**Title:** Patterns of socio-cognitive stratification and perinatal risk in the child brain.

**Authors:** Dag Alnæs[a*], Tobias Kaufmann[a], Andre F. Marquand[b,c,d], Stephen M. Smith[e], Lars T. Westlye[a,f]

[a] Norwegian Centre for Mental Disorders Research (NORMENT), Institute of Clinical Medicine, University of Oslo & Division of Mental Health and Addiction, Oslo University Hospital, Norway
[b] Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[c] Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, the Netherlands
[d] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, UK
[e] Oxford University Centre for Functional MRI of the Brain (FMRIB), Oxford, UK
[f] Department of Psychology, University of Oslo, Norway

***Correspondence:** Dag Alnæs*, email:* dag.alnas@psykologi.uio.no*, Oslo University Hospital, PO Box 4956 Nydalen, 0424 Oslo, Norway, Phone: +47 23 02 73 50, Fax: +47 23 02 73 33*

**Disclosures:** The authors declare no conflict of interest

**Acknowledgements:** D.A. is funded by the South- Eastern Norway Regional Health Authority (2019107). T.K. is funded by the Research Council of Norway (276082). A.F.M. gratefully acknowledges support from the Dutch Organisation for Scientific Research via a Vernieuwingsimpuls VIDI fellowship (016.156.415) and a Wellcome Trust Innovator award (215698/Z/19/Z). L.T.W. is funded by the European Research Council under the European Union's Horizon 2020 research and innovation program (ERC Starting Grant 802998), the Research Council of Norway (249795), the South-East Norway Regional Health Authority (2019101), and the Department of Psychology, University of Oslo.

**Author contributions**: D.A. and L.T.W. conceived of the study; D.A. performed the analysis with contributions from L.T.W., A. F. M., S.M. and T.K.; D.A. wrote the first draft with contributions from L.T.W and T.K., and all authors contributed to the final manuscript.

**Abstract**

The expanding behavioral repertoire of the developing brain during childhood and adolescence is shaped by complex brain-environment interactions and flavored by unique life experiences. Our multivariate brain-behavior mapping in 8,622 children aged 9-11 years across 585 brain imaging phenotypes and 601 cognitive, behavioral, psychosocial and socioeconomic measures revealed robust brain patterns reflecting socio-cognitive stratification, and traces of perinatal complications in brain morphology and white matter.

## Introduction

The complexity and idiosyncratic characteristics of the human mind originates in an intricate web of interactions between genes, brain circuits, behaviors, economic, social and cultural factors during childhood and adolescence. The major life changes associated with the transition into young adulthood offer opportunities for adaptation and growth, but also increased susceptibility to detrimental perturbations, such as the characteristics of social and parental relationships, family environment, quality of schools and activities, economic security, urbanization and pollution, drugs, cultural practices, and values, that all act in concert with our genetic architecture and biology. A multidimensional understanding of the interplay of these factors is paramount to identify the constituents of healthy development and to identify factors that may precede maladjustment and mental illness.

Population-based neuroimaging now allows us to take a birds-eye view on this stupendous multiplicity, and to bring hitherto unseen patterns into focus[1]. The Adolescent Brain Cognitive Development (ABCD) study[2] provides brain images of more than 10,000 children aged 9-11 years across the US and includes a broad range of cognitive, behavioral, clinical, psychosocial and socioeconomic measures. Analogous to the accumulation of small effects in the genetics of complex human traits and polygenic disorders, while each neuroimaging feature typically explains a minute amount of unique variance in behavioural outcome[3, 4], their combined predictive value is non-negligible, including predictive patterns for individuals[5, 6] and characteristics such as age[7, 8], cognitive ability[9] and psychopathology[9]. Combining levels of information using latent-variable approaches which model all available information may reveal interpretable patterns among multiple imaging and variables such as cognition and socio-demographics[4, 10]. One recent example revealed that a wide range of cognitive, clinical and lifestyle measures constitute a "positive-negative" dimension associated with adult brain network connectivity[1].

Here we used an analogous approach in 8,622 children aged 9-11 from the ABCD-study, combining canonical correlation analysis (CCA) with independent component analysis (ICA) to derive population-level modes of co-variation, linking behavioral, psychosocial, socioeconomic and demographical variables (behavioral measures) to a wide set of neuroimaging phenotypes. Each resulting mode represents an association between a linear combination of behavioral measures with a separate combination of imaging features that show similar variation across participants[10]. CCA was performed after data reduction with principal component analysis, and robustness and reliability of the identified modes were assessed using cross-validation and

permutation testing to avoid overfitting. To express results in the original variable-space, CCA-ICA subject weights where correlated back into the original data.

## Results

We identified four distinct modes of co-variation, linking brain-features with cognition, socioeconomic status (SES) and lifestyle factors, perinatal events and cognitive, motor and speech development, and urbanicity and onset of puberty **(Fig. 1)**. Canonical correlations (*r=.68, r=.50, r=.49, r=44*) where were highly significant and robust as assessed by 10-fold cross-validation and permutation (out-of-sample *r=.63, r=.40, r=.41, r=.32*, all *permuted-p < 0.01* respectively, Supplementary Fig. 1).

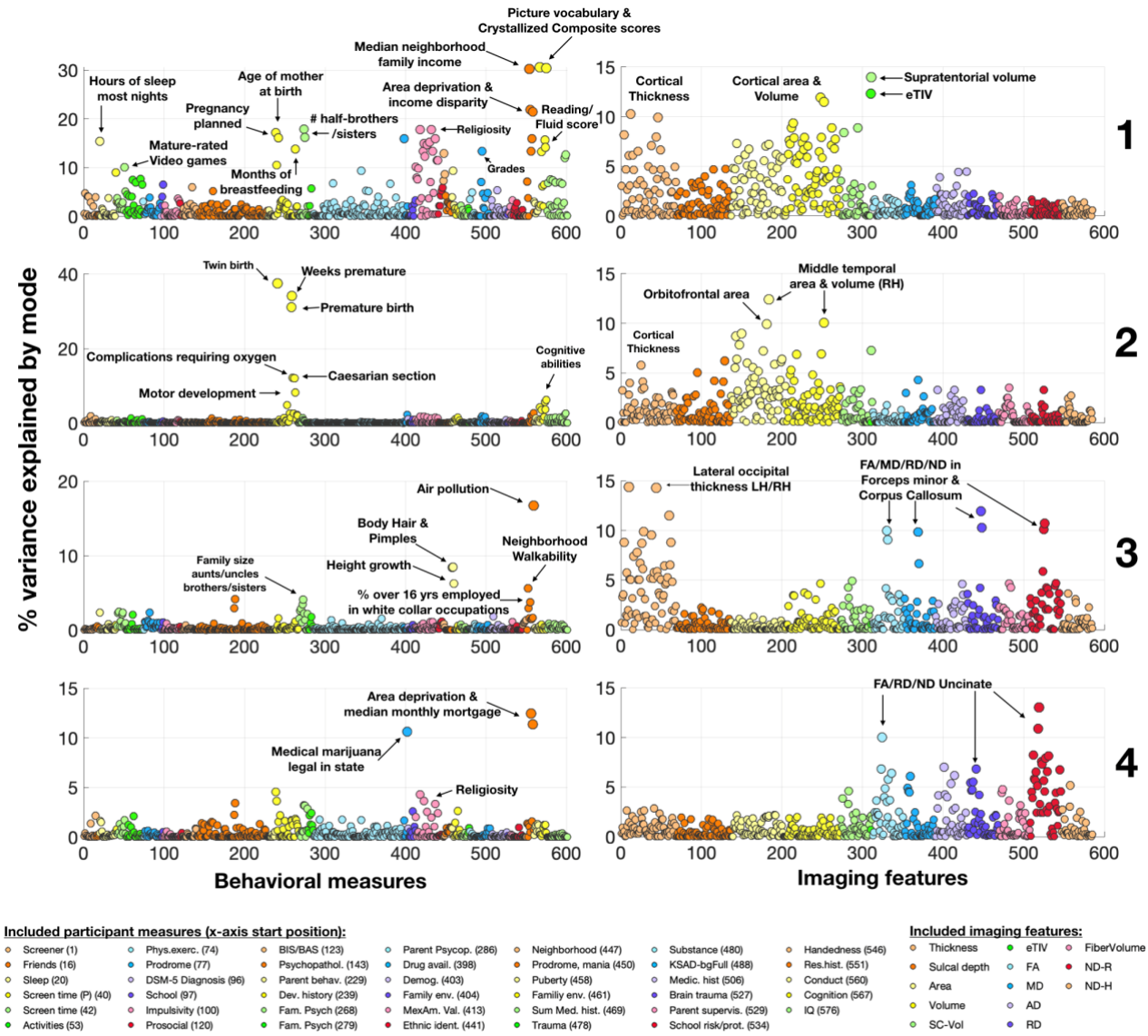



***Fig. 1:*** Each row represents a CCA-ICA mode. Each mode represents an association between a linear combination of behavioral measures (left) with a separate combination of imaging features (right). Numbers on x-axes represent

the numbered behavioral measures / imaging features. Behavioral measures legend shows starting location on x-axis for each measure (in parenthesis). Y-axes shows the explained variance in original adjusted data using the subject weights from the ICA performed on the CCA scores.

Mode 1 links cognitive ability and grades, to area deprivation index (ADI), reflecting median household income, income disparity, home ownership percentage, and neighborhood violence

