SUPPLEMENTARY MATERIALS

Table of Contents

Participants:

A total of 9,498 participants 8-22 years of age received cognitive assessment and clinical phenotyping as part of the Philadelphia Neurodevelopmental Cohort (PNC), a large community-based sample of youth (1). From the pool of 9,498 participants, 6,476 were ineligible for the current study for 1) medical disorders that could impact brain functioning (n=2,347), 2) missing age at clinical assessment (n=92), or 3) missing depression or overall psychiatric sub-score (n=87); several subjects were ineligible based on multiple criteria. Of the remaining participants, 712 met screening criteria for a lifetime history of a major depressive episode (referred to as depressed youth, or DY) and 2,310 were typically developing (TD) youth with no psychiatric diagnosis. Our analysis evaluated a final sample of 712 depressed youth and 712 typically developing youth (total n=1,424).

Of the 1,424 individuals in the final group, a subset of participants (n=368, TD=200) also completed *n*-back functional magnetic resonance imaging (fMRI) and passed strict quality control including T1 structural and motion exclusion.

Matching Procedure:

Using the R package Matchit, depressed youth were matched on age and sex with typically developing youth. Given that not all participants underwent neuroimaging, the match was performed in multiple steps to allow us to enrich our typically developing group for children who were in the subset that obtained neuroimaging. All depressed youth were included in the final sample. First, depressed youth with imaging were matched with TD youths with imaging.

Next, depressed youth without imaging were matched with the remaining TDs with imaging that were not matched in the first step. Youths with poor T1 quality were excluded. The results from both matches were combined, yielding a group of 712 DYs and 712 TDs. After additional quality assessment of the *n*-back task-related imaging data, 368 youth (DY =168, TD = 200) were included in the functional imaging analysis.

Clinical Assessment:

As described in detail in our previous work, assessment of lifetime psychopathology was conducted using GOASSESS, a structured screening interview administered to probands (age 11-22 years) and collateral informants of probands (age 8-17 years), based on a modified version of the Kiddie-Schedule for Affective Disorders and Schizophrenia and Diagnostic and Statistical Manual of Mental Disorders, 4th edition, Text Revision criteria (2). The GOASSESS interview assesses lifetime occurrence of mood (major depressive episode, mania), anxiety (agoraphobia, generalized anxiety, panic, specific phobia, social phobia, separation anxiety, posttraumatic stress), behavioral problems (oppositional defiant, attention deficit/hyperactivity, conduct), psychosis, eating disorder (anorexia, bulimia), and suicidal symptoms. Among the GOASSESS questions, 107 screening items administered to all participants were used for the current investigation. Of note, due to comorbidity, participants may be represented in more than one diagnostic category. The median interval of time between clinical assessment and neuroimaging was 2 months.

Bachelor- and master-level assessors underwent a 25-hr training protocol developed and implemented by the PNC Clinical Core; it included didactic sessions, assigned readings, and supervised pairwise practice. Assessors were certified for independent assessments through a

standardized procedure requiring observation by a certified clinical observer who rated the proficiency of the assessor on a 60-item checklist of interview procedures. The number of certified assessors was 55. Additionally, responses coded in GOASSESS by the assessor were required to correspond to responses coded by a certified clinical observer. Assessors who did not achieve these standards were required to undergo repeat observation until the passing criteria were met. Assessor drift was monitored and corrected through periodic review of audio-recordings of real interviews, and re-training and re-certification was conducted at data collection mid-point. Assessors were assigned a maximum of 10 interviews a week, with the goal of completing 5–7 interviews per week. To maximize the quality of interview data, each assessment underwent a computerized error-checking algorithm that identified areas requiring assessors' attention, and a standardized post-administration review process by certified clinical reviewers. Results were reported to assessors and supervisors. A computerized chart review module provided management tools for the comprehensive review process for supervisors, reviewers, and assessors, as well as an automated check to ensure that all steps were completed successfully. Data were checked and corrected prior to final inclusion in the dataset.

Cognitive Assessment:

Cognition was assessed using the University of Pennsylvania Computerized Neurocognitive Battery (CNB), which has been described previously (3). Briefly, 14 cognitive tests evaluating aspects of cognition, including executive control, episodic memory, complex reasoning, social cognition, and sensorimotor speed, were administered in a fixed order. Except for two sensorimotor tests that only measure speed, each test provides measures of both accuracy and speed, yielding 26 total measures (abstraction/mental flexibility, attention, working memory, verbal memory, face memory, spatial memory, language/verbal reasoning, nonverbal reasoning,

spatial reasoning, emotion recognition, emotion discrimination, age discrimination, motor, sensorimotor). Academic skills were measured with the Wide Range Achievement Test, 4th Edition (WRAT-4) reading subscale with total subscale scores reported as T-scores. Youth performance on each measure was transformed into a Z-score, which was used for further analysis.

Clinical Factor Analysis:

To provide a dimensional summary of psychopathology, we used an exploratory factor analysis to derive latent factors from the item-level data from the GOASSESS interview (2, 4). Previous confirmatory bifactor analyses performed on the GOASSESS utilized all 112 item-level symptom questions. In the current study, to prevent circularity between our inclusion criteria and outcome measures, we modified this confirmatory bifactor analysis to exclude five depression items. Specifically, these five depression items were also used to establish the categorical diagnosis of major depressive disorder used in sample selection. Thus, by definition, all depressed youth presented with these symptoms. Because our goal was to evaluate dimensions of psychopathology independent of our sample construction criteria, we excluded these five items. This exploratory factor analysis yielded four correlated dimensions of psychopathology including factors for anxious-misery (31 items), psychosis (26 items), externalizing behavior (25 items), and fear (25 items). We then used a confirmatory bifactor analysis implemented in Mplus11 to orthogonally model the four factors plus overall psychopathology, which represents the symptoms common across all psychiatric disorders.

