

How does machine learning bias in prediction of behavioral phenotypes relate to individual characteristics?

Jianxiao Wu^{1,2,T,*}, Jingwei Li^{1,2,T}, Simon B. Eickhoff^{1,2}, B.T. Thomas Yeo^{3,4-8}, Sarah Genon^{1,2,*}

¹Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany

²Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

³Centre for Sleep & Cognition & Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

⁴Department of Electrical and Computer Engineering, National University of Singapore, Singapore

⁵N.1 Institute for Health, National University of Singapore, Singapore

⁶Department of Medicine, Human Potential Translational Research Programme & Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, Singapore

⁷Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore

⁸Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

^TThese authors contributed equally

*Correspondence authors: Jianxiao Wu (j.wu@fz-juelich.de); Sarah Genon (s.genon@fz-juelich.de)

Supplementary methods

Scan information

The ABCD imaging data were collected via three types of 3-tesla scanners: GE, Philips, and Siemens with harmonized protocols (1,2). Each structural MRI had an isotropic resolution of 1mm. Each rs-fMRI run had an isotropic resolution of 2.4mm, collected using a multi-band sequence with $TR = 0.8s$. For every participant, two sessions of two 5min rs-fMRI runs were collected.

The HCP-D imaging data were collected using Siemens Prisma 3T MRI scanners from four acquisition sites across the United States (3,4). All structural MRI scans had a 0.8mm isotropic resolution with $TR = 2.5s$. All rs-fMRI scans had an isotropic resolution of 2mm with $TR = 0.8s$. The length of rs-fMRI scans differed regarding age to adapt to the specific challenge of developmental imaging: for participants older than eight years, four 6.5-minute runs were collected; for children between five and seven years old, six 3.5-minute runs were collected.

The HCP-YA imaging data were acquired on a customized Siemens 3T Skyra at Washington University (St Louis) using a multi-band sequence (5,6). The structural images were 0.7mm isotropic. The resting-state fMRI (rs-fMRI) data were 2mm isotropic with $TR = 0.72s$. Two sessions of rs-fMRI data were collected on consecutive days for each participant, and each session consisted of one or two runs. The length of each rs-fMRI scan was 14.4 min (1200 frames).

Preprocessing of rs-fMRI data

ABCD: After basic preprocessing with fMRIPrep, the resting-state functional MRI (rs-fMRI) data were further processed with the following steps: removal of starting frames identified as non-steady state frames, outlier frames detection (i.e., censoring), ICA-AROMA denoising for surface data, and nuisance regression. During outlier frame detection, censored frames included volumes with $FD > 0.3mm$ (7) or DVARS (root-mean-square of voxel-wise differentiated signal) > 50 (8), along with one frame before and two frames after. Remaining segments with fewer than five contiguous frames were also censored. Rs-fMRI runs with at least half of all frames flagged as censored frames were discarded. After censoring, the participants with less than 4-minute rs-fMRI data were excluded. The FD threshold used for ABCD was chosen following Chen et al. (9) to achieve the balance between adequate motion censoring and a large sample size. The DVARS thresholds were selected so that the number of censored frames due to DVARS was roughly the same as the number of frames due to FD. ICA-AROMA denoising was done as part of the preprocessing in fMRIPrep. The denoising components and filters in MNI152 space were applied to the surface data in fsLR space using the same procedure as the denoising for volumetric data. The nuisance regression step removed white matter signals, cerebrospinal fluid (CSF) signal, global signals, and their derivatives, as well as 24 motion parameters.

HCP-D and HCP-YA: The ICA-FIX processed rs-fMRI data further underwent a nuisance regression step, removing white matter signals, cerebrospinal fluid signals, global signals, and their derivatives.

Data split schemes for cross-validation

ABCD: Data from the 21 acquisition sites were first combined into 10 sets to have more balanced sample sizes. Next, every seven sets were iteratively selected as the training set while the remaining three sets were left as the test set (9,10).

HCP-YA: A 10-fold cross-validation was used while preserving family structures (10,11).

HCP-D: Since the data were collected from only four acquisition sites, only a 4-fold cross-validation was possible. This means that, in every iteration, the data from three sites served as the training data, and the remaining site as the test set.

Machine learning methods

For all datasets, we regressed commonly used confounding variables from the behavioral variables before prediction, including age, sex/gender, education (or parental education for ABCD), intracranial volume, FD, and DVARS. To prevent data leakage, the regression coefficients were estimated only from the training set in each cross-validation fold and applied to the test set.

The SVR algorithm was implemented using the Matlab function “fitrlinear”. Default option was used for the hyperparameter determining the error tolerance, epsilon, which was a ratio of the interquartile range of the target variable in the training set.