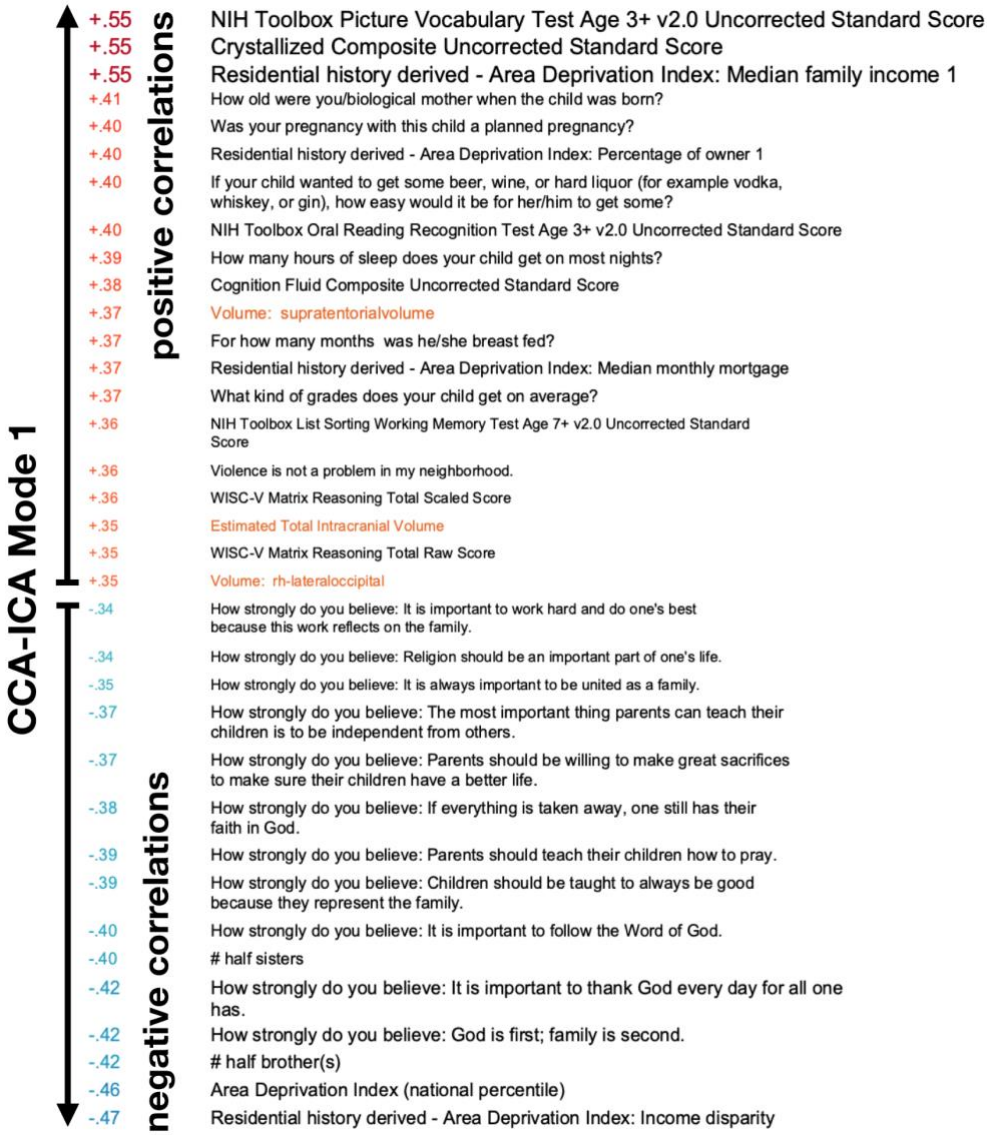
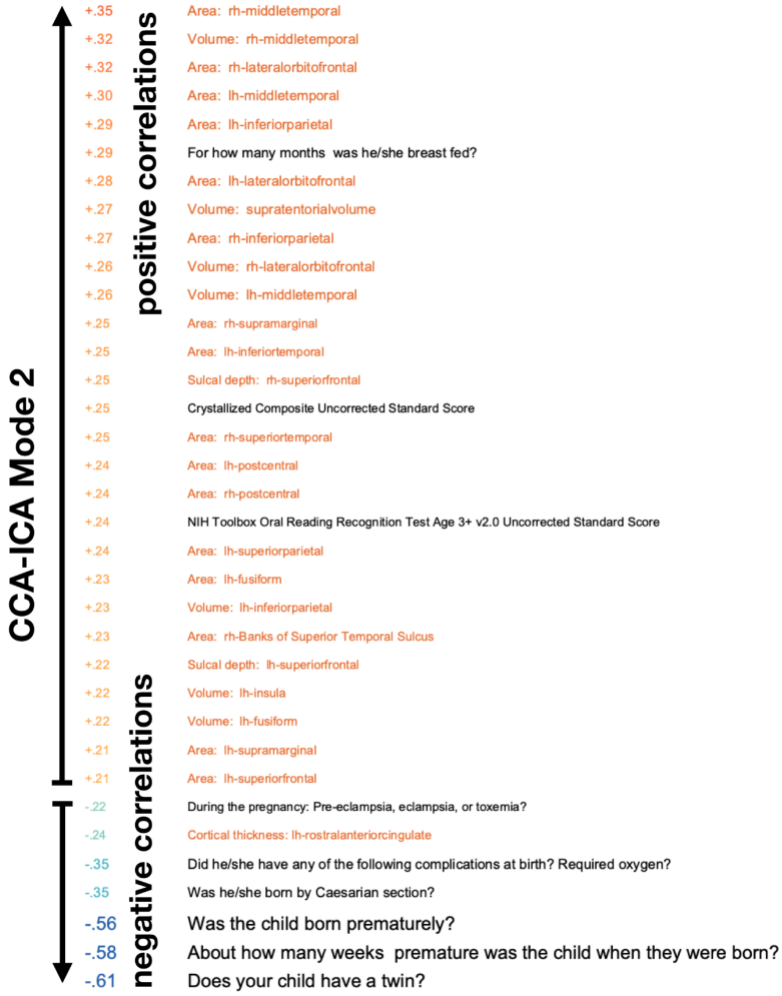



**Fig. 2:** A dimensional axis of socio-cognitive stratification and brain structure in children. For each original variable/questionnaire item their correlations with CCA-ICA Mode 1 subject weights are shown, forming a dimension of socio-cognitive stratification which is associated with total intracranial volume and regional cortical volumes and thickness. Numbers on the left are correlations, hot/cold colors represent positive/negative correlations, respectively, arrows indicate increasing positive/negative correlations. Text on the right represents the behavioral measures (black) and imaging features (orange). The top 35 items are shown, for a full list of all measures and associated correlations see Supplementary Table 1.

and crime **(Fig. 2)**. The mode further links maternal age at birth and whether the pregnancy was planned, number of half-siblings, sleep duration, and religiosity, jointly forming a dimension of socio-cognitive stratification. This dimension is robustly associated with cortical thickness, sulcal depth, area and volume, with brain volume, lateral occipital and middle temporal cortical volumes and bilateral lingual thickness among the highest-loading imaging features.

Mode 2 links perinatal factors and obstetric complications to cognitive ability and brain morphology in late childhood **(Fig. 3)**. Having a twin, premature birth, birth complications requiring oxygen, Caesarian section, (pre-)eclampsia, toxemia and jaundice, as well as duration of breast feeding, are linked to cognition, parent reported motor and speech development, and cortical morphometry and white matter diffusion measures in several brain regions, with middle temporal, lateral orbitofrontal and inferior parietal volumes and area as the highest-loading imaging features.

**Fig. 3:** Mode 2 links obstetric and perinatal complications with cognitive ability and cortical area, volume and thickness. Numbers on the left are correlations between each participant measure and MRI feature with Mode 2 CCA-ICA subject weights. Hot/cold colors represent positive/negative correlations, respectively. Text on the right represent the participant measures (black) and imaging features (orange). The top 35 items are shown, for a full list of all measures and associated correlations see Supplementary Table 2




Mode 3 links air pollution ($PM_{2.5}$), neighborhood walkability and being overweight with emerging signs of puberty such as body hair, growth-spurt and pimples, to cortical thickness and diffusion indices, with thickness in lateral occipital, middle frontal and superior parietal regions, and fractional anisotropy (FA), radial diffusivity (RD), mean diffusivity (MD) and neurite density (ND) of the forceps minor and the corpus callosum among the highest loading imaging features (Supplementary Fig. 2A, Supplementary Table 3).

Mode 4 links median monthly mortgage, national percentile of ADI and living in a state with legalized medical marihuana, with white matter diffusion indices, with bilateral ND and FA of the left uncinate fasciculus among the highest loading imaging features (Supplementary Fig. 2B, Supplementary Table 4).

## Discussion

We found four robust modes of child population co-variation, each linking behavioral, clinical, psychosocial, socioeconomic and demographical measures to neuroimaging in more than 8,600 children aged 9-11 years. The first mode captured a pattern of socio-cognitive stratification which was associated with brain volume and regional measures of cortical thickness, area and volume. Conceptually the mode shares similarities with a positive-negative population mode linked to brain connectivity[1] and structure[11] in adults. Similar to adults, the child mode captures a positive manifold among several cognitive measures. Further, the mode is linked to several positive and negative life-events, constituting important factors for neurodevelopment that these children receive from their parents, both in terms of genetic and epigenetic inheritance of cognitive ability, but also of living in high- or low-income neighborhoods, the neighborhood level of violence, parents' level of religious beliefs, which is known to be associated with social deprivation, family-planning and later pregnancies, which is associated with higher education and SES. The neurotypical developmental trajectory at this age is characterized by apparent cortical thinning, likely partly reflecting synaptic pruning[12] and myelination[13]. Thicker cortex with higher SES is consistent with reports of accelerated brain maturation in children from low-SES families[14, 15]. Indeed, across species and in humans, early life adversity is associated with accelerated maturation of neural systems, possibly at a cost of increased risk for later mental health problems[16].

The second mode links obstetric complications and early life factors such as duration of breast feeding and motor development, with cognitive ability, cortical surface area, thickness and volume in late childhood. Obstetric complications increase the risk of later cognitive deficits and mental disorders[17]. The present results support that children with a history of obstetric and perinatal complications show delayed brain development and are consistent with reports associating birth weight with cortical area and brain volume in childhood and adolescence[18].

The third mode is related to urbanicity, with an inverse relationship between air pollution and walkability. Low-pollution, high-walkability "sweet-spot" neighborhoods are skewed toward high-

SES households, contrasted with high-pollution, low-walkability "sour-spot" neighborhoods associated with mid-level income[19]. Associations between neighborhood walkability and overweight[20], and between overweight and onset of puberty[21] are consistent with previous reports, and here we document an association with cortical thickness, and FA, ND, RD and MD in inter-hemispheric connections, including corpus callosum and forceps minor.

The fourth mode also reflects socioeconomic factors, capturing an inverse relationship between neighborhood median monthly mortgage, and neighborhood area deprivation, as well as living in a state with legalized marijuana. This pattern is likely not related to cannabis use, rather it may capture individual factors related to residing in liberal versus conservative areas and possibly other tendencies or parental factors determining demographical patterns. The link to white matter indices is interesting, yet, as with the other modes, the cross-sectional and the non- experimental design warrant caution. People and their brains, genes and environments are not varying randomly, but are highly correlated[22], likely along multiple dimensions, which complicates causal and mechanistic inference. This is especially relevant for population-based neuroimaging, in which subtle confounds can induce spurious associations[4]. These general caveats notwithstanding, these valuable resources represent an unprecedented opportunity to reveal co-varying patterns of socio-demographics, cognitive abilities, mental health, and brain imaging data, beyond simple bivariate associations, which are potentially highly informative of the biology and sociology of brain development and psychological adaptation.

In contrast to the standard regression approach which models one outcome-variable at a time and typically includes only a few covariates the combined multivariate approach employed here considers the full pattern of co-variability between variables. The current approach is therefore well suited for capturing population patterns by maximizing statistical power. However, it does not allow for interpretation of specific associations between pairs of variables. All the patterns are purely correlational and also treated analytically and reported as such. It is also entirely possible, and highly probable, that these patterns are further correlated with other important phenomena not measured or included in the current analysis. The current approach also effectively captures differential patterns involving the same measures. For example, higher cortical thickness, indicative of delayed maturation, is independently associated both with socio-cognitive stratification, higher cognitive ability and SES, as well as with obstetric and perinatal complications, lower cognitive ability, and delayed speech and motor development. The independent and co-existing associations with brain structure emphasize the importance of

multidimensional considerations for understanding child and adolescent neurodevelopment and support that political decision aiming to improve health outcomes and adaptation during transformative life phases should be based on interdisciplinary perspectives integrating social, psychological and biological sciences.

## References


1. Smith, S.M., *et al. Nature neuroscience* **18**, 1565 (2015).
2. Casey, B.J., *et al. Developmental Cognitive Neuroscience* **32**, 43-54 (2018).
3. Paulus, M.P. & Thompson, W.K. *JAMA Psychiatry* **76**, 353-354 (2019).
4. Smith, S.M. & Nichols, T.E. *Neuron* **97**, 263-268 (2018).
5. Finn, E.S., *et al. Nature neuroscience* **18**, 1664 (2015).
6. Kaufmann, T., *et al. Nature neuroscience* **20**, 513 (2017).
7. Kaufmann, T., *et al. Nature neuroscience* (2019).
8. Franke, K. & Gaser, C. *Frontiers in Neurology* **10**, 789 (2019).
9. Alnæs, D., *et al. JAMA Psychiatry* (2018).
10. Miller, K.L., *et al. Nature neuroscience* **19**, 1523 (2016).
11. Llera, A., Wolfers, T., Mulders, P. & Beckmann, C.F. *Elife* **8**, e44443 (2019).
12. Gilmore, J.H., Knickmeyer, R.C. & Gao, W. *Nature Reviews Neuroscience* **19**, 123 (2018).
13. Natu, V.S., *et al. Proceedings of the National Academy of Sciences* **116**, 20750 (2019).
14. LeWinn, K.Z., Sheridan, M.A., Keyes, K.M., Hamilton, A. & McLaughlin, K.A. *Nature Communications* **8**, 874 (2017).
15. Piccolo, L.R., *et al. PloS one* **11**, e0162511 (2016).
16. Callaghan, B.L. & Tottenham, N. *Curr Opin Behav Sci* **7**, 76-81 (2016).
17. Mezquida, G., *et al. The Journal of Nervous and Mental Disease* **206** (2018).
18. Walhovd, K.B., *et al. Proceedings of the National Academy of Sciences* **109**, 20089 (2012).
19. Marshall, J.D., Brauer, M. & Frank, L.D. *Environ Health Perspect* **117**, 1752-1759 (2009).
20. Stowe, E.W., Hughey, S.M., Hallum, S.H. & Kaczynski, A.T. *Childhood Obesity* (2019).
21. Reinehr, T. & Roth, C.L. *The Lancet Child & Adolescent Health* **3**, 44-54 (2019).
22. Turkheimer, E. in *Philosophy of behavioral biology.* 43-64 (Springer Science, New York, NY, US, 2012).