***n*-back Task:**

Subjects completed a fractal version of the *n*-back task during their fMRI scan (4,5). During the task, a fractal was presented for 500 ms followed by a 2500 ms interstimulus interval.

This task was used to probe working memory and had 3 conditions: 0-, 1-, and 2-back. During the 0-back, subjects responded by pressing a button when the fractal presented matched a predefined fractal. During the 1-back condition, subjects responded when the fractal presented was the same as the one preceding it. During the 2-back condition, subjects responded when the fractal was identical to the one two before it. Each condition consisted of three 20-trial blocks, each preceded by a 9s instruction period, with a target to foil ratio of 1:3. The task included a total of 45 targets and 135 foils, as well as three 24 s blocks of rest during which a fixation crosshair was displayed.

_n_-back Performance Analysis:

Correct responses, false positives, and median response time to correct responses were calculated for all *n*-back loads. As previously, to relate task performance to the neuroimaging data, task performance was summarized using the signal detection measure d' (6,7). This measure considers both correct responses and false positives to limit the influence of response bias.

Image Acquisition and Preprocessing:

Acquisition

Imaging data were acquired on a single 3T Siemens TIM Trio whole-body scanner using a 32-channel head coil. A magnetization-prepared rapid acquisition gradient echo T1-weighted (MPRAGE) image (TR, 1810 ms; TE, 3.51 ms; TI, 1100 ms; FOV, 180 × 240 mm; matrix, 192 × 256; 160 slices; slice thickness/gap, 1/0 mm; flip angle, 9°; effective voxel resolution, 0.9 × 0.9 × 1 mm) and B0 field map (TR, 1000 ms; TE1, 2.69 ms; TE2, 5.27 ms; 44 slices; slice thickness/gap, 4/0 mm; FOV, 240 mm; effective voxel resolution, 3.8 × 3.8 × 4 mm) were acquired to aid spatial normalization to standard space and application of distortion correction

procedures, respectively. Functional images were then obtained using a whole-brain, single-shot, multislice, gradient-echo echoplanar sequence (231 volumes; TR, 3000; TE, 32 ms; flip angle, 90°; FOV, 192 × 192 mm; matrix 64 × 64; 46 slices; slice thickness/gap 3/0 mm; effective voxel resolution, 3.0 × 3.0 × 3.0 mm).

Preprocessing

As previously described, fMRI data were pre-processed with FSL, including skull removal with BET, slice time correction, motion-correction with MCFLIRT, spatial smoothing (6 mm FWHM), and mean-based intensity normalization. Subject-level timeseries analyses were carried out using FILM3 (FMRIB's Improved Linear Model) with local autocorrelation correction (5). The three condition blocks (0-back, 1-back, and 2-back) were modeled using a canonical (double-gamma) hemodynamic response function, with six motion parameters and the instruction period included as nuisance covariates. The rest condition served as the unmodeled baseline. The median functional and anatomical volumes were co-registered using boundary-based registration with integrated distortion correction using FUGUE. The anatomical image was normalized to a custom 1mm template using the top-performing diffeomorphic SyN registration of ANTs (8). All transformed images (distortion correction, co-registration, normalization, and down-sampling to $2mm^3$) were concatenated so that only one interpolation was required. The statistical maps for the contrast of interest (2-back > 0-back) were then used in the group-level analyses. This contrast was implemented using the task module of XCP (9).

Construction of Functional Regions of Interest

Functional regions of interest were delineated from the 2-back > 0-back map from the complete subsample of youth who underwent *n*-back imaging and met quality control (n=951) as previously described (4). Specifically, to isolate core regions of the executive network with a high degree of anatomic specificity, the 2-back>0-back map was thresholded at z > 20; clusters of <100 voxels were discarded. This high threshold was selected because at lower thresholds substantial volumes of white matter were included due to spatial smoothing, the very high statistical power of the large sample, and the robust nature of the contrast. Next, a watershed algorithm implemented in MATLAB was applied to parse confluent regions of interest. The watershed procedure separates contiguous regions of voxels into subregions by first identifying local maxima, each of which becomes a peak within a subregion. Subregion boundaries are defined by the watershed algorithm, which computes how water would drain into the inverted topology of the activation map. Last, a second threshold on spatial extent (k < 50 voxels) was used to remove undesirably small subregions by absorbing them into the nearest neighboring suprathreshold subregion. When this procedure was applied to the activated contrast of 2-back > 0-back, a set of 21 functional ROIs was produced within the executive network (see Results) that corresponded to a high degree with previously published meta-analyses of working memory (10,11). Finally, signal change in the 2-back > 0-back contrast in each of these 21 regions of interests was extracted. The final regions of interest included the right and left crus I, right and left crus II, right and left dorsolateral prefrontal cortices (anterior region), right and left dorsolateral prefrontal cortices (posterior region), right and left dorsal frontal gyri (part of middle frontal gyrus), right and left precuneus, right and left thalamus, right and left insula, right and left frontal poles, right and left parietal cortices, and dorsal anterior cingulate gyrus.

