821 (2000).

4. Rubia, K., Smith, A., Brammer, M., Toone, B., Taylor, E. Abnormal brain activation during inhibition and error detection in medication-naive adolescents with ADHD. *American Journal of Psychiatry* **162**, 1067–75 (2005).

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

6. Grosbras, M. -H. & Paus, T. Brain networks involved in viewing angry hands or faces. *Cerebral Cortex*.**16**, 1087–96 (2006).

7. Costa, P. & McCrae, R. Domains and facets: hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment* **64**, 21–50 (1995).

8. Watson, D. & Clark, L.A. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. *Journal of Personality* **60,** 441-476 (1992).

9. Woicik, P., Stewart, S. H., Pihl, R. O., Conrod, P. J. The Substance Use Risk Profile Scale: a scale measuring traits linked to reinforcement-specific substance use profiles. *Addictive Behaviors* **34**, 1042-1055 (2009).

10. Cloninger, C. R. The temperament and character inventory-revised. *St Louis, MO: Center for Psychobiology of Personality, Washington University* (1999).

11. Wechsler, David. "Wechsler intelligence scale for children–Fourth Edition (WISC-IV)." *San Antonio, TX: The Psychological Corporation* (2003).

12. Kirby, K. N., Petry, N. M., Bickel, W. K. Heroin addicts discount delayed rewards at higher rates than non-drug using controls. *Journal of Experimental Psychology: General* **128,** 78-87 (1999).

13. Newcomb, M. D., Huba, G. J., Bentler, P. M. A multidimensional assessment of stressful life events among adolescents: Derivation and correlates. *Journal of Health and Social Behavior* **22,** 400-415 (1981).

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

15. Selzer, M. L. The Michigan alcoholism screening test: the quest for a new diagnostic instrument. *American Journal of Psychiatry* **127**, 1653–1658 (1971).

16. Saunders, J. B. *et al.* Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* **88**, 791-804 (1993).

17. Petersen, A. C., Crockett, L., Richards, M. A self-report measure of pubertal status: Reliability, validity, and initial norms. *Journal of Youth and Adolescence* **17,** 117-133 (1988).

18. Hibell, B. *et al.* The 1995 ESPAD report: alcohol and other drug use among students in 26 European countries. *Stockholm: Swedish Council for Information on Alcohol and Other Drugs* (1997).

19. Psytools ® (John Rogers, Delosis Ltd, London, UK).

20. Shaun, P. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-575 (2007).

21. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** (2005).

22. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267-288 (1996).

23. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67 (1970).