## Online Methods

*ABCD data access:* We accessed MRI, behavioral and clinical data from ABCD Annual curated release 2.0.1. The data as well as release notes including documentation of measures, scanning protocols and imaging QC can be accessed using the following NIMH data archive DOI: http://dx.doi.org/10.15154/1503209.

*Behavioral, clinical, cognitive and demographical data:* Tabulated data was imported and processed using R (https://cran.r-project.org). We accessed data from 11,853 participants. Supplementary Table 5 lists the behavioral measures included the analysis. We used the function 'nearZeroVar' from the R-package 'caret' (v. 6.0-81, https://github.com/topepo/caret/) to identify and exclude any continuous variables with zero or near-zero variance, and categorical variables with a ratio of > .95 for the most common compared to the second most common response. For each remaining variable we derived robust z-scores by calculating each scores absolute deviation from the median absolute deviation[23] (MAD), and removed values with a z > 3 (3 x MAD). We then excluded variables with less than 90% of retained datapoints, before excluding subjects with less than 90% retained data across the retained variables. The remaining subjects (n=11,816) were included for further analysis.

*MRI Imaging derived phenotypes:* We accessed T1 and T2 (n=11,534) and DWI (n=11,400) tabulated data from ABCD curated release 2.0.1. Supplementary Table 6 lists the MRI features included in the analysis. We included participants which passed quality assurance using the recommended QC parameters (T1: n=11,359, T2: n=10,476, DWI: n=10,414) described in the ABCD 2.0.1 Imaging Instruments Release Notes and whom had all included modalities available (n=9,811). ABCD preprocessing and QC steps are described in detail in the methodological reference for the ABCD Study by Hagler et al[24]. For each included imaging phenotype, we calculated the median absolute deviation (MAD) for each score, and removed values with MAD > 3. Subjects with less than 90% of features retained in any of the imaging modalities, and features with less than 90% of retained subjects were excluded from analysis. The remaining subjects (n=9,016) were included for further analysis.

*Canonical correlation analysis:* We performed CCA[25] using MATLAB R2019b. Participants with both MRI and behavioral variables (n=8,622) were included. We applied a rank-based normal transformation to the behavioural/clinical data using 'palm_inormal' from FSLs PALM[26] (v. 0.52, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM). Next, we residualized all measures with respect to

age and sex using linear models. Imaging phenotypes were also residualized for site/scanner, and volumetric features were also corrected for estimated total intracranial volume (eTIV from Freesurfer). eTIV was also included as variable in the analysis to capture associations with global volume, in addition to the eTIV-corrected volumes capturing associations with regional specificity. For both MRI and behavioral measures, missing values were imputed with 'knnimpute', replacing missing data based on the k nearest-neighbour columns based on Euclidian distance (k=3), before being z-normalized and submitted (separately for imaging and behavioural data) to PCA (Supplementary Fig. 3), to avoid issues with rank deficiency and to increase robustness of estimated modes by avoiding fitting to noise. We extracted the first 200 components for both the imaging and behavioural data, and submitted these to CCA.

*Cross-validation:* To assess the reliability and generalizability of the resulting CCA-modes we performed the following 10-fold cross-validation procedure: For each iteration (n=100) of the cross-validation loop the dataset was randomly divided into 10 folds. While keeping each fold (10% of participants) out once we submitted the remaining data (90% of participants) to PCA (separately for imaging and MRI data) and then to CCA. Next we multiplied the kept-out behavioral measure and MRI feature matrices with the estimated PCA coefficient matrices, before multiplying the resulting PCA scores with the canonical coefficients and then correlated the resulting CCA scores. Finally, we took the average of these canonical correlations across the 10 folds (Supplementary Fig. 1). This procedure was repeated 100 times to derive mean canonical correlations for kept-out data, and used in the next step for calculating p-values. We also correlated the CCA subject measure and MRI coefficients derived for kept-out participants, with those from the full analysis.

*Permutation testing*: To assess significance of the resulting CCA-modes, we ran 1000 iterations of the same 10-fold cross-validation procedure described above, but with the order of participants of the imaging phenotype matrix randomly permuted in each iteration, and collecting canonical correlations for the kept rather than the kept-out data to account for overfitting by the CCA. We then collected the maximum canonical correlation across CCA-modes (i.e. mode 1) for each permutation to form a null-distribution to calculate familywise error corrected (FWE) p-values. P-values for each of the CCA-modes were calculated by dividing the count of permuted maximum R-values (including the observed value) >= the mean cross-validated R-value by the number of permutations. CCA-modes with a corrected p-value < .01 was included for further analysis (Supplementary Fig. 1).

*CCA-ICA:* We extracted the CCA-scores for the four significant modes and submitted them to independent component analysis (ICA), following procedures described by Miller et al[10], using fastICA[27] estimating four components (the number of significant modes). To assess the reliability and generalizability of the ICA decomposition we reran 100 iterations of the 10-fold cross-validation procedure described above, this time including ICA estimation after the CCA step, and then correlated subject-weights derived from kept-out data to those from the full analysis (Supplementary Fig. 1). To assess and plot the significant CCA-ICA modes in the original variable space, we correlated the subject weights for each CCA-ICA mode with the original age- and sex (+ eTIV) adjusted matrices and expressed these association in terms of explained variance by squaring these correlations. For each significant mode of population covariation, we also plotted the variable text/descriptions for the 35 variables with the highest explained variance in the original adjusted data (List of all variables and associated descriptions, correlations and ICA weights can be found in Supplementary Tables 1-4). ICA subject weight histograms are shown in Supplementary Fig. 4. The explained variance of single variables ranged between 12 and 40% for the most highly involved items on these population modes, which is in a similar range as reported employing a similar approach in the adult UK Biobank sample[10].

*Consistently across sex and parent-ascribed race:* To assess whether any of the modes show specific patterns related to sex or parent-ascribed race, we computed CCA-ICA item correlations by sex (Supplementary Fig. 5-8) and parent-ascribed race (Supplementary Fig. 9-12), and compared these subject-weight-variable correlations with those estimated for the full analysis. Correlations for the four modes ranged between r=1 - r=.97 for sex and between r=.99 - r=.71 for parent-ascribed race.

*Consistency across sites:* To further assess the generalizability and robustness of the CCA-ICA patterns we computed site/scanner-wise CCA-ICA variable-correlations, and performed correlations comparing these to the full model (Supplementary Fig. 13-16). The pattern of the four modes are mostly consistent across sites, but with some sites deviating more from the full analysis modes than others (r=.95 - r=.31). All imaging phenotypes were adjusted for site, however, ABCD collects data at 21 sites across the continental US (https://abcdstudy.org/about) and some degree of population-level demographical differences are expected.

*Alternative approach without imputation:* To ensure that the results were not affected by the imputation procedure for missing data points, we also used an alternative and previously described approach[1] in which we estimated the subject x subject covariance matrix, ignoring missing values, before projecting this approximated covariance matrix to the nearest positive-definite covariance matrix using the MATLAB tool 'nearestSPD' (https://www.mathworks.com/matlabcentral/fileexchange/42885-nearestspd), thereby avoiding the need for imputation of missing values. The correlations between the CCA-ICA subject weights for the first four modes between the original analysis using imputation and this approach were r= .99, r=.97, r=.97 and r=.97, respectively.

*Alternative number of PCA components:* To investigate whether choosing a stricter criterion of inclusion of PCA components to CCA-ICA impacted the results, we reran the analysis with 100 PCA components, and compared the resulting CCA-ICA subject weights for the first four modes with those from the original analysis, yielding correlations of r=.98, r=.91, r=.91 and r=.94, respectively.

*Adjusting data for age$_2$:* To address the possibility of non-linear relationships between age and subject measures / MRI features we reran analysis with age$_2$ added along with the original confound variables, and compared the resulting CCA-ICA subject weights for the first four modes with those from the original analysis, yielding correlations of r=.99 for all four modes.

**Online methods references:**


1. Smith, S.M., *et al. Nature neuroscience* **18**, 1565 (2015).
10. Miller, K.L., *et al. Nature neuroscience* **19**, 1523 (2016).
23. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. *Journal of Experimental Social Psychology* **49**, 764-766 (2013).
24. Hagler, D.J., *et al. NeuroImage*, 116091 (2019).
25. Krzanowski, W.J. *Principles of multivariate analysis: a user's perspective* (Oxford University Press, Inc., 1988).
26. Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M. & Nichols, T.E. *NeuroImage* **92**, 381-397 (2014).
27. Hyvarinen, A. *IEEE Transactions on Neural Networks* **10**, 626-634 (1999).