Permutation Testing:

Permutation testing is a widely used methodology for evaluating statistical significance when the null distribution of data is unknown (12). For null distribution of the subtypes, half of the healthy controls were randomly assigned to the control group (n=356 of total 712) and another half to the pseudo-patient group (n=356 of total 712). These samples were permuted 50 times and HYDRA was run each time. To fairly compare these results with the real-patient results, half of the healthy controls (n=356 of total 712) and half of the real-patient group (n=356 of total 712) were selected. These sample selections were also permuted 50 times and HYDRA was run each time. Finally, the clustering reproducibility assessed as ARI was compared between the subtypes derived from the null distribution and real-patient samples. The ARI for the 3 subtypes was significantly higher in the real-patient samples compared to that of the null distribution ($P_{fdr} = 0.011$) while the ARIs of other subtypes were not significantly different ($P_{fdr} = 0.192$).

Resting-state Functional Connectivity Analyses

Resting-state fMRI acquired as part of the PNC was available in a subset of the participants considered in this report (n=333 total, TD = 180, S1 = 61, S2 = 48, S2 = 44). We used these data to evaluate differences in functional connectivity between our data-driven cognitive subtypes. Below, we describe the image acquisition, image processing, and statistical testing used to evaluate differences between cognitive subtypes using resting-state functional connectivity.

Image Acquisition

Resting-state functional magnetic resonance imaging (rs-fMRI) datasets were acquired as part of the Philadelphia Neurodevelopmental Cohort (13). Structural and functional subject data were acquired on a 3T Siemens Tim Trio scanner with a 32-channel head coil (Erlangen, Germany), as previously described (13,14). High-resolution structural images were acquired in order to facilitate alignment of individual subject images into a common space. Structural images were acquired using a magnetization-prepared, rapid-acquisition gradient-echo (MPRAGE) T1-weighted sequence (T_R = 1810ms; T_E = 3.51 ms; FoV = 180 × 240 mm; resolution 0.9375 × 0.9375 × 1 mm). Approximately 6 minutes of task-free functional data were acquired for each subject using a blood oxygen level-dependent (BOLD-weighted) sequence (T_R = 3000 ms; T_E = 32 ms; FoV = 192 × 192 mm; resolution 3 mm isotropic; 124 volumes). In order to minimize motion, prior to data acquisition subjects' heads were stabilized in the head coil using one foam pad over each ear and a third over the top of the head. During the resting-state scan, a fixation cross was displayed as images were acquired. Subjects were instructed to stay awake, keep their eyes open, fixate on the displayed crosshair, and remain still.

Image Processing

Task-free functional images were processed using one of the top-performing pipelines for removal of motion-related artifact (15). Preprocessing steps included correction for distortions induced by magnetic field inhomogeneities using FSL's FUGUE utility, removal of the 4 initial volumes of each acquisition, realignment of all volumes to a selected reference volume using MCFLIRT (16), removal of and interpolation over intensity outliers in each voxel's time series using AFNI's 3DDESPIKE utility, demeaning and removal of any linear or quadratic trends, and

co-registration of functional data to the high-resolution structural image using boundary-based registration (17). The artefactual variance in the data was modelled using a total of 36 parameters, including the six framewise estimates of motion, the mean signal extracted from eroded white matter and cerebrospinal fluid compartments, the mean signal extracted from the entire brain, the derivatives of each of these nine parameters, and quadratic terms of each of the nine parameters and their derivatives. Both the BOLD-weighted time series and the artefactual model time series were temporally filtered using a first-order Butterworth filter with a passband between 0.01 and 0.08 Hz (18).

Measures of Connectivity

All functional connectivity networks were built using the residual timeseries (following de-noising). The functional connectivity between any pair of brain regions was operationalized as the Pearson correlation coefficient between the mean activation timeseries extracted from those regions. For each parcellation, an $n \times n$ weighted adjacency matrix encoding the connectome was thus obtained, where n represents the total number of nodes (or parcels) in that parcellation. Nodal strength was defined as the sum of Pearson correlation coefficients for each node (i.e. summing correlation coefficients by column in the adjacency matrix). We also constructed subject-level functional networks that included an *a priori* assignment of nodes to network communities (19,20). To obtain within-network and between-network strength measures, Pearson correlation coefficients of relevant connections were averaged.

We assessed six different combinations of parcellations and network measures. First, we evaluated the nodal strength of each of the 200 nodes described in the parcellation developed by Schaefer *et al.* (19); nodal strength was compared between subtypes. Second, we mapped the 200

nodes into the seven canonical systems defined by Yeo *et al*. (frontoparietal, ventral attention, visual, somatomotor, dorsal attention, limbic, and default mode) (20) and evaluated network strength within and between each network. Third, we also evaluated within- and between-network strength in the 17-network solution defined by Yeo *et al*. Next, these three analyses were also repeated using the higher resolution atlas of 400 nodes provided by Schaefer *et al.*

Statistical Testing

For all analyses, we used a general linear model to test how well subtypes predicted the outcome of interest (nodal or network strength), where subtype was modeled as a factor. As in the other analyses in the manuscript, we included mean in-scanner motion as an additional covariate to control for the potentially confounding effects of motion on image quality. An omnibus ANOVA testing for group differences was corrected for multiple comparisons by controlling the False Discovery Rate (FDR, $Q < 0.05$).