For comparison, we implemented kernel ridge regression (KRR) with a nonlinear kernel. Similar to previous work, a similarity-based kernel was used where similarity between subjects was computed as Pearson’s correlation between the functional connectivity features of pairs of subjects. The regularization hyperparameter, lambda, was determined via nested cross-validation.

Cramer’s V between two categorical interindividual characteristic variables

Suppose the two variables are A and B with categories of $1, 2, \dots, I$ and $1, 2, \dots, J$. We first need to calculate the chi-squared statistic: $\chi^2 = \sum_{i \in \{1, 2, \dots, I\}, j \in \{1, 2, \dots, J\}} \left[\left(n_{i,j} - \frac{n_{i,*} \times n_{*,j}}{n} \right)^2 / \frac{n_{i,*} \times n_{*,j}}{n} \right]$, where n is the total number of observations, $n_{i,*}$ is the number of observations that $A = i$, $n_{*,j}$ is the number of observations that $B = j$, and $n_{i,j}$ is the number of observations that $A = i$ and $B = j$. Cramer’s V can then be calculated as $V = \sqrt{(\chi^2/n)/\min(I-1, J-1)}$.

Pseudo R-squared and Partial R-squared

The pseudo R-squared (R_{pseudo}^2) of a general linear model (GLM) simulates the coefficients of determination (R^2) in linear regression to measure the goodness of fit. The commonly used form is computed in the form of a likelihood ratio test between the GLM to examine and a null model, where only a constant was fitted. In our R implementation, the alternative expression using deviance was used. Using either likelihood (L) or deviance (D), the pseudo R-squared is calculated as follows:

$$R_{pseudo}^2 = 1 - \frac{\ln(L_{examined\ model})}{\ln(L_{null})} = \frac{D_{null} - D_{examined\ model}}{D_{null}}$$

To estimate variable importance in a GLM in the presence of collinearity among feature variables, we use the partial R-squared (*partial R*²) between the full model and the reduced model. This measures the improvement in variance explained of the full model over the reduced model, as a proportion of the variance not explained in the reduced model. In other words, this metric represents the percentage improvement in variance explained the examined feature brings. The partial R-squared is calculated as follows:

$$partial\ R^2 = \frac{R_{pseudo}^2(full) - R_{pseudo}^2(reduced)}{1 - R_{pseudo}^2(reduced)}$$