# Supplementary information

## Contents:

### 1. Supplementary Figures

Supplementary Fig. 1: CCA-ICA permutation and cross-validation

Supplementary Fig. 2: Item text plots for mode 3 & 4

Supplementary Fig. 3: Principal component analysis

Supplementary Fig. 4: CCA-ICA subject weights, correlation and distributions

Supplementary Fig. 5: CCA-ICA Mode 1 by Sex

Supplementary Fig. 6: CCA-ICA Mode 2 by Sex

Supplementary Fig. 7: CCA-ICA Mode 3 by Sex

Supplementary Fig. 8: CCA-ICA Mode 4 by Sex

Supplementary Fig. 9: CCA-ICA Mode 1 by Parent-ascribed race

Supplementary Fig. 10: CCA-ICA Mode 2 by Parent-ascribed race

Supplementary Fig. 11 CCA-ICA Mode 3 by Parent-ascribed race

Supplementary Fig. 12: CCA-ICA Mode 4 by Parent-ascribed race

Supplementary Fig. 13: CCA-ICA Mode 1 by Site

Supplementary Fig. 14: CCA-ICA Mode 2 by Site

Supplementary Fig. 15: CCA-ICA Mode 3 by Site

Supplementary Fig. 16: CCA-ICA Mode 4 by Site

### 2. Supplementary Tables

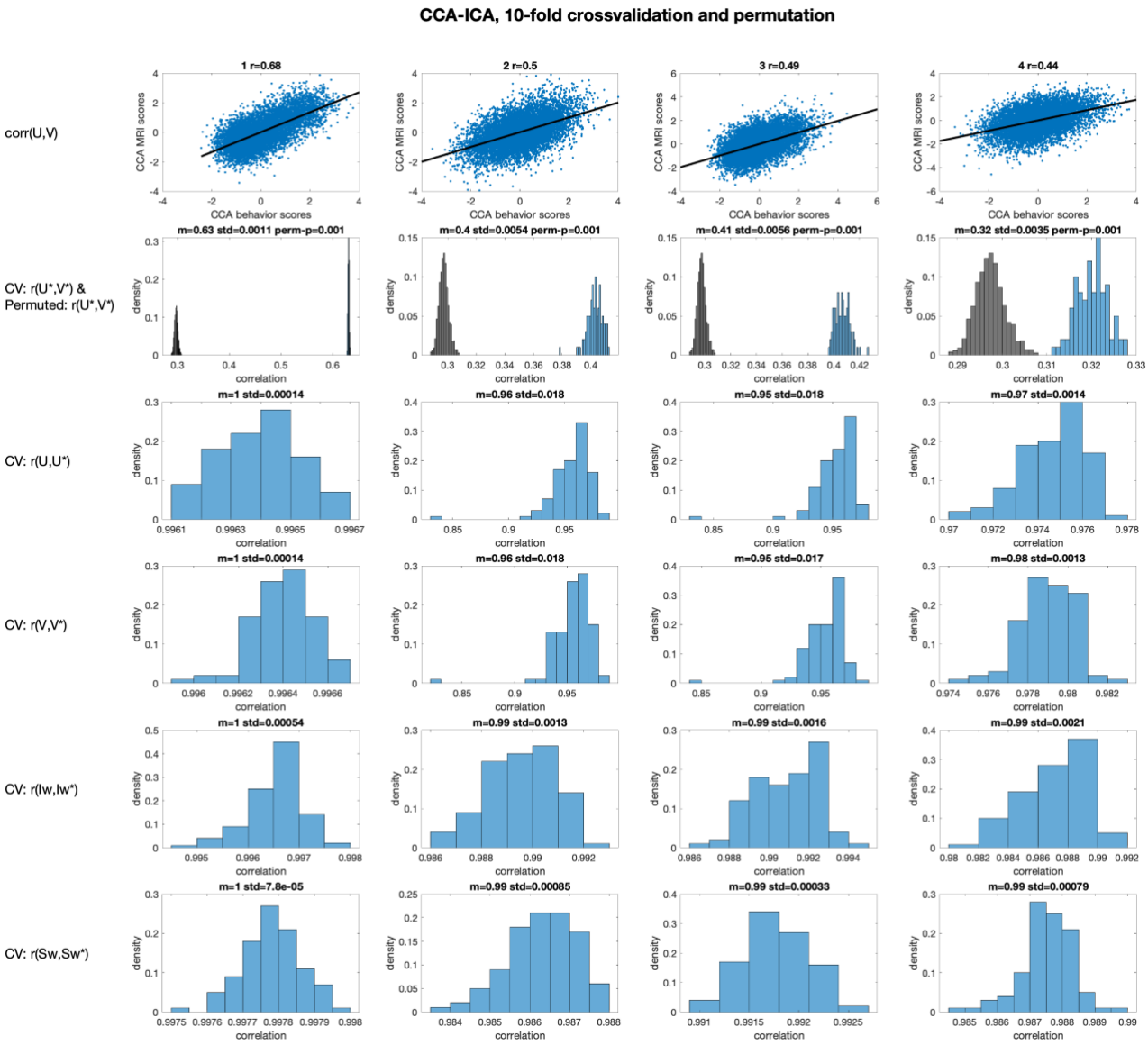
Tables available using this link to a OSF.io data repository (Center for Open Science) for the current paper.

**Supplementary Tables 1-4:** List of all included variables, their description, ICA-weights and correlations with CCA-ICAs Mode 1-4 subject weights.

**Supplementary Table 5:** Lists the included behavioral variable names, and their associated descriptions and instrument.

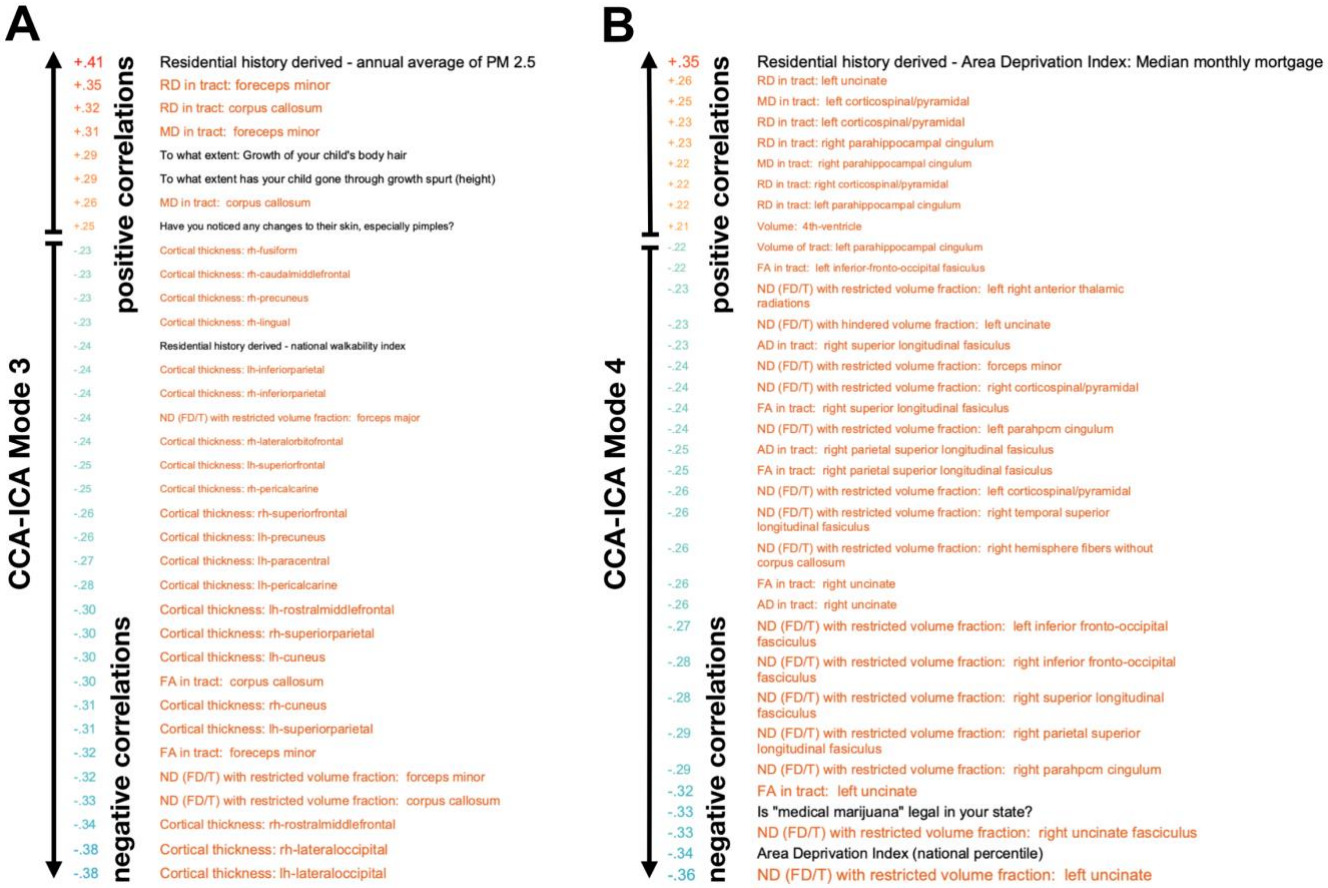
**Supplementary Table 6:** Lists the included MRI variable names, and their associated descriptions.

# 1. Supplementary Figures

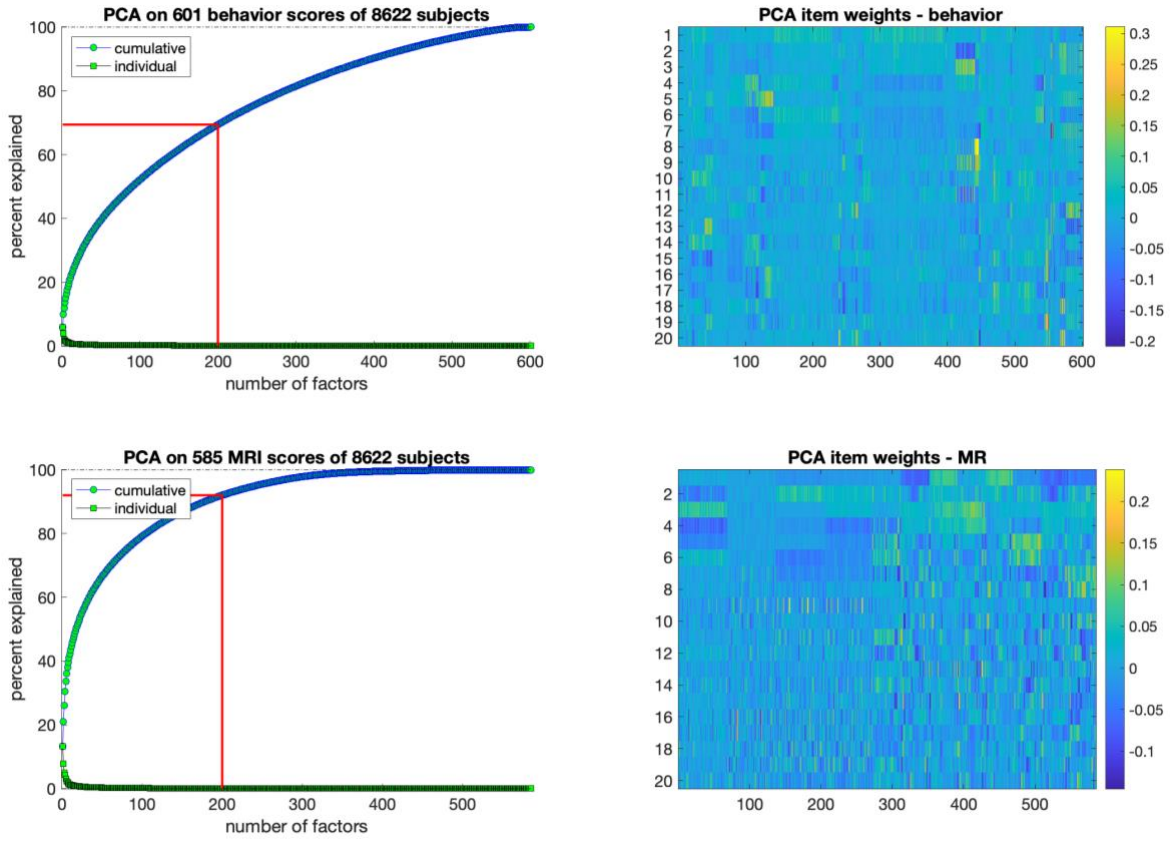



**Supplementary Fig. 1:** The four columns show results for CCA and CCA-ICA modes 1-4, respectively. **Row 1** shows the scatter-plots of behavioral and imaging scores from the in-sample CCA, and the title shows the canonical correlation for the full in-sample analysis (100% of data). **Row 2** Shown in blue is the distribution of out-of-sample canonical correlation values across 100 iterations of a 10-fold cross-validation procedure. Each fold (10% of participants) was kept out once, while estimating PCA and CCA using the remaining folds (90% of participants). Each of the 100 correlations plotted here is the mean canonical correlation between CCA weights derived for kept-out participants across the 10 folds in each of the 100 iterations. Title displays the mean out-of-sample canonical correlations, along with the standard deviation. The null-distribution generated by the permutation procedure is shown in black in the same plot: we ran 1000 iterations of the 10-fold cross-validation procedure, randomizing the rows (participants) of the imaging feature-matrix for each run. To correct for familywise error (FWE) we collected the largest canonical correlation (i.e. the value for mode 1) to form the null-distribution, and used this to calculate p-values using the mean canonical correlation for kept-out participants from the cross-validation procedure. **Row 3** shows the correlations between CCA behavioral measure-weights derived during cross-validation for folds kept out of the estimation, with those from the full analysis. **Row 4** shows the correlations between CCA imaging feature-

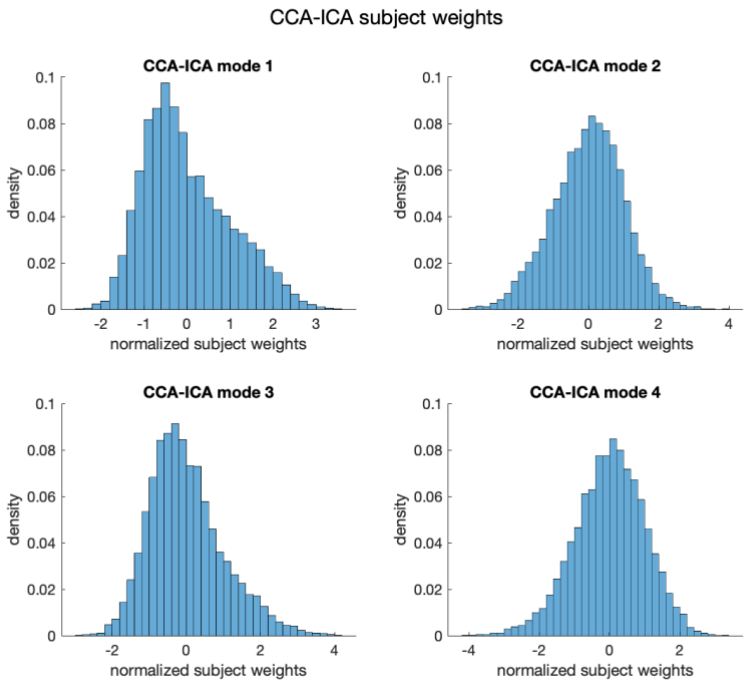
weights derived during cross-validation for folds kept out of the estimation, with those from the full analysis. **Row 5** shows the correlations between ICA behavioral-measure and imaging-feature weights derived during cross-validation for folds kept out of the estimation, with those from the full analysis. **Row 6** shows the correlations between ICA subject weights derived during cross-validation for folds kept out of the estimation, with those from the full analysis. Y-axis in all histograms represent densities.