Sensitivity Analyses:

Sensitivity analyses excluding participants who were taking psychotropic medications at the time of clinical assessment were conducted to ensure that our results were not driven by medication effects. For the clinical data, 308 were excluded (final n=1116). The majority of the subjects excluded were in the DY group (n=217). Interestingly, 91 participants in the typically developing group were also on psychoactive substances. This fact reflects real world circumstances, where patients may be prescribed psychoactive substances in the absence of a formal psychiatric assessment and diagnosis. For the imaging analysis, 65 were excluded (final n=303). All methods were identical to whole group analyses as detailed in the main paper and this supplement.

References

1. Satterthwaite TD, Connolly JJ, Ruparel K, Calkins ME, Jackson C, Elliott MA, et al. The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth. Neuroimage. 2016;124:1115–1119.

2. Calkins ME, Merikangas KR, Moore TM, Burstein M, Behr MA, Satterthwaite TD, et al. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. Journal of Child Psychology and Psychiatry. 2015;56(12):1356–1369.

3. Moore TM, Reise SP, Gur RE, Hakonarson H, Gur RC. Psychometric properties of the Penn Computerized Neurocognitive Battery. Neuropsychology. 2015;29(2):235.

4. Satterthwaite TD, Wolf DH, Erus G, Ruparel K, Elliott MA, Gennatas ED, et al. Functional maturation of the executive system during adolescence. Journal of Neuroscience. 2013;33(41):16249–16261.

5. Shanmugan S, Wolf DH, Calkins ME, Moore TM, Ruparel K, Hopson RD, et al. Common and dissociable mechanisms of executive system dysfunction across psychiatric disorders in youth. Am J Psychiatry. 2016 01;173(5):517–26.

6. Snodgrass JG, Corwin J. Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General. 1988;117(1):34–50.

7. Shamosh NA, DeYoung CG, Green AE, Reis DL, Johnson MR, Conway ARA, et al. Individual differences in delay discounting: relation to intelligence, working memory, and anterior prefrontal cortex. Psychol Sci. 2008 Sep 1;19(9):904–11.

8. Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage. 2014 Oct 1;99:166–79.

9. Ciric R, Rosen AFG, Erus G, Cieslak M, Adebimpe A, Cook PA, et al. Mitigating head motion artifact in functional connectivity MRI. Nat Protoc. 2018 Dec;13(12):2801–26.

10. Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human Brain Mapping. 2005;25(1):46–59.

11. Rottschy C, Langner R, Dogan I, Reetz K, Laird AR, Schulz JB, et al. Modelling neural correlates of working memory: A coordinate-based meta-analysis. NeuroImage. 2012 Mar 1;60(1):830–46.

12. Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping. 2002 Jan;15(1):1–25.

13. Satterthwaite TD, Elliott MA, Ruparel K, Loughead J, Prabhakaran K, Calkins ME, et al. Neuroimaging of the Philadelphia Neurodevelopmental Cohort. Neuroimage. 2014;86:544–553.

14. Xia CH, Ma Z, Ciric R, Gu S, Betzel RF, Kaczkurkin AN, et al. Linked dimensions of psychopathology and connectivity in functional brain networks. Nature Communications. 2018 Aug 1;9(1):3003.

15. Ciric R, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage. 2017;154:174–187.

16. 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 1;17(2):825–41.

17. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. NeuroImage. 2009 Oct 15;48(1):63–72.

18. Hallquist MN, Hwang K, Luna B. The nuisance of nuisance regression: Spectral misspecification in a common approach to resting-state fMRI preprocessing reintroduces noise and obscures functional connectivity. NeuroImage. 2013 Nov 15;82:208–25.

19. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex. 2018 01;28(9):3095–114.

20. Thomas Yeo BT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol. 2011 Sep;106(3):1125–65.

Supplementary Tables and Figures

Supplementary Table 1: Effect sizes of accuracy domain in the CNB. Effect sizes are reported as Cohen's *d*. Effect sizes are inflated due to the use of cognitive data as features in the HYDRA clustering procedure. CNB – Computerized Neurocognitive Battery. HYDRA – Heterogeneity through Discriminative Analysis.

	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
ABF	-0.26	0.50	0.24	0.76	0.51	-0.26
ATT	-0.06	0.27	0.02	0.31	-0.08	-0.23
WM	-0.06	0.50	0.29	0.54	0.36	-0.21
VMEM	-0.34	0.08	0.35	0.43	0.71	0.24
FMEM	-0.18	0.44	-0.01	0.66	0.18	-0.47
SMEM	-0.39	0.48	0.14	0.94	0.59	-0.35
LAN	-0.35	0.74	0.34	1.12	0.76	-0.39
NVR	-0.70	0.73	0.62	1.71	1.59	-0.14
SPA	-0.44	0.55	0.45	1.07	0.98	-0.10
EID	-0.09	0.16	0.07	0.24	0.16	-0.10
EDI	-0.52	0.38	0.33	0.95	0.96	-0.06
ADI	-0.52	-0.13	0.37	0.42	0.92	0.50
Overall Accuracy	-0.65	0.76	0.54	1.58	1.49	-0.26

ABF-Abstraction/Mental Flexibility, ATT-Attention, WM-Working Memory, VMEM-Verbal Memory, FMEM-Face Memory, SMEM-Spatial Memory, LAN-Language/Verbal Reasoning, NVR-Nonverbal Reasoning, SPA-Spatial Reasoning, EID-Emotion Recognition, EDI-Emotion Discrimination, ADI-Age Discrimination.

