Supplementary Methods

Inclusion/Exclusion Criteria

Infants were enrolled in the NIH-funded, Autism Center of Excellence Infant Brain Imaging Study (IBIS) via one of four clinical sites (University of Washington, University of North Carolina, Washington University in St. Louis, and The Children's Hospital of Philadelphia). High risk (HR) infants had at least one older sibling with a clinical diagnosis of ASD (confirmed by medical records); Low risk (LR) infants had an older sibling with typical development and no first or second-degree relatives with ASD, intellectual disability, or neurological or psychiatric disorders. Infants were excluded from enrolling for any of the following: known genetic conditions associated with ASD, medical conditions affecting growth, development or cognition, sensory impairments, birth weight < 2,000 grams, gestational age < 36 weeks, history of perinatal adversity or in-utero exposure to neurotoxins, contraindication for MRI, predominant home language other than English, or having been adopted. Participants (N = 432) were included in the current analysis if they had data available on the questionnaire used to measure sleep (the Infant Behavior Questionnaire-Revised), diagnostic data at 24 months, and met quality control standards for imaging data from at least one timepoint (6, 12, 24 months; see Table 1 in main text). Three participants scored >3 SD above the mean average ISOP score (2 HR-NonASD, 1 HR-ASD). Inclusion/exclusion of these three outlying participants did not alter the pattern of observed findings.

Behavioral assessment

HR and LR infants were assessed at 6, 12 and 24 months with MRI scans, a parent-report measure of adaptive skills (the Vineland Scales II) (1) and standardized assessment of cognitive

ability (the Mullen Scales of Early Learning) (2). At 24 months, diagnostic classification according to the DSM-IV-TR (3) criteria (Autistic Disorder or Pervasive Developmental Disorder, Not Otherwise Specified) was assigned by an expert clinician at each site, and confirmed by a second, independent clinician, using all clinical, behavioral, and questionnaire data available, including the Autism Diagnostic Interview-Revised (ADI-R) (4) and the Autism Diagnostic Observation Schedule (ADOS) (5). Approximately half of the infants (n = 226) participated in a second diagnostic visit at 36 months. Infants who met DSM-IV-TR criteria for ASD or PDD-NOS at either 24 or 36 months were included in the ASD group (additional details reported in Table S1). This yielded three outcome groups: (1) high risk infants who met criteria for ASD (HR-ASD = 71) (2) high risk infants who did not meet criteria for ASD (HR-NonASD = 234); and (3) low risk infants who did not meet criteria for ASD (LR = 127). Three LR infants meeting criteria for ASD and were excluded from the analysis because they were too few in number to constitute a comparison group.

Behavioral measures

Autism symptoms: The ADOS (6) is a semi-structured, standardized measure of social relatedness, communication, play, and repetitive behaviors administered by an examiner trained to research standards. Calibrated severity scores (7) for social affect and restricted/repetitive behaviors were used for these analyses.

Cognitive ability: The Mullen Scales of Early Learning (Mullen, 1995) were used to measure child cognitive ability. The Mullen is a standardized developmental test for children ages birth to 68 months. Standard scores on the Early Learning Composite (normed to M 100; SD 15), which

combines scores across 4 subscales (Fine Motor, Visual Reception, Expressive Language, and Receptive language), and standardized subscale scores were utilized for these analyses.

Adaptive skills: The Vineland Scales of Adaptive Behavior II (Vineland; 1; 8) are a parent interview assessing social, communication, motor, and daily living skills. Standardized subscale scores (Socialization, Motor Skills, Communication, Daily Living Skills), were utilized for these analyses.

MRI Acquisition & Processing

MRI Acquisition

Pediatric imaging was completed during natural sleep at each clinical site using identical 3-T Siemens TIM Trio scanners (Siemens Medical Solutions, Malvern, PA) equipped with 12-channel head coils. The imaging protocol included 1) a localizer scan, 2) 3D T1 MPRAGE: TR = 2400ms, TE=3.16ms, 160 sagittal slices, FOV=256, voxel size = $1mm^3$, 3) 3D T2 FSE TR = 3200ms, TE=499ms, 160 sagittal slices, FOV=256, voxel size = $1mm^3$, and 4) a 25 direction DTI: TR = 12800ms, TE=102ms, slice thickness = 2mm isotropic, variable *b* value = maximum of $1000s/mm^2$, FOV = 190.

A number of quality control procedures were employed to assess scanner stability and reliability across sites, time, and procedures. Geometric phantoms were scanned monthly and human phantoms (two adult subjects) were scanned annually to monitor scanner stability at each site across the study period. Details on the stability procedures for IBIS and scanner quality control checks are described elsewhere (9).

Image Preprocessing

Raw T1 and T2 images were visually inspected by a single expert rater at the Data Coordinating Center for overall image quality, including motion and other image artifacts. Images that were deemed of poor quality (having heavy artifacts in several slices, or moderate artifacts in many slices) were marked as failing quality control and were not included in any downstream image processing. Images with minimal artifacts (ranging from pristine images to those containing moderate or minor artifacts in a few to several slices) passed this initial visual quality inspection and proceeded to downstream image processing pipelines. The subcortical segmentation pipeline required that both the T1 and T2 image were of usable quality. A total of 99% (T1) and 98% (T2) of acquired scans passed initial visual quality control step and were eligible for processing to generate subcortical structural segmentations. There were no differences in the proportion of scans that passed quality control by outcome group (i.e., HR-ASD, HR-NonASD, or LR; T1 $X^2(2) = 1.7$, $p = .4$; T2 $X^2(2) = 1.4$, $p = .5$.)