**Supplementary Fig. 2:** Item text plot showing the correlation of original adjusted behavioral measures and imaging features with (A) CCA-ICA Mode 1 subject weights and (B) CCA-ICA Mode 1 subject weights. Numbers on the left are correlations, hot/cold colors represent positive/negative correlations, respectively, arrows indicate increasing positive/negative correlations. Text on the right represent the behavioral measures (black) and imaging features (orange). The top 35 items are shown, for a full list of all measures and associated correlations for modes 3 and 4, see Supplementary Tables 3 and 4.

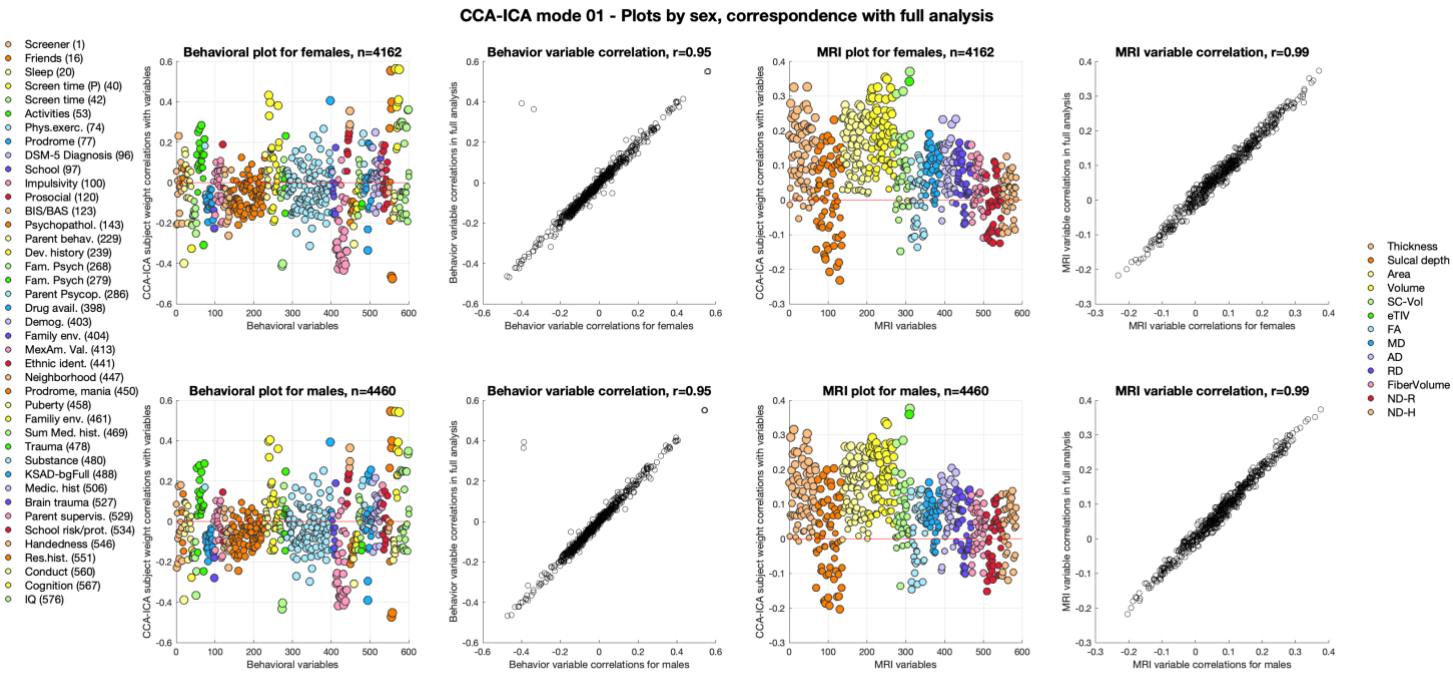



**Supplementary Fig. 3:** PCA-analysis. Top and bottom panels show behavior/clinical/demographics, and imaging phenotypes, respectively. Left panels: Cumulative explained variance as a function of included eigenvalues/components. Red line demarks the 200 included PCA-components. Right panels: PCA-item weights for the first 20 components.



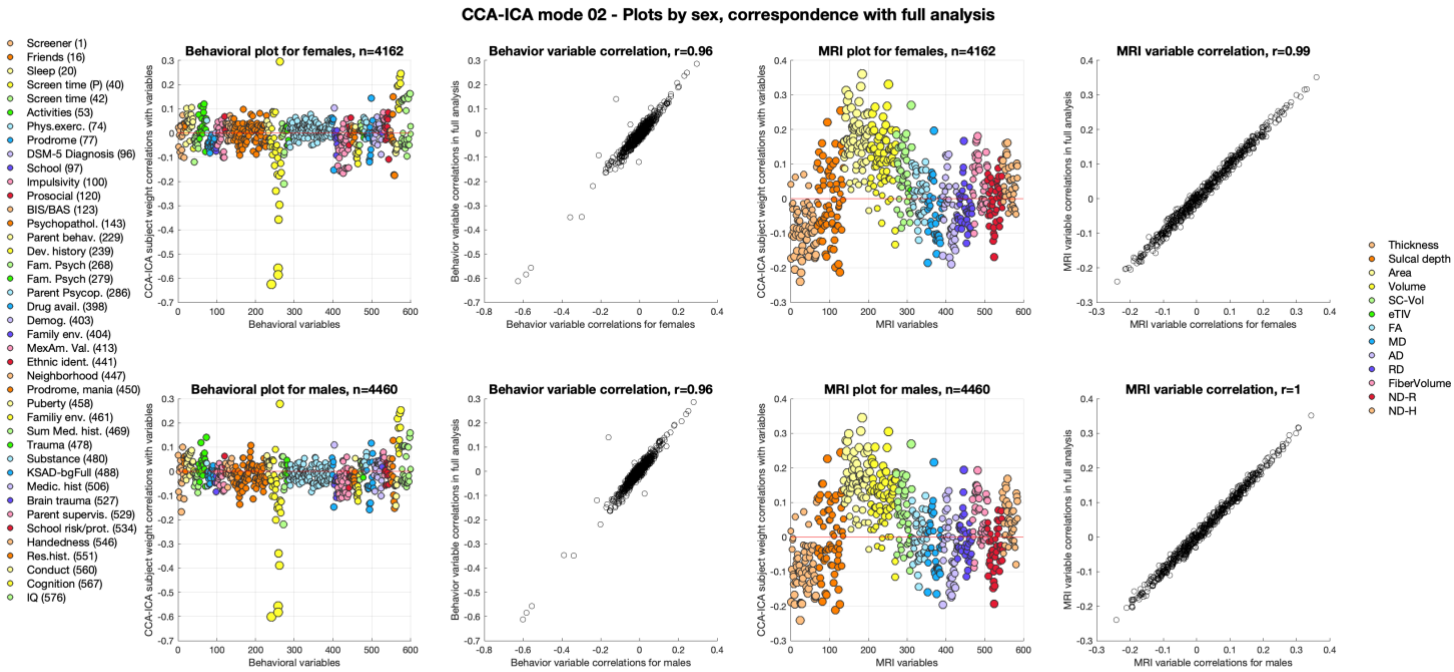

**Supplementary Fig. 4:** Histograms of CCA-ICA subject weights for the four significant modes.

## Supplementary Fig. 5 - 8 – Mode 1 - 4 by sex

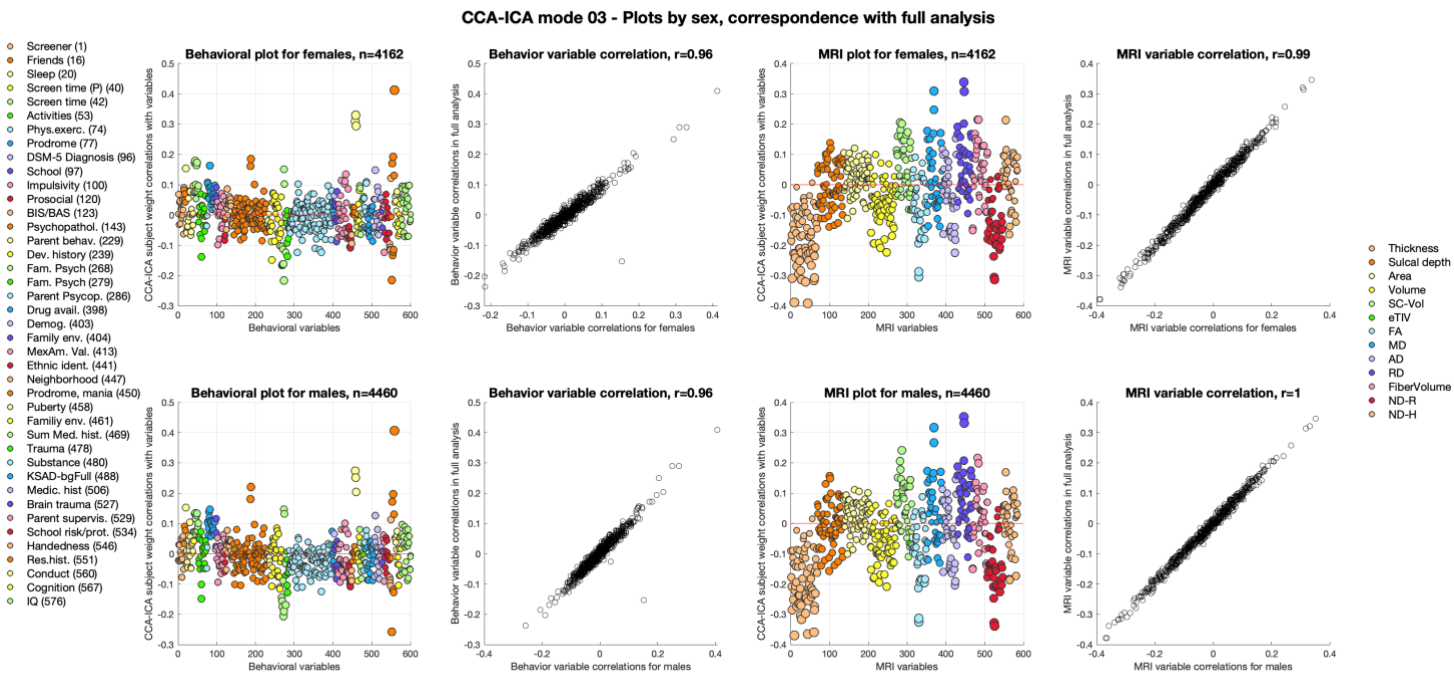



**Supplementary Fig. 5:** Colored plots show the correlation between the CCA-ICA mode 1 subject weights and the adjusted original behavioral and MRI data, plotted separately for females (F, top panels) and males (M, bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the