Supplementary Table 2: Effect sizes of speed domain in the CNB. Effect sizes are reported as Cohen's *d*. Effect sizes are inflated due to the use of cognitive data as features in the HYDRA clustering procedure. CNB – Computerized Neurocognitive Battery. HYDRA – Heterogeneity through Discriminative Analysis.

	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
ABF	-0.16	0.58	-0.33	0.78	-0.21	-0.99
ATT	-0.09	0.36	-0.35	0.47	-0.29	-0.75
WM	-0.16	0.46	-0.25	0.61	-0.11	-0.68
VMEM	-0.10	0.58	-0.46	0.64	-0.43	-0.94
FMEM	-0.12	0.38	-0.64	0.56	-0.70	-1.18
SMEM	-0.17	0.46	-0.61	0.64	-0.58	-1.11
LAN	-0.12	0.69	-0.29	0.76	-0.22	-0.88
NVR	0.46	-0.18	-0.64	-0.64	-1.12	-0.62
SPA	0.03	0.36	-0.49	0.32	-0.57	-0.85
EID	-0.11	0.67	-0.55	0.79	-0.53	-1.25
EDI	0.07	0.48	-0.68	0.42	-0.85	-1.26
ADI	0.08	0.43	-0.60	0.36	-0.82	-1.18
MOT	-0.26	0.33	0.02	0.57	0.30	-0.31
SM	0.04	0.40	-0.53	0.36	-0.66	-0.93
Overall Speed	-0.11	0.76	-0.78	0.97	-0.93	-1.79

ABF-Abstraction/Mental Flexibility, ATT-Attention, WM-Working Memory, VMEM-Verbal Memory, FMEM-Face Memory, SMEM-Spatial Memory, LAN-Language/Verbal Reasoning, NVR-Nonverbal Reasoning, SPA-Spatial Reasoning, EID-Emotion Recognition, EDI-Emotion Discrimination, ADI-Age Discrimination, MOT-Motor, SM-Sensorimotor.

Supplementary Table 3: *Post hoc* analyses of clinical factor scores. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Anxious-Misery	<0.001	<0.001	<0.001	<0.001	0.030	0.375	0.724
Externalizing	<0.001	<0.001	<0.001	<0.001	0.701	0.776	0.207
Fear	<0.001	0.017	<0.001	0.041	<0.001	1	<0.001
Overall Psychopathology	<0.001	<0.001	<0.001	<0.001	0.313	0.957	0.671

Supplementary Table 4: Effect sizes of differences between subtypes in clinical factor scores. Effect sizes are reported as Cohen's *d*.

	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Anxious-Misery	-0.92	-0.68	-0.78	0.24	0.13	-0.10
Externalizing	-0.54	-0.66	-0.46	-0.09	0.08	0.17
Fear	-0.22	-0.64	-0.22	-0.39	0.01	0.40
Overall Psychopathology	-1.60	-1.77	-1.66	-0.16	-0.05	0.11

Supplementary Table 5: *Post hoc* analyses of clinical factor scores determined by 112 items. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Anxious-Misery	<0.001	<0.001	<0.001	<0.001	0.001	0.110	0.427
Psychosis	0.050	0.646	0.019	0.924	0.441	0.981	0.281
Externalizing	<0.001	<0.001	<0.001	<0.001	0.847	0.951	0.570
Fear	<0.001	0.008	<0.001	0.078	<0.001	0.970	<0.001
Overall Psychopathology	<0.001	<0.001	<0.001	<0.001	0.066	0.811	0.455

Supplementary Table 6: *Post hoc* analyses of state and trait anxiety. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
State Anxiety	0.001	0.025	0.016	0.079	0.992	1	0.986
Trait Anxiety	<0.001	<0.001	<0.001	<0.001	0.971	0.975	1

Supplementary Table 7: Effect sizes of differences between subtypes in STAI scores. Effect sizes are reported as Cohen's *d*. STAI - State-Trait Anxiety Inventory.

	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
State Anxiety	-0.41	-0.46	-0.40	-0.05	0.02	0.06
Trait Anxiety	-0.81	-0.75	-0.78	0.07	0.07	-0.002

Supplementary Table 8: *Post hoc* analyses of *n*-back activation. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Right Crus II	0.043	0.490	0.195	0.215	0.031	0.036	1.000
Left Anterior DLPFC	0.043	0.633	0.070	0.213	0.017	0.056	0.989
Dorsal Anterior Cingulate	0.050	0.892	0.044	0.359	0.031	0.219	0.895
Left Dorsal Frontal	0.043	0.346	0.225	0.281	0.022	0.031	1.000
Left Precuneus	0.043	0.998	0.009	0.472	0.028	0.523	0.581
Right Precuneus	0.043	0.978	0.032	0.051	0.045	0.062	1.000

Supplementary Table 9: *Post hoc* sensitivity analyses of clinical factor scores. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Anxious-Misery	<0.001	<0.001	<0.001	<0.001	0.097	0.895	0.440
Externalizing	<0.001	<0.001	<0.001	<0.001	0.409	0.888	0.120
Fear	<0.001	0.003	<0.001	0.179	0.045	0.764	0.003
Overall Psychopathology	<0.001	<0.001	<0.001	<0.001	0.229	0.954	0.568

Supplementary Table 10: Effect sizes in sensitivity analyses in clinical factor scores. Effect sizes are reported as Cohen's *d*.