Supplementary figures

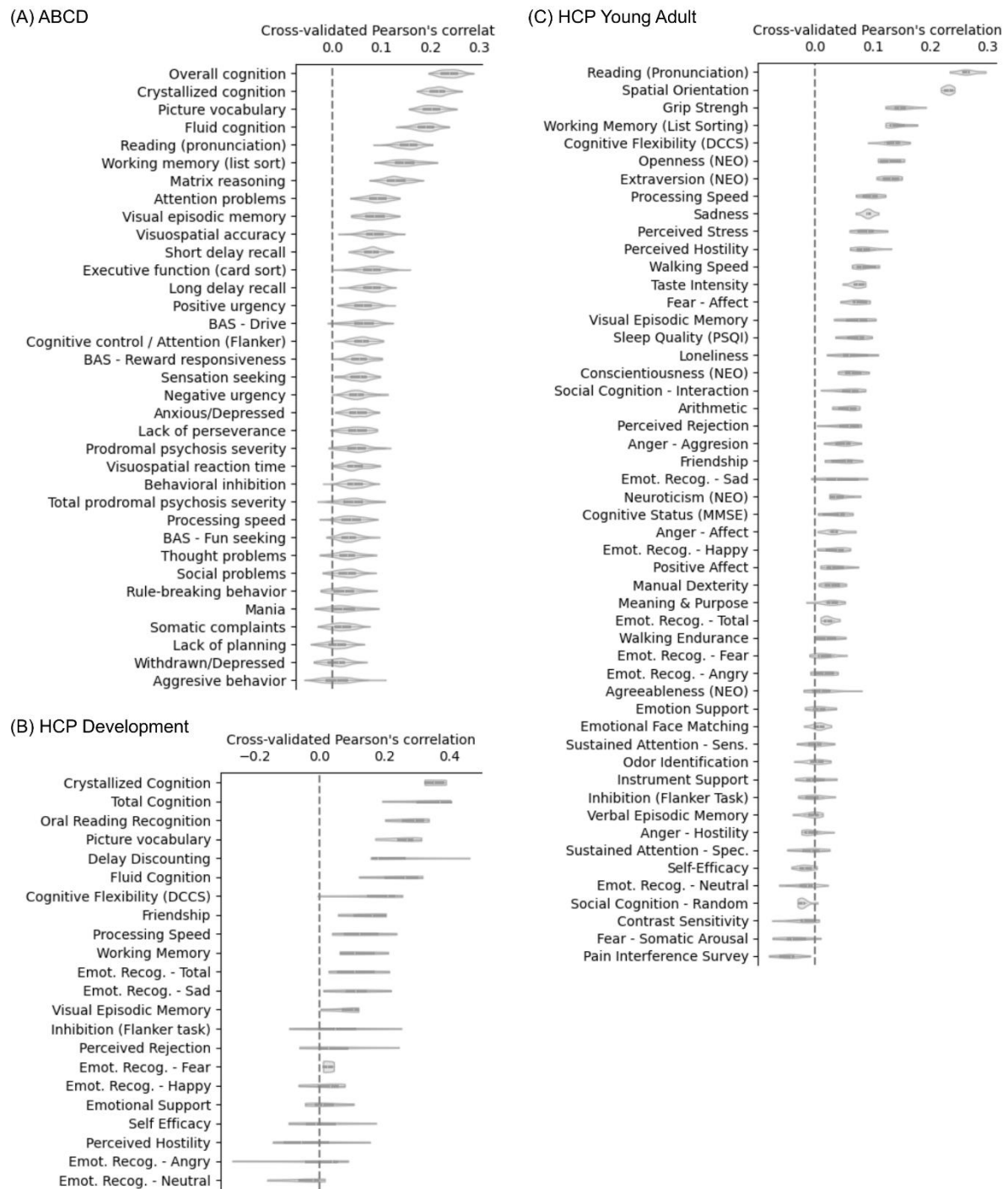
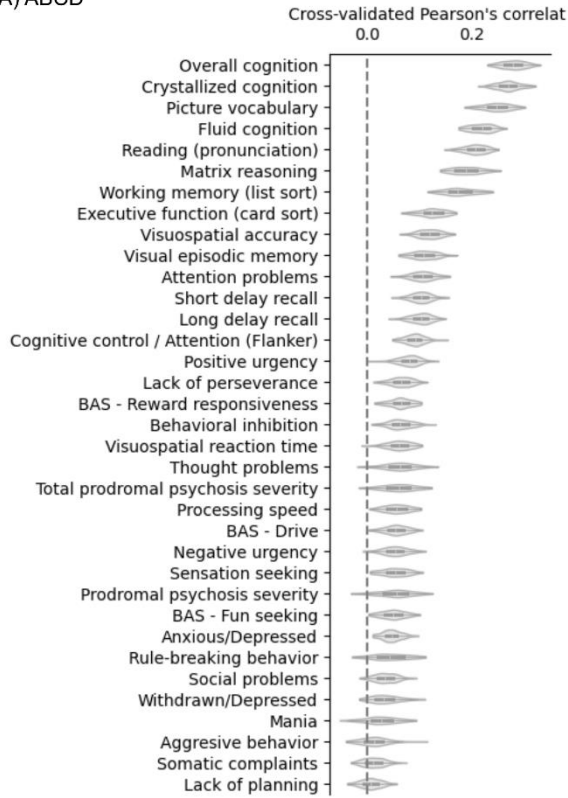
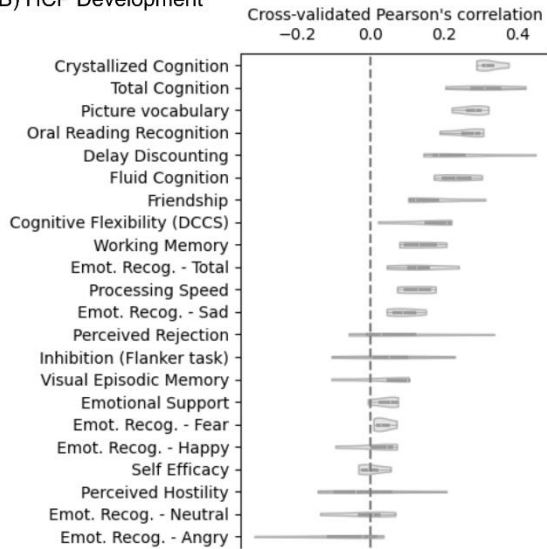


Figure S1. Prediction accuracy of all behavioral measures predicted using support vector regression (SVR) quantified by Pearson's correlation between predicted and original behavioral scores in the (A) ABCD, (B) HCP Development, and (C) HCP Young Adult datasets.