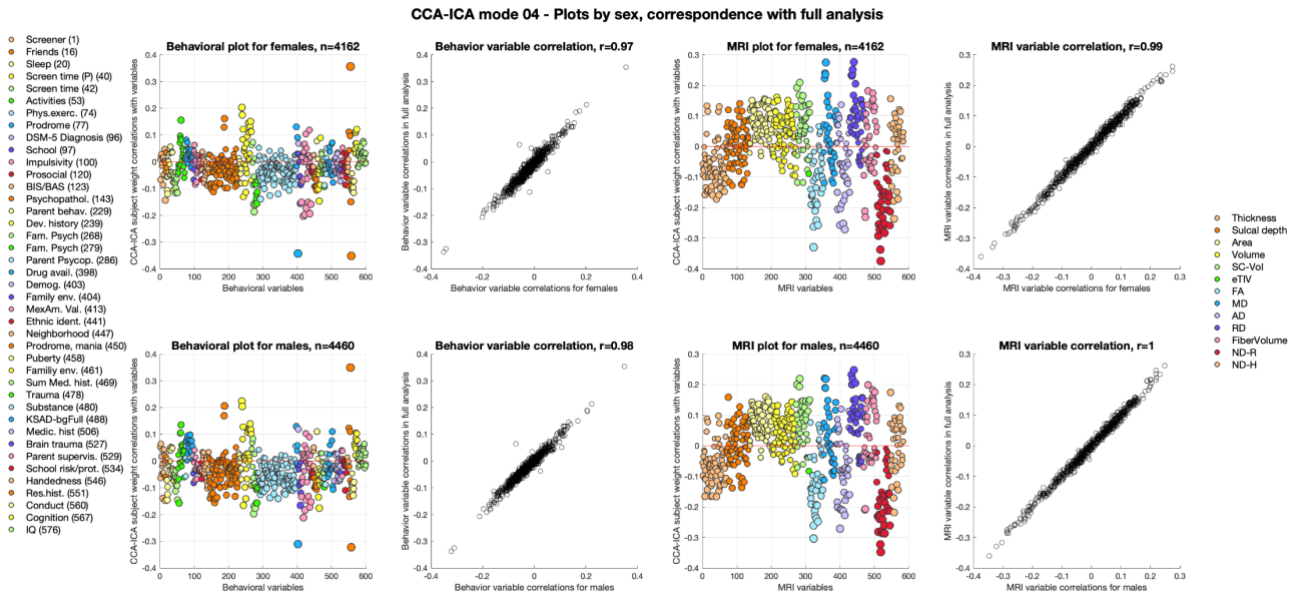
adjusted original behavioral and MRI data, plotted against the r-values for sex-specific CCA-ICA subject weight correlations.




**Supplementary Fig. 6:** Colored plots show the correlation between the CCA-ICA mode 2 subject weights and the adjusted original behavioral and MRI data, plotted separately for females (F, top panels) and males (M, bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data, plotted against the r-values for sex-specific CCA-ICA subject weight correlations.



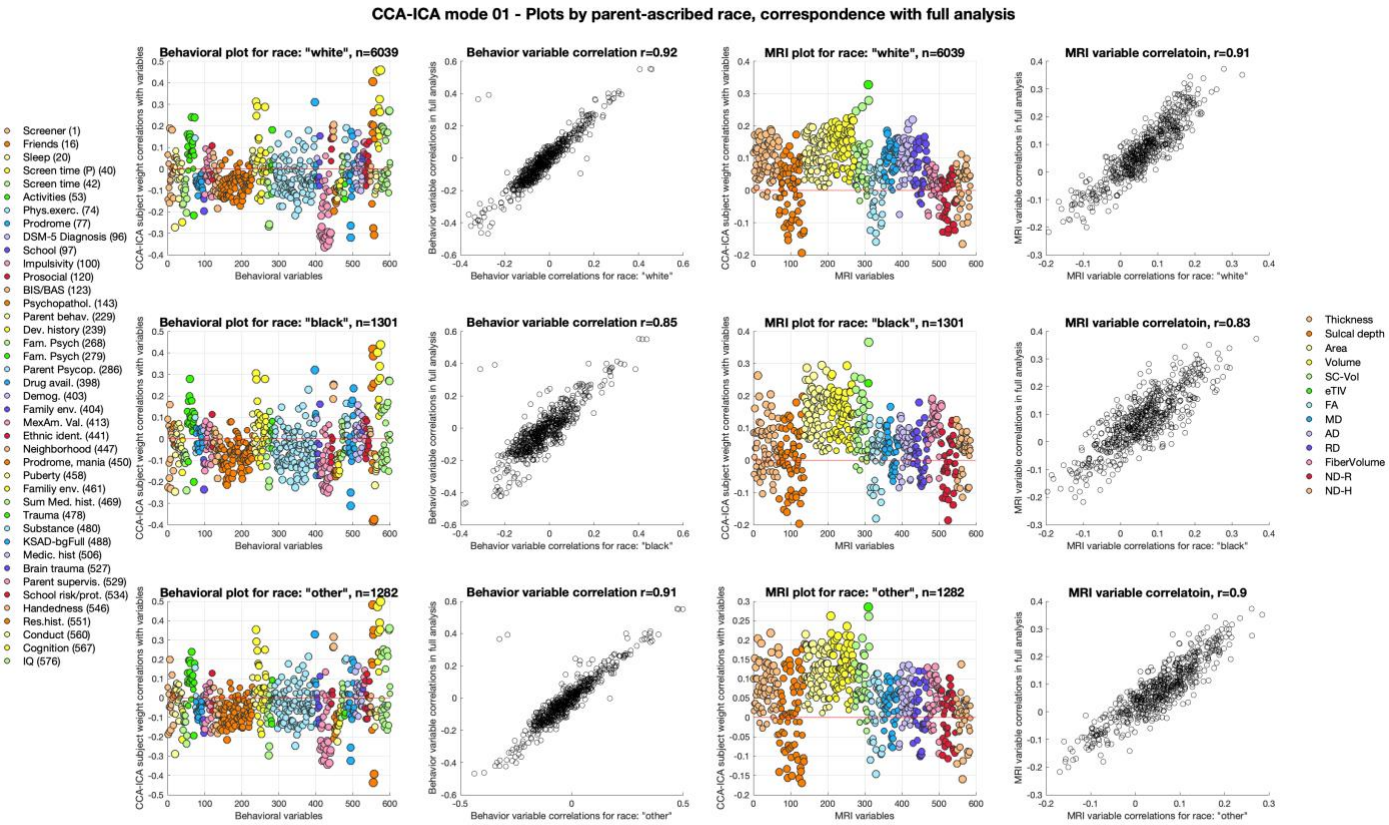

**Supplementary Fig. 7:** Colored plots show the correlation between the CCA-ICA mode 3 subject weights and the adjusted original behavioral and MRI data, plotted separately for females (F, top panels) and males (M, bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data, plotted against the r-values for sex-specific CCA-ICA subject weight correlations.



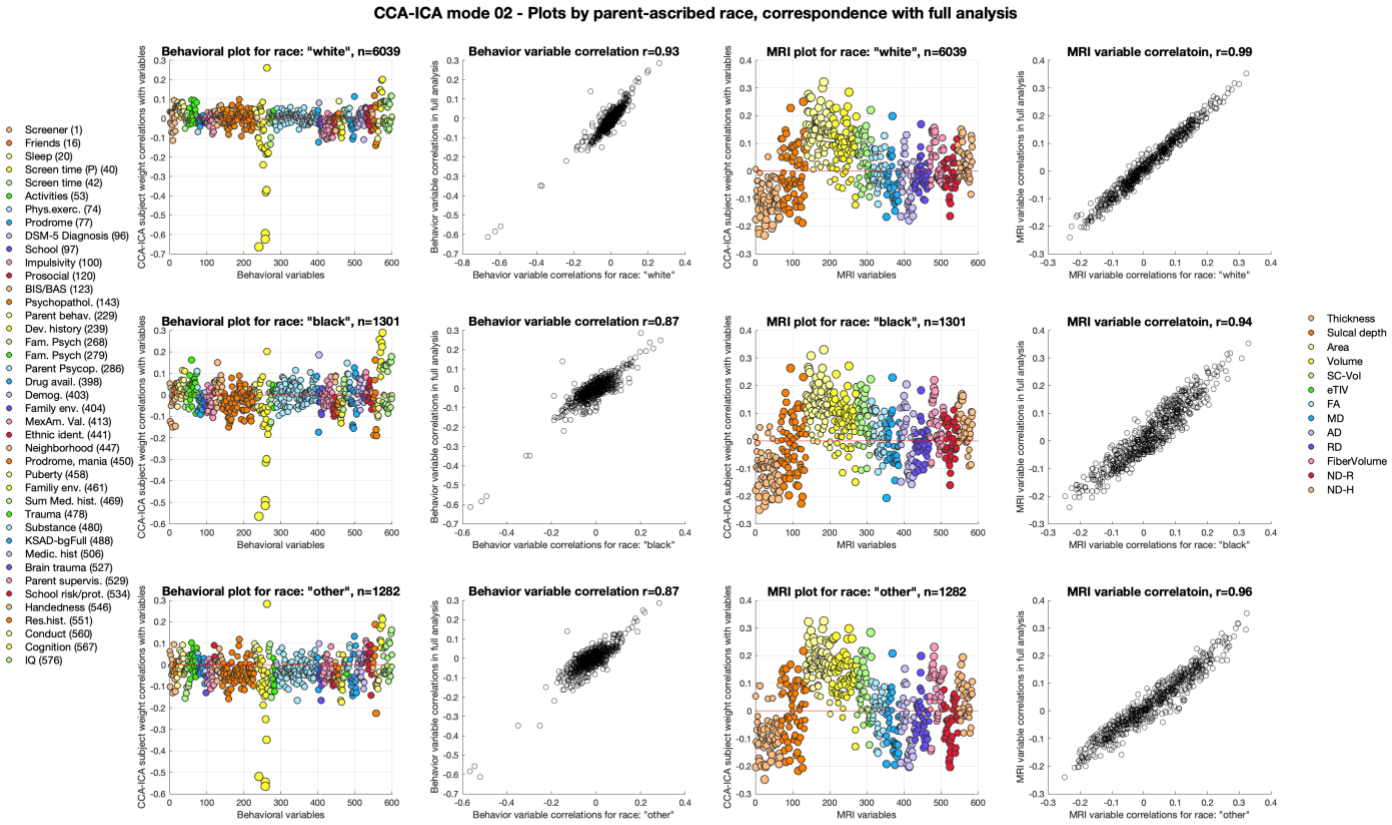

**Supplementary Fig. 8:** Colored plots show the correlation between the CCA-ICA mode 4 subject weights and the adjusted original behavioral and MRI data, plotted separately for females (F, top panels) and males (M, bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the

adjusted original behavioral and MRI data, plotted against the r-values for sex-specific CCA-ICA subject weight correlations.