	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Anxious-Misery	-0.85	-0.62	-0.79	0.24	0.08	-0.17
Externalizing	-0.47	-0.65	-0.40	-0.15	0.07	0.23
Fear	-0.31	-0.58	-0.20	-0.26	0.11	0.37
Overall Psychopathology	-1.54	-1.78	-1.62	-0.21	-0.06	0.15

Supplementary Table 11: *Post hoc* sensitivity analyses of state and trait anxiety. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
State Anxiety	0.001	0.031	0.029	0.058	1	1	1
Trait Anxiety	<0.001	<0.001	<0.001	<0.001	0.989	0.992	1

Supplementary Table 12: Effect sizes in sensitivity analyses state and trait anxiety scores. Effect sizes are reported as Cohen's *d*.

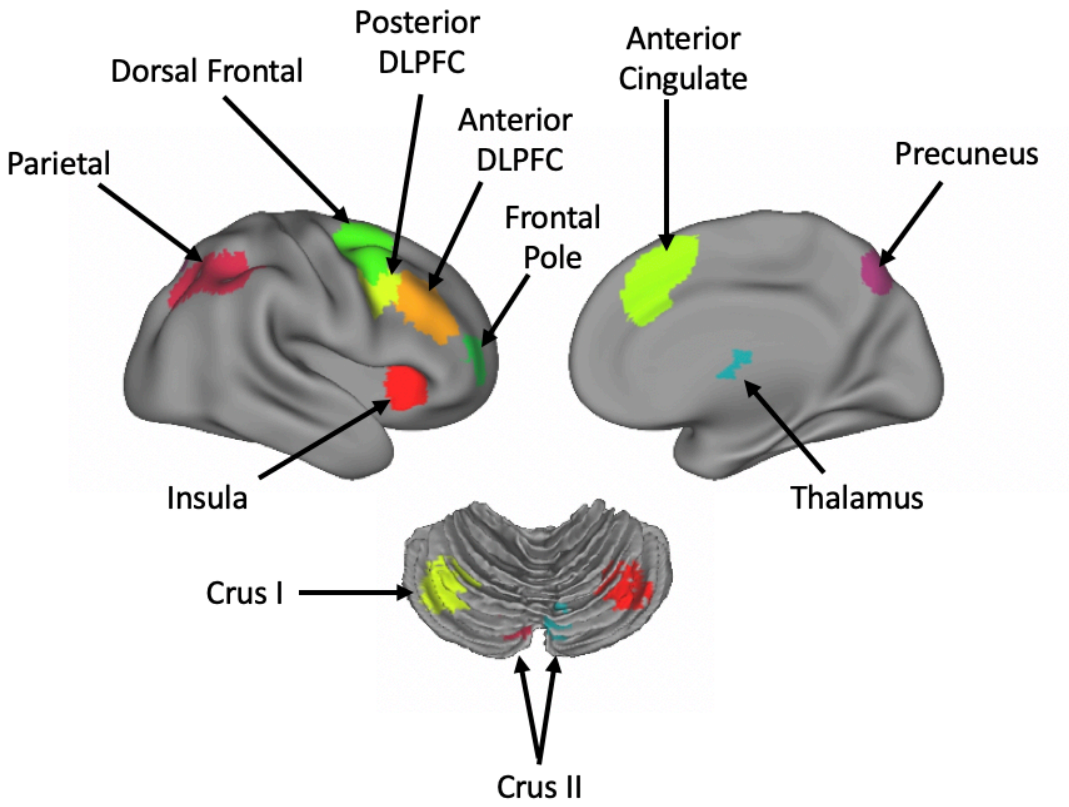
	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
State Anxiety	-0.47	-0.46	-0.47	-0.01	0.003	-0.004
Trait Anxiety	-0.73	-0.84	-0.87	-0.06	-0.06	0.003

Supplementary Table 13: *Post hoc* sensitivity analyses of *n*-back activation. Group statistics were corrected for multiple comparisons by controlling the False Discovery Rate ($Q < 0.05$). Pairwise contrasts are reported as *p*-values, and were adjusted via the Tukey method.