(A) ABCD



(B) HCP Development



(C) HCP Young Adult

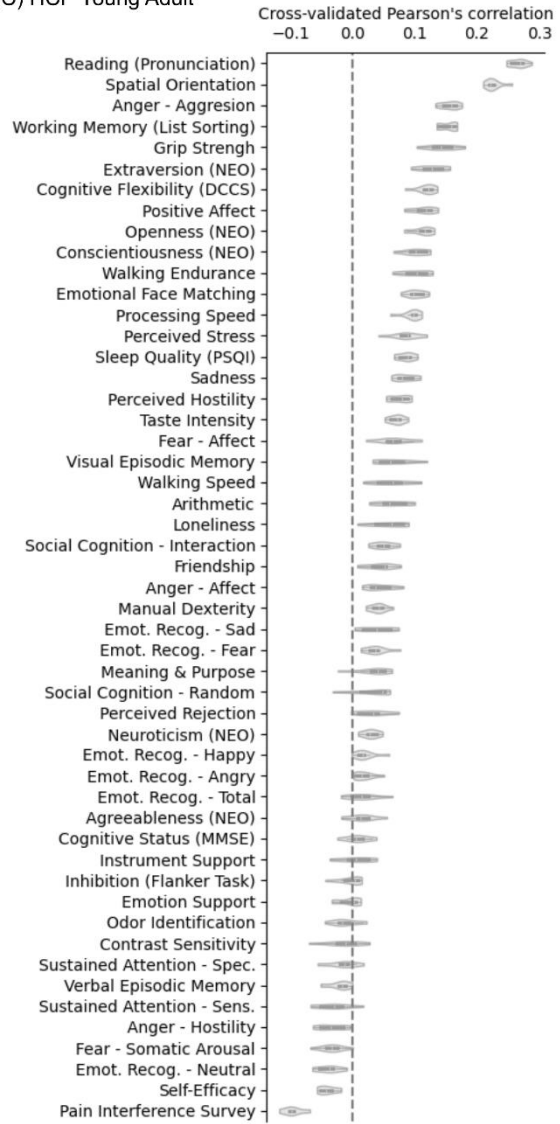
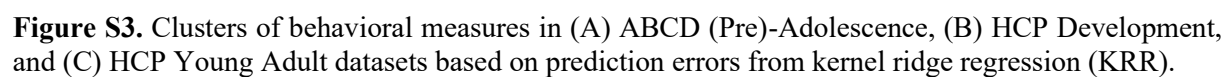
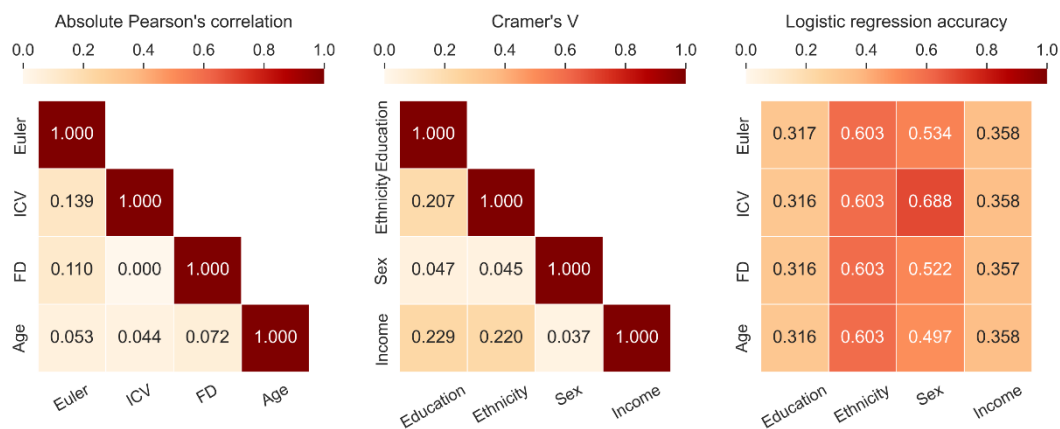


Figure S2. Prediction accuracy of all behavioral measures predicted using kernel ridge regression (KRR) quantified by Pearson's correlation between predicted and original behavioral scores in the (A) ABCD, (B) HCP Development, and (C) HCP Young Adult datasets.