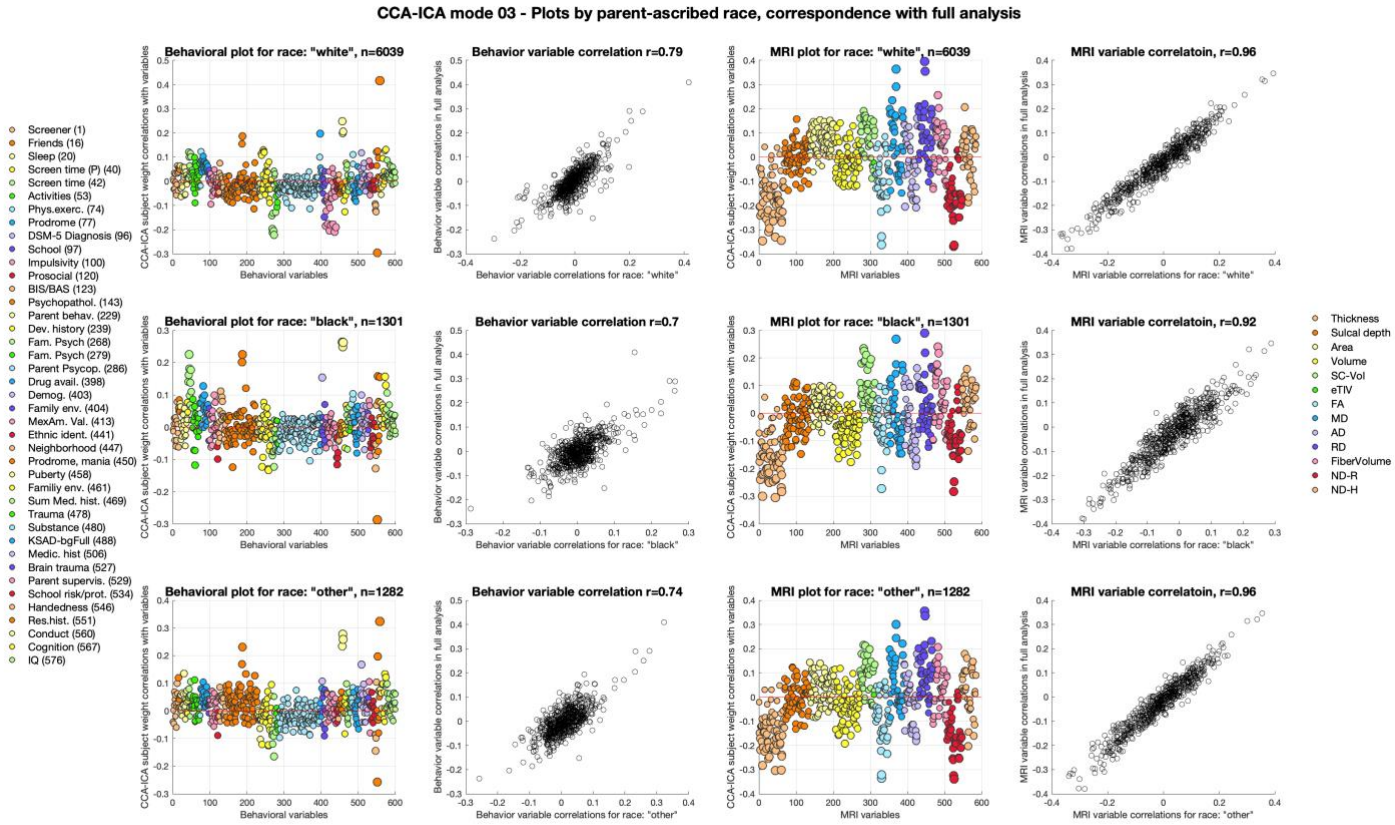
## Supplementary Fig. 9 - 12 – Mode 1 - 4 by parent-ascribed race



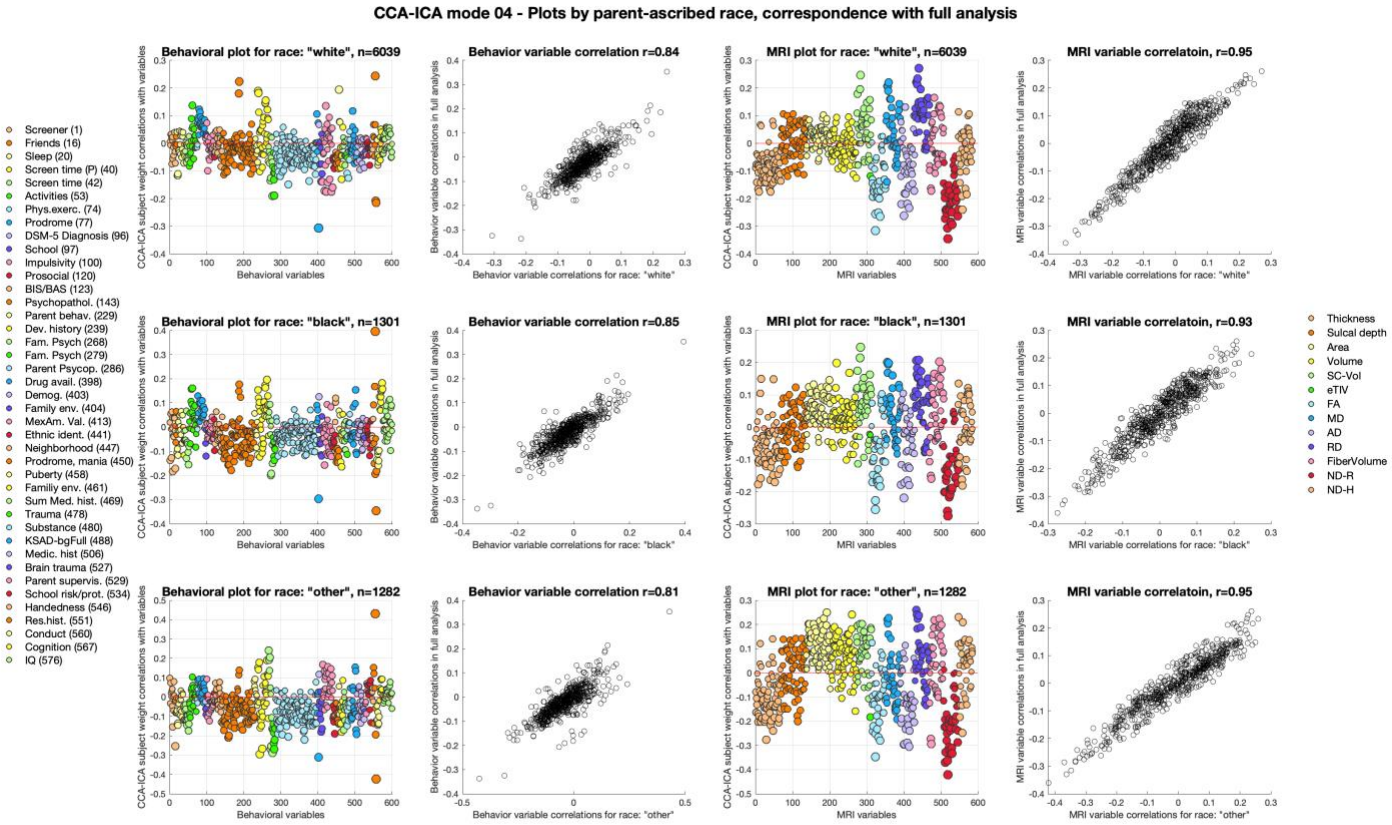

**Supplementary Fig. 9:** Colored plots show the correlation between the CCA-ICA mode 1 subject weights and the adjusted original behavioral and MRI data, plotted separately for parent-ascribed race ("white" in top panels, "black" in middle panels, "other" in bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for parent-ascribed race-specific CCA-ICA subject weight correlations.




**Supplementary Fig. 10:** Colored plots show the correlation between the CCA-ICA mode 2 subject weights and the adjusted original behavioral and MRI data, plotted separately for parent-ascribed race ("white" in top panels, "black" in middle panels, "other" in bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for parent-ascribed race-specific CCA-ICA subject weight correlations.

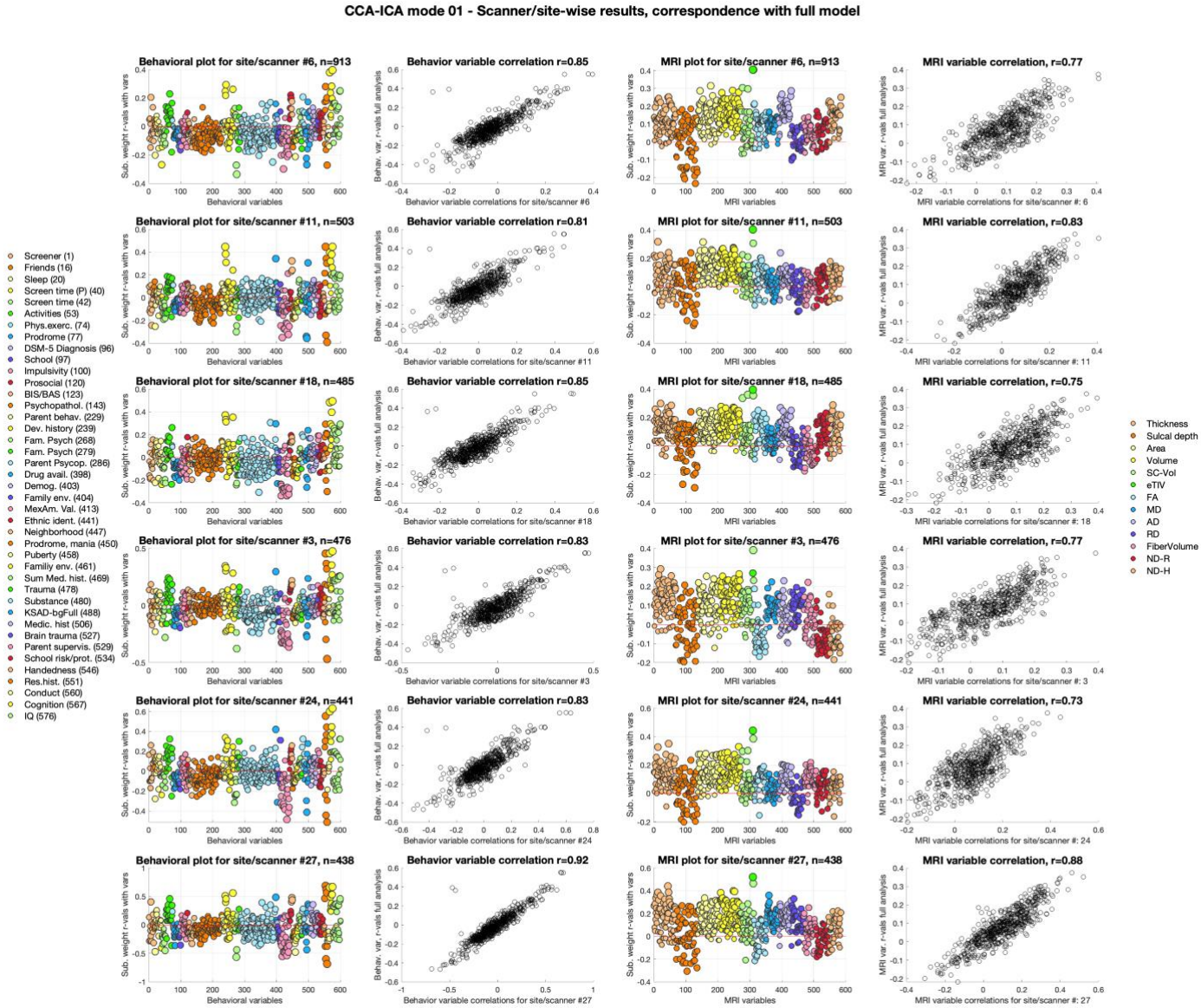



**Supplementary Fig. 11:** Colored plots show the correlation between the CCA-ICA mode 3 subject weights and the adjusted original behavioral and MRI data, plotted separately for parent-ascribed race ("white" in top panels, "black" in middle panels, "other" in bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for parent-ascribed race-specific CCA-ICA subject weight correlations.