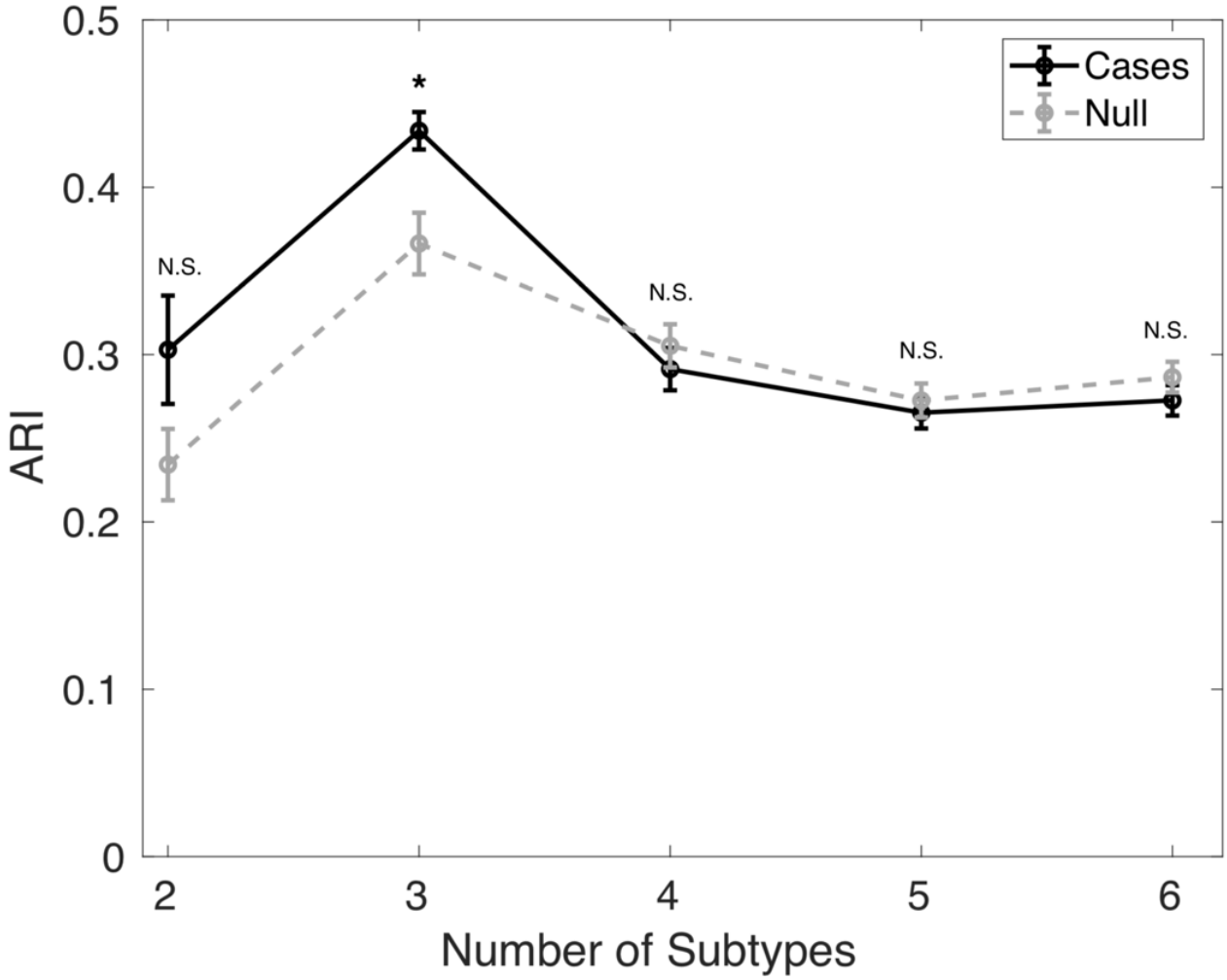
	Pr(>F)	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Right Crus I	0.046	0.893	0.053	0.222	0.047	0.152	0.988
Right Crus II	0.046	0.982	0.204	0.030	0.246	0.052	0.845
Left Anterior DLPFC	0.046	0.945	0.093	0.061	0.101	0.065	0.988
Dorsal Anterior Cingulate	0.0496	1	0.069	0.117	0.178	0.225	1
Left Dorsal Frontal	0.046	0.751	0.369	0.053	0.161	0.024	0.806
Left Parietal	0.046	0.882	0.164	0.035	0.115	0.029	0.902
Left Precuneus	0.046	0.998	0.012	0.178	0.095	0.407	0.925
Right Precuneus	0.0487	1	0.068	0.085	0.216	0.219	0.999

Supplementary Table 14: Effect sizes in sensitivity analyses in *n*-back analysis. Effect sizes are reported as Cohen's *d*.