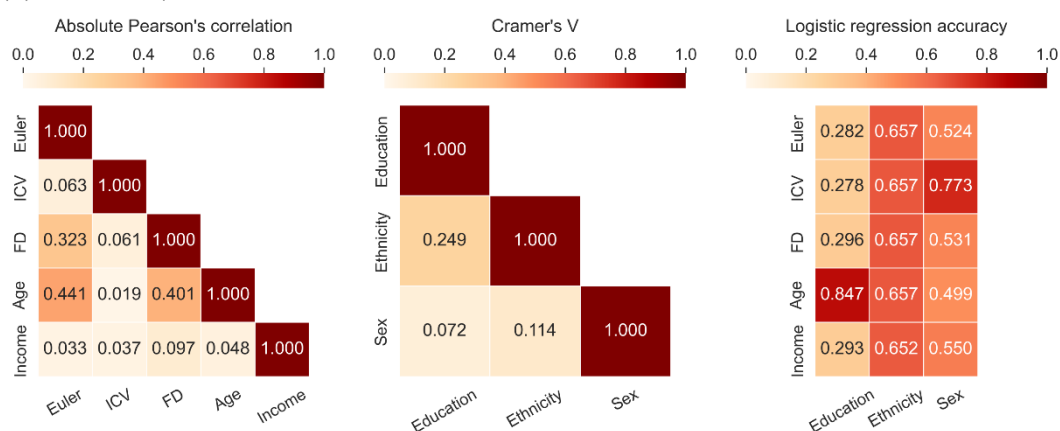
(B) HCP Development



(A) ABCD



(B) HCP Development



(C) HCP Young Adult

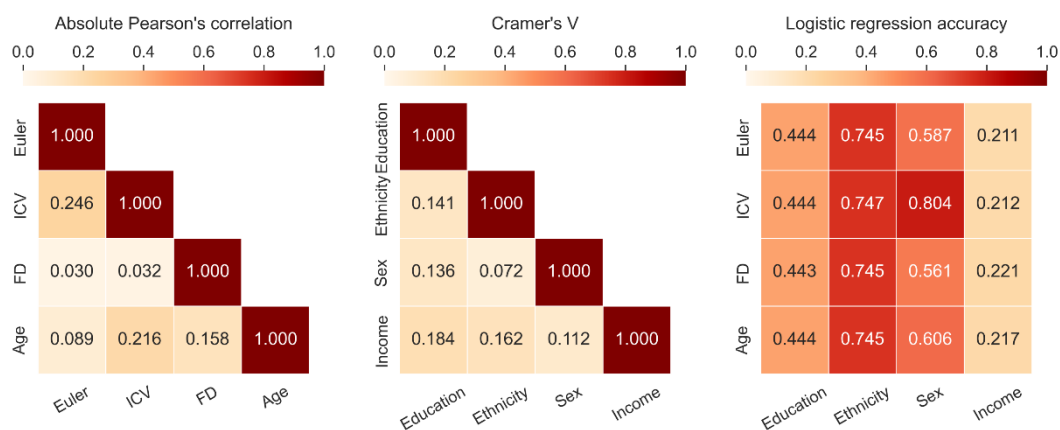
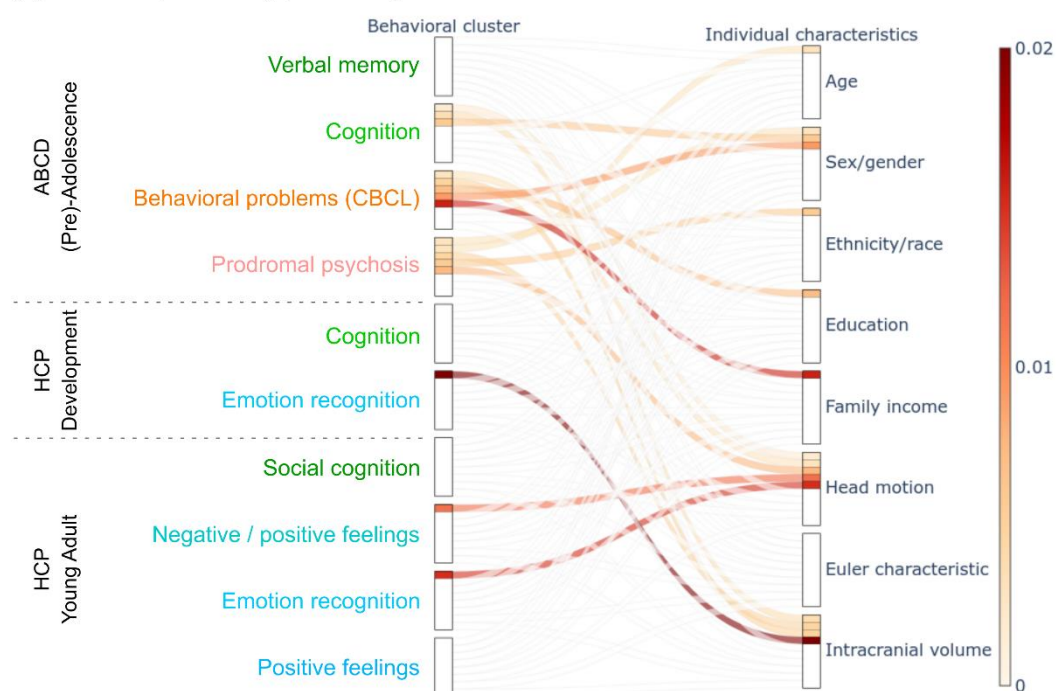