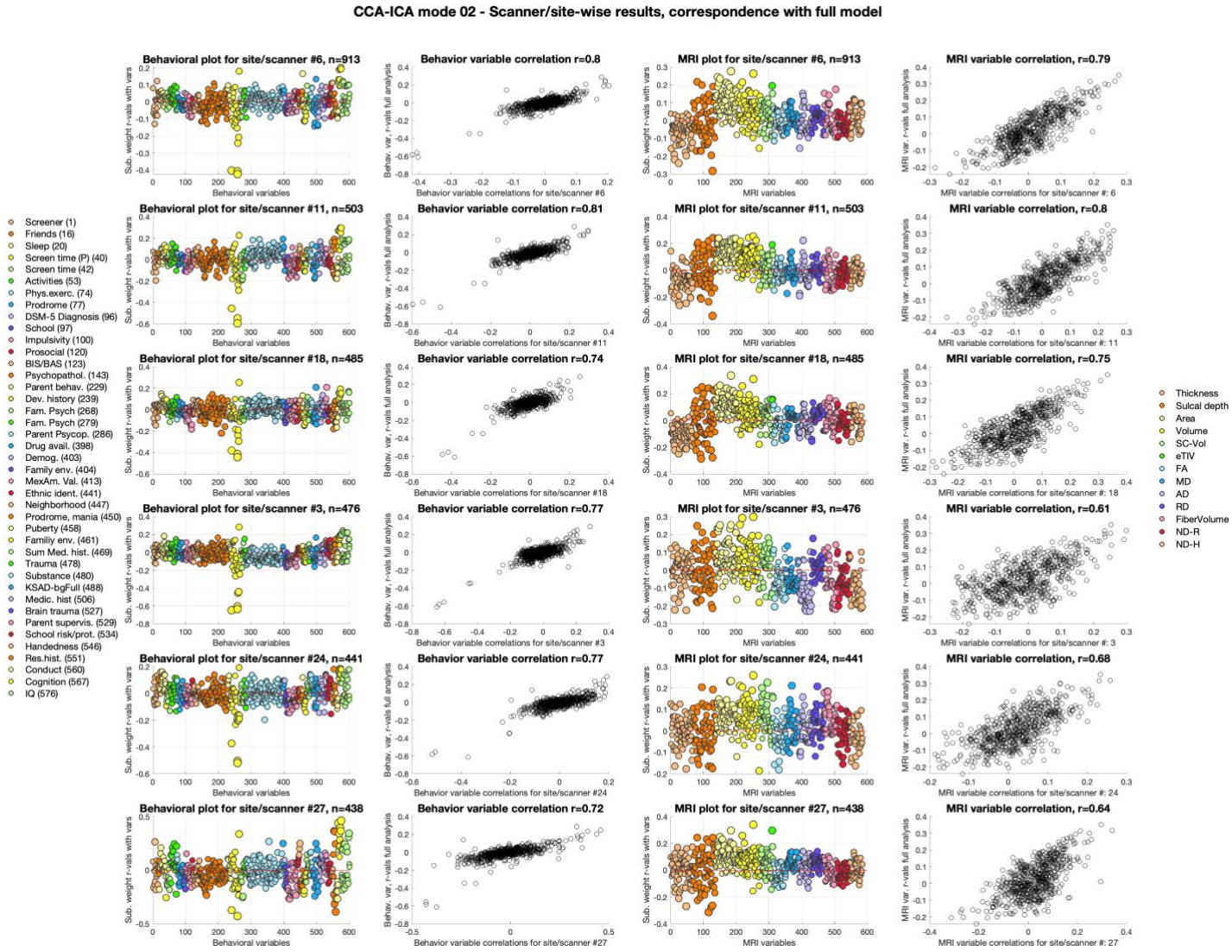



**Supplementary Fig. 12:** Colored plots show the correlation between the CCA-ICA mode 4 subject weights and the adjusted original behavioral and MRI data, plotted separately for parent-ascribed race ("white" in top panels, "black" in middle panels, "other" in bottom panels), while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for parent-ascribed race-specific CCA-ICA subject weight correlations.

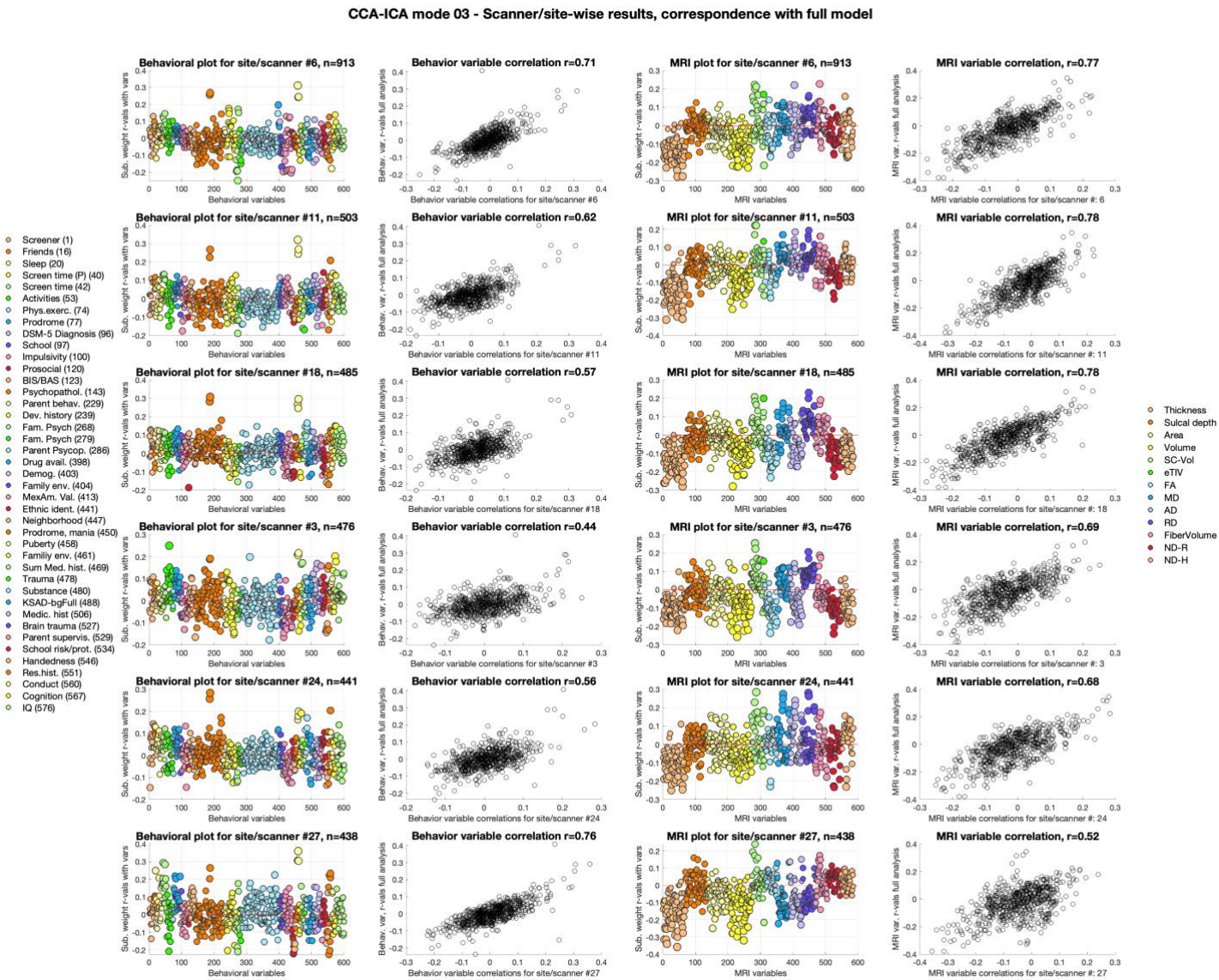
## Supplementary Fig. 13-16 – Mode 1 – 4 by site/scanner



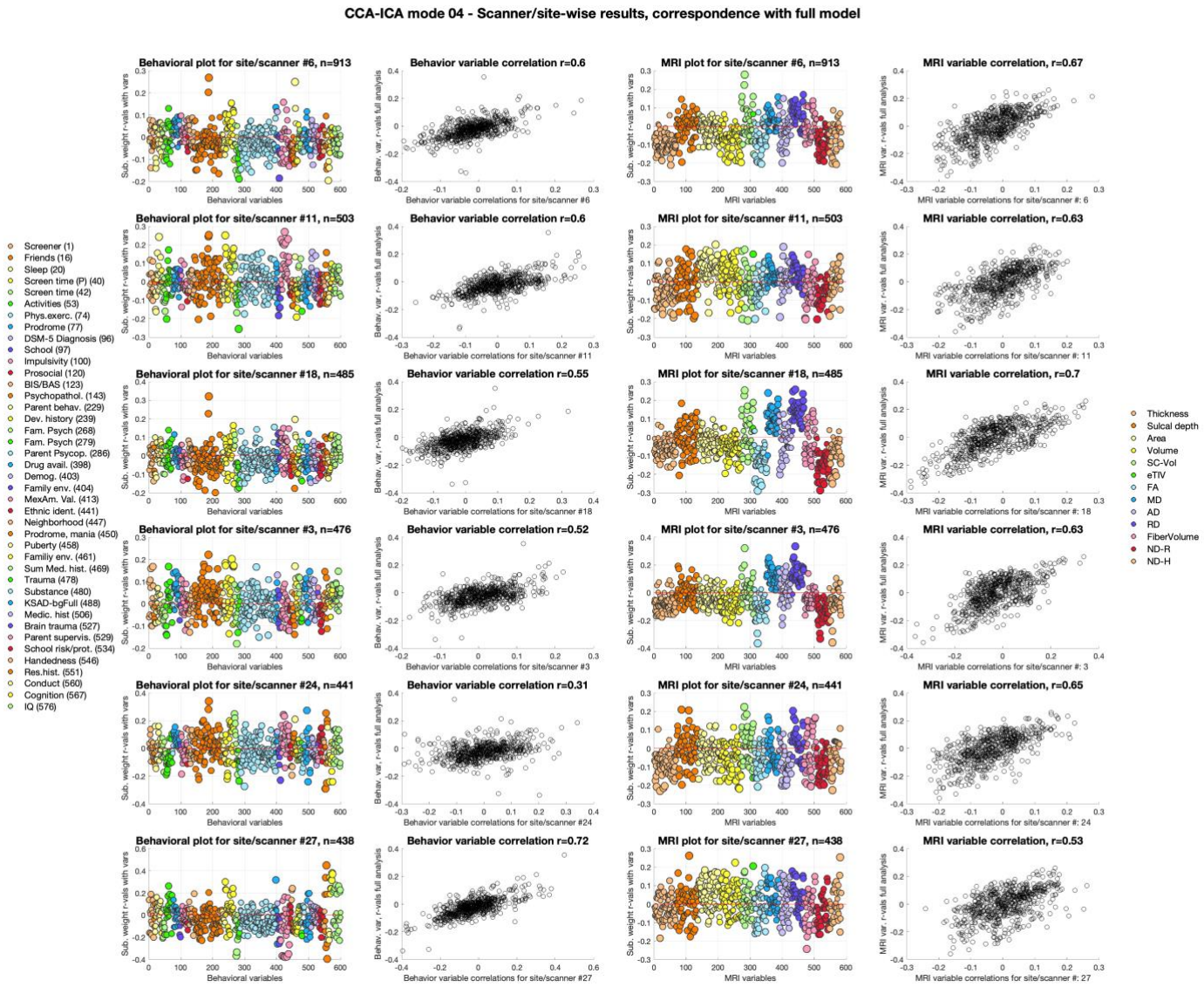

**Supplementary Fig. 13:** Colored plots show the correlation between the CCA-ICA mode 1 subject weights and the adjusted original behavioral and MRI data, plotted separately for the sites/scanner constituting >= 5% of the total sample, in descending order, while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for scanner/site-specific CCA-ICA subject weight correlations.




**Supplementary Fig. 14:** Colored plots show the correlation between the CCA-ICA mode 2 subject weights and the adjusted original behavioral and MRI data, plotted separately for the sites/scanner constituting >= 5% of the total sample, in descending order, while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for scanner/site-specific CCA-ICA subject weight correlations.




**Supplementary Fig. 15:** Colored plots show the correlation between the CCA-ICA mode 3 subject weights and the adjusted original behavioral and MRI data, plotted separately for the sites/scanner constituting >= 5% of the total sample, in descending order, while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for scanner/site-specific CCA-ICA subject weight correlations.




**Supplementary Fig. 16:** Colored plots show the correlation between the CCA-ICA mode 4 subject weights and the adjusted original behavioral and MRI data, plotted separately for the sites/scanner constituting >= 5% of the total sample, in descending order, while scatter-plots shows the r-values for correlations between all CCA-ICA subject weights and the adjusted original behavioral and MRI data plotted against the r-values for scanner/site-specific CCA-ICA subject weight correlations.