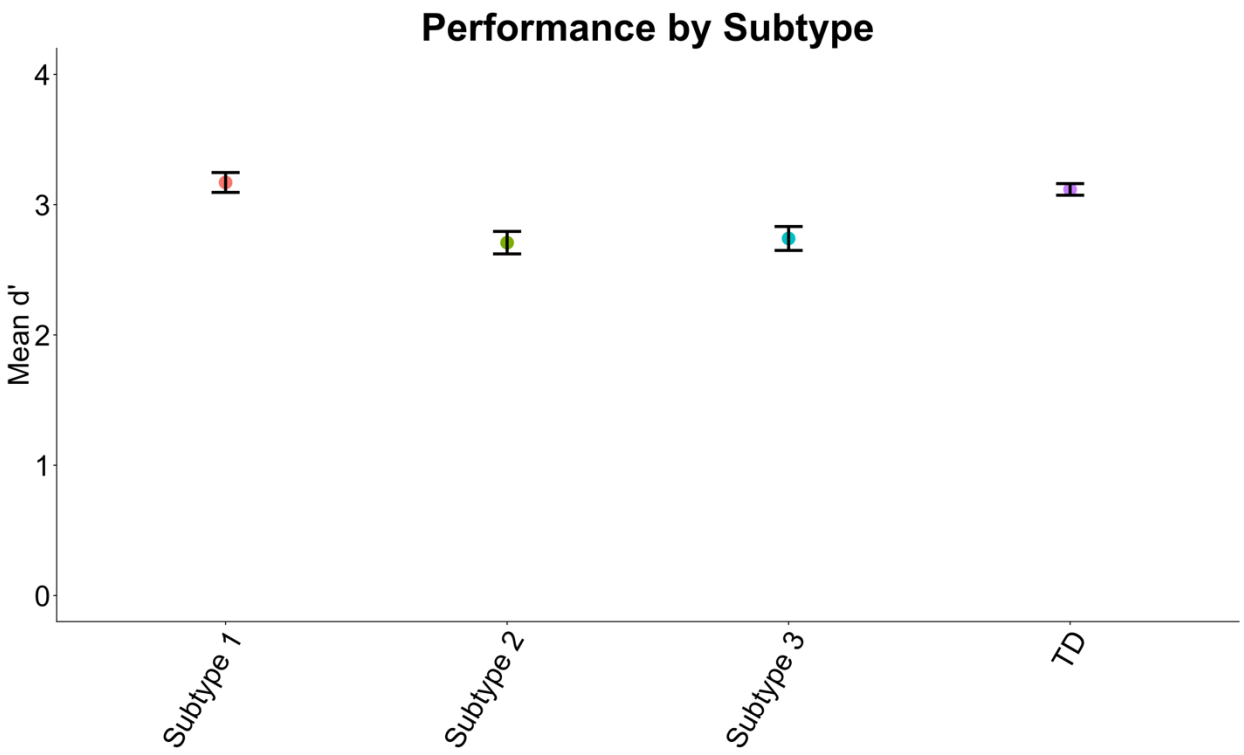
	TD vs. Subtype 1	TD vs. Subtype 2	TD vs. Subtype 3	Subtype 1 vs. Subtype 2	Subtype 1 vs. Subtype 3	Subtype 2 vs. Subtype 3
Right Crus I	-0.14	0.42	0.34	0.66	0.55	-0.10
Right Crus II	-0.07	0.33	0.49	0.43	0.58	0.21
Left Anterior DLPFC	-0.12	0.38	0.45	0.53	0.64	0.67
Dorsal Anterior Cingulate	-0.05	0.41	0.42	0.45	0.49	-0.006
Left Dorsal Frontal	-0.20	0.28	0.48	0.45	0.69	0.18
Left Parietal	-0.15	0.35	0.50	0.50	0.69	0.15
Left Precuneus	-0.01	0.51	0.37	0.54	0.40	-0.17
Right Precuneus	-0.02	0.41	0.44	0.42	0.48	0.02



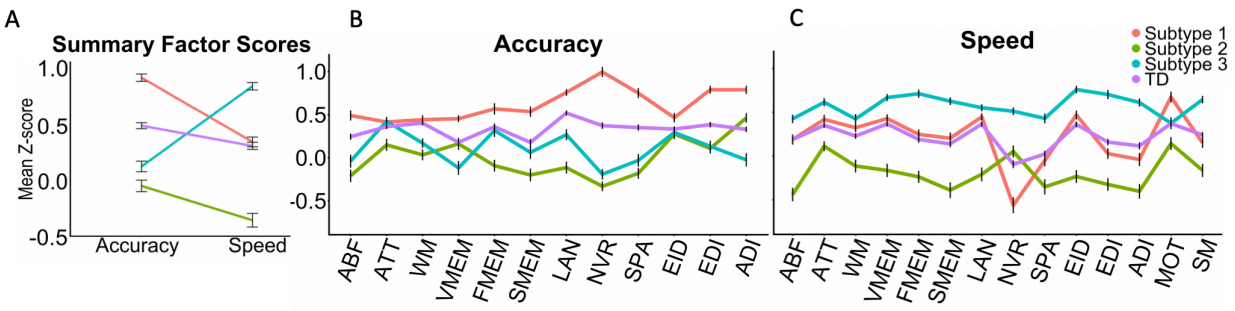
Supplementary Figure 1: Twenty-one functionally defined regions of interest. As described in previous work, the executive network was parsed into 21 functional regions of interest by applying a watershed algorithm to the map of the 2-back > 0-back contrast using an initial threshold of $z > 20$ (4).



Supplementary Figure 2: Permutation testing results. Statistical significance of ARI for the number of clusters (subtypes) via comparison with the null distribution. Note * represents $P_{fdr} = 0.011$ and N.S. represents Not Significant ($P_{fdr} = 0.192$).



Supplementary Figure 3: D' performance analysis during n-back. To relate task performance to the neuroimaging data, task performance was summarized using the signal detection measure *d'* (6,7). This measure considers both correct responses and false positives to limit the influence of response bias. Between subtype differences in d' mapped on to neuroimaging results, with Subtype 1 having the highest d' score, followed by TDs, then Subtype 3 and lastly Subtype 2. Error bars represent standard error of the mean.



Supplementary Figure 4: Neurocognitive profiles in sensitivity analysis (no medications). Subtypes revealed by HYDRA differ in neurocognitive profile. (**A**) Three neurocognitive signatures emerged in depressed youth: Subtype 1 had preserved cognition, with high accuracy and speed; Subtype 2 had impaired cognition, with low accuracy and speed; Subtype 3 was impulsive, with high speed but low accuracy. (**B-C**) Patterns were largely consistent for all measures of accuracy (panel **B**) and speed (panel **C**).

HYDRA-Heterogeneity through Discriminative Analysis, ABF-Abstraction/Mental Flexibility, ATT-Attention, WM-Working Memory, VMEM-Verbal Memory, FMEM-Face Memory, SMEM-Spatial Memory, LAN-Language/Verbal Reasoning, NVR-Nonverbal Reasoning, SPA-Spatial Reasoning, EID-Emotion Recognition, EDI-Emotion Discrimination, ADI-Age Discrimination, MOT-Motor, SM-Sensorimotor