Figure S4. Heatmap showing the relationships between each pair of scan-related and sociodemographic variables. Left: Pearson's correlation between any two continuous variables. Middle: Cramer's V between any two categorical variables. Right: Accuracy when classifying a categorical variable using a continuous variable by logistic regression. Note: the numbers across these three matrices are not comparable even though the color scales are all between 0 and 1.

(A) Variable importance by partial R-squared



(B) Variable importance by BIC

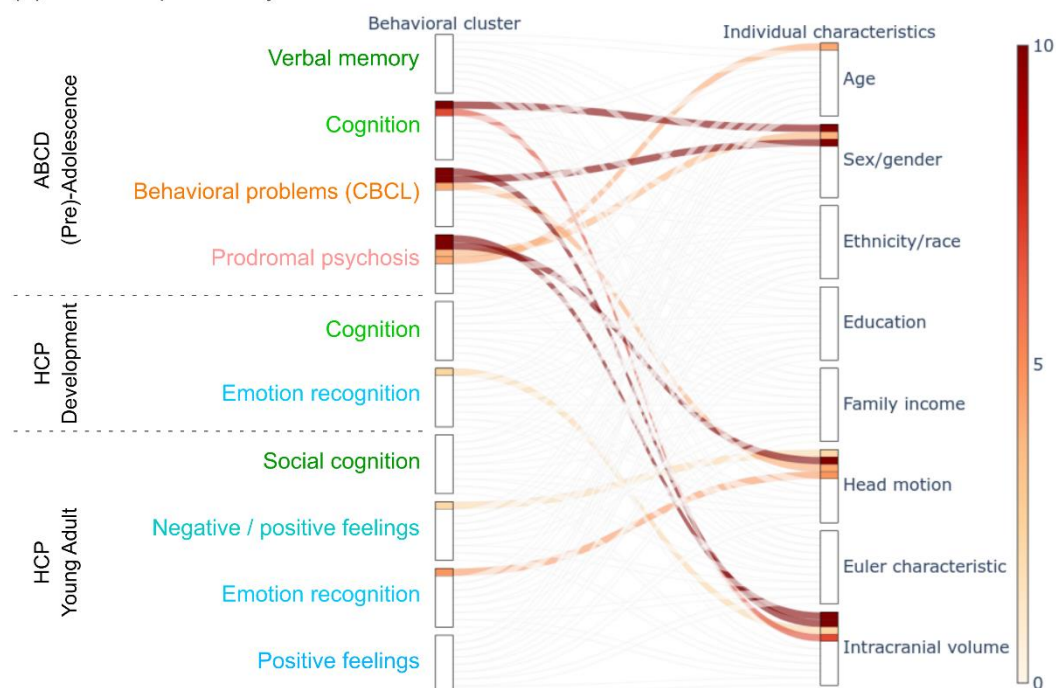


Figure S5. Importance of each individual characteristic in predicting prediction errors of each behavioral cluster defined based on (A) partial R-squared and (B) Bayesian Information Criteria (BIC). Prediction errors were obtained from the KRR prediction results. Colors closer to red indicate higher importance, while white lines represent the lack of importance.

Supplementary tables

Table S1. Summary of ABCD behavioral measures with the descriptive names used in the manuscript.

Scale	Description	ABCD field
Rey Auditory Verbal Learning Test (RAVLT)	Short delay recall	pea_ravlt_sd_trial_vi_tc
	Long delay recall	pea_ravlt_ld_trial_vii_tc
Wechsler Intelligence Scale for Children-V (WISC-V)	Matrix reasoning	pea_wiscv_trs
NIH Toolbox	Cognitive control / Attention (Flanker)	nihtbx_flanker_uncorrected
	Working memory (list sort)	nihtbx_list_uncorrected
	Executive function (card sort)	nihtbx_cardsort_uncorrected
	Reading (pronunciation)	nihtbx_reading_uncorrected
	Processing speed	nihtbx_pattern_uncorrected
	Visual episodic memory	nihtbx_picture_uncorrected
	Picture vocabulary	nihtbx_picvocab_uncorrected
	Fluid cognition	nihtbx_fluidcomp_uncorrected
	Crystallized cognition	nihtbx_cryst_uncorrected
	Overall cognition	nihtbx_totalcomp_uncorrected
Little Man Task	Visuospatial accuracy	lmt_scr_perc_correct
	Visuospatial reaction time	lmt_scr_rt_correct
Achenbach Child Behavior Check List	Anxious/Depressed	cbcl_scr_syn_anxdep_r
	Withdrawn/Depressed	cbcl_scr_syn_withdep_r
	Somatic complaints	cbcl_scr_syn_somatic_r
	Social problems	cbcl_scr_syn_social_r
	Thought problems	cbcl_scr_syn_thought_r
	Attention problems	cbcl_scr_syn_attention_r
	Rule-breaking behavior	cbcl_scr_syn_rulebreak_r
	Aggressive behavior	cbcl_scr_syn_aggressive_r
Parent General Behavior Inventory	Mania	pgbi_p_ss_score
Pediatric Psychosis Questionnaire – Brief Version	Total prodromal psychosis symptoms	pps_y_ss_number
	Prodromal psychosis severity	pps_y_ss_severity_score
	Negative urgency	upps_y_ss_negative_urgency
Modified UPPS-P for Children from PhenX	Positive urgency	upps_y_ss_positive_urgency
	Lack of planning	upps_y_ss_lack_of_planning
	Lack of perseverance	upps_y_ss_lack_of_perseverance
	Sensation seeking	upps_y_ss_sensation_seeking
	Behavioral inhibition	bis_y_ss_bis_sum
Behavioral Inhibition & Activation	BAS - Reward responsiveness	bis_y_ss_bas_rr
	BAS – Drive	bis_y_ss_bas_drive
	BAS - Fun seeking	bis_y_ss_bas_fs

Table S2. Lookup table showing the original HCP-D behavioral variable names with the corresponding descriptive labels used in the manuscript.

Description	HCP-D CSV filename: field
Fluid Cognition	cogcomp01: nih_fluidcogcomp_ageadjusted
Crystallized Cognition	cogcomp01: nih_crycogcomp_ageadjusted
Total Cognition	cogcomp01: nih_totalcogcomp_ageadjusted
Cognitive Flexibility (DCCS)	dccs01: nih_dccs_ageadjusted
Delay Discounting	deldisk01: auc_40000
Emot. Recog. - Total	er4001: er40_c_cr
Emot. Recog. - Angry	er4001: er40_c_ang
Emot. Recog. - Fear	er4001: er40_c_fear
Emot. Recog. - Happy	er4001: er40_c_hap
Emot. Recog. - Neutral	er4001: er40_c_noe
Emot. Recog. - Sad	er4001: er40_c_sad
Inhibition (Flanker task)	flanker01: nih_flanker_ageadjusted
Working Memory	lswmt01: age_corrected_standard_score

Oral Reading Recognition	orrt01: read_acss
Processing Speed	pcps01: nih_patterncomp_ageadjusted
Visual Episodic Memory	psm01: nih_picseq_ageadjusted
Self Efficacy	self_effic01: nih_tlbx_rawscore
Emotional Support	tlbx_emsup01: nih_tlbx_rawscore
Friendship	tlbx_friend01: nih_tlbx_rawscore
Perceived Hostility	tlbx_perhost01: nih_tlbx_rawscore
Perceived Rejection	tlbx_rej01: nih_tlbx_tscore
Picture vocabulary	tpvt01: tpvt_acss

Table S3. Lookup table showing the original HCP-YA behavioral variable names with the corresponding descriptive labels used in the manuscript.

Description	HCP-YA field
Visual Episodic Memory	PicSeq_Unadj
Cognitive Flexibility (DCCS)	CardSort_Unadj
Inhibition (Flanker Task)	Flanker_Unadj
Reading (Pronunciation)	ReadEng_Unadj
Processing Speed	ProcSpeed_Unadj
Spatial Orientation	VSLOT_TC
Sustained Attention – Sens.	SCPT_SEN
Sustained Attention – Spec.	SCPT_SPEC
Verbal Episodic Memory	IWRD_TOT
Working Memory (List Sorting)	ListSort_Unadj
Cognitive Status (MMSE)	MMSE_Score
Sleep Quality (PSQI)	PSQI_Score
Walking Endurance	Endurance_Unadj
Walking Speed	GaitSpeed_Unadj
Manual Dexterity	Dexterity_Unadj
Grip Strength	Strength_Unadj
Odor Identification	Odor_Unadj
Pain Interference Survey	PainInterf_Tscore
Taste Intensity	Taste_Unadj
Contrast Sensitivity	Mars_Final
Emotional Face Matching	Emotion_Task_Face_Acc
Arithmetic	Language_Task_Math_Avg_Difficulty_Level
Social Cognition – Random	Social_Task_Perc_Random
Social Cognition – Interaction	Social_Task_Perc_TOM
Agreeableness (NEO)	NEOFAC_A
Openness (NEO)	NEOFAC_O
Conscientiousness (NEO)	NEOFAC_C
Neuroticism (NEO)	NEOFAC_N
Extraversion (NEO)	NEOFAC_E
Emot. Recog. – Total	ER40_CR
Emot. Recog. – Angry	ER40ANG
Emot. Recog. – Fear	ER40FEAR
Emot. Recog. – Happy	ER40HAP
Emot. Recog. - Neutral	ER40NOE
Emot. Recog. – Sad	ER40SAD
Anger – Affect	AngAffect_Unadj
Anger – Hostility	AngHostil_Unadj
Anger – Aggression	AngAggr_Unadj
Fear – Affect	FearAffect_Unadj
Fear – Somatic Arousal	FearSomat_Unadj
Sadness	Sadness_Unadj
Meaning & Purpose	MeanPurp_Unadj
Positive Affect	PosAffect_Unadj
Friendship	Friendship_Unadj
Loneliness	Loneliness_Unadj
Perceived Hostility	PercHostil_Unadj
Perceived Rejection	PercReject_Unadj

Emotional Support	EmotSupp_Unadj
Instrument Support	InstruSupp_Unadj
Perceived Stress	PercStress_Unadj
Self-Efficacy	SelfEff_Unadj

Table S4. Association between the important individual characteristics (sex/gender, head motion, and intracranial volume) and prediction errors of behavioral clusters from the SVR models. For sex/gender, associations were assessed by two-sample t-test and effect sizes were represented by the Cohen's d. For head motion and intracranial volume, associations were assessed by both Pearson's correlation and Spearman's correlation.

Cohort	Behavioral cluster	Sex/gender	Head motion	Intracranial volume
ABCD (Pre)-Adolescence	Cognition	errors higher in male Cohen's d = 0.097	-	Pearson's R = -0.034 Spearman's R = (n.s.)
	Behavioral problems (CBCL)	errors higher in male Cohen's d = 0.173	Pearson's R = 0.094 Spearman's R = 0.101	Pearson's R = -0.071 Spearman's R = -0.068
	Prodromal psychosis	errors higher in male Cohen's d = 0.081	Pearson's R = 0.099 Spearman's R = 0.102	-
HCP Development	Emotion recognition	-	-	Pearson's R = -0.112 Spearman's R = (n.s.)
HCP Young Adult	Emotion recognition	-	Pearson's R = 0.110 Spearman's R = 0.081	-

1. Karcher NR, Barch DM. The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*. 2021 Jan;46(1):131–42.
2. Casey BJ, Cannonier T, Conley MI, Cohen AO, Barch DM, Heitzeg MM, et al. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cogn Neurosci*. 2018 Aug;32:43–54.
3. Somerville LH, Bookheimer SY, Buckner RL, Burgess GC, Curtiss SW, Dapretto M, et al. The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5-21 year olds. *Neuroimage*. 2018 Dec;183:456–68.
4. Harms MP, Somerville LH, Ances BM, Andersson J, Barch DM, Bastiani M, et al. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *Neuroimage*. 2018 Dec;183:972–84.
5. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K, et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage*. 2013 Oct 15;80:62–79.
6. Smith SM, Beckmann CF, Andersson J, Auerbach EJ, Bijsterbosch J, Douaud G, et al. Resting-state fMRI in the Human Connectome Project. *Neuroimage*. 2013 Oct 15;80:144–68.
7. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. 2002 Oct;17(2):825–41.

8. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*. 2012 Feb 1;59(3):2142–54.
9. Chen J, Tam A, Kebets V, Orban C, Ooi LQR, Asplund CL, et al. Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nat Commun*. 2022 Apr 25;13(1):2217.
10. Li J, Bzdok D, Chen J, Tam A, Ooi LQR, Holmes AJ, et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv*. 2022 Mar 18;8(11):eabj1812.
11. Li J, Kong R, Liégeois R, Orban C, Tan Y, Sun N, et al. Global signal regression strengthens association between resting-state functional connectivity and behavior. *Neuroimage*. 2019 Aug 1;196:126–41.