T1- and T2-weighted images that passed visual quality control underwent distortion correction, mutual registration, transformation to stereotactic space, and CSF/brain tissue segmentation. Specifically, all images were corrected for geometric distortions (10) and intensity non-uniformity (11). T2-weighted images underwent linear, rigid registration to the corresponding T1-weighted images via mutual information registration (12). Subsequently, both T1- and T2-weighted images were transformed to stereotactic space based on the registration of the T1 scan. The skull was extracted using a "majority voting approach" between the T1 atlas mask, T2 atlas mask, and the T1 and T2 images jointly via FSL Brain Extraction Tool (13). All corrected and skull-stripped T1 and T2 images were used as input for an expectation,

maximization-based, tissue segmentation tool (AutoSeg pipeline) (14) to obtain white matter, gray matter and CSF (15).

Segmentation of Subcortical Brain Structures

A graph-based, multi-atlas method developed by investigators in the IBIS Network was employed to segment the subcortical structures (14). A brief summary is presented here, see Wang et al. (14) for complete details. First, all atlases and participant MR images were paired and co-registered via symmetric diffeomorphic registration using the ANTS (Advanced Normalization ToolS) registration tool (16). Second, a directed graph with edge weights based on intensity and shape similarity was constructed between all atlases and the participant MR image (14). Third, the shortest path from each atlas to the participant image was computed (with atlases sharing the same shortest paths combined into the same cluster), and the atlas closest to the participant for each cluster was selected as the neighboring template (14). Finally, the final segmentation was produced by fusing the propagated label files of the neighboring templates via weighted majority voting (14). The caudate segmentations were additionally refined by using the lateral ventricles as a mask (17).

Atlas templates were derived from 16 cases at each time point (6, 12, and 24 months), which were manually segmented by a single anatomical expert, used as training images in the multi atlas segmentation, and then applied to all 6 month and 12-24 month data sets. The multi-atlas segmentation method was validated in a leave-one-out validation analysis that achieved high Dice coefficients for all structures (mean=91.47%, SD=.03, range=87.20-96.00%). All subcortical segmentations underwent visual quality inspection by one of two trained experimenters (blind to diagnosis, risk status, sex, scan site), and over 98% of scans met quality inspection criteria for

inclusion in subsequent analyses. There was no difference in the proportion of scans that met quality inspection criteria for subcortical segmentation by outcome group (i.e., HR-ASD, HR-NonASD, or LR; $X^2(2) = 1.4$, $p = .5$). Combining high-quality segmentation data with scores from requisite behavioral measures (IBQ-R) yielded a total of 328 6-month, 317 12-month, and 287 24-month scans for the current analysis (see Table 1 in main text). The complete Autoseg software pipeline for multi-atlas-based segmentation is publicly available on the NIH NITRC website (Neuroimaging Informatics Tools and Resources Clearinghouse) at http://www.nitrc.org/projects/autoseg.

Statistical Analysis

In the first analysis phase, ISOP score was examined in relation to trajectories of subcortical volumes from 6-24 months of age. Linear mixed-effect models were used to predict bilateral volumes for each subcortical structure (hippocampus, amygdala, caudate, globus pallidus, putamen, thalamus). This analytic method is suitable for an unbalanced design and allows for missing values in a longitudinal study. All tests were two-tailed with $\alpha = 0.05$, correcting for multiple comparisons (across six p-values, as six structures were tested) using the Benjamini & Hochberg (18) method for false discovery rate correction. Volumes were summed across hemispheres consistent with previous work showing no laterality effects in subcortical volumes in this sample (19). In each model predicting bilateral subcortical volume, individual intercepts were included as a random effect. Diagnostic group, ISOP score, age, quadratic effect of age (age^2), sex, and group interactions with each were included as fixed effects. Quadratic effect of age was included in all models as it improved model fit for each subcortical area. Total cerebral volume was included as a covariate given its relationship to subcortical volumes and to

control for possible group differences in brain size. Scan site was also included as a control variable.

In the second analysis phase, any subcortical structure that showed a significant association with ISOP was subjected to additional follow-up analyses to determine the strength and specificity of the finding. First, laterality effects were examined to determine if the relationship held across hemispheres. Second, models were repeated controlling for cognitive ability at 24 months, to determine whether the relationship between sleep and subcortical volume existed independent of cognitive functioning level. This step was important to include as sleep in infancy has been associated with later cognitive functioning (20). Third, subcortical structure volume was tested for relationships with other subscales of the IBQ-R to determine whether the relationship was unique to the sleep items, or to infant temperament more broadly. Fourth, significant sleep-subcortical relationships were tested at individual timepoints in development, using ISOP scores at either 6 months or 12 months, and controlling for cognitive ability at that same timepoint. All analyses were performed in R. Analysis code is publicly available at http://faculty.washington.edu/kmacd/sleep_subcortical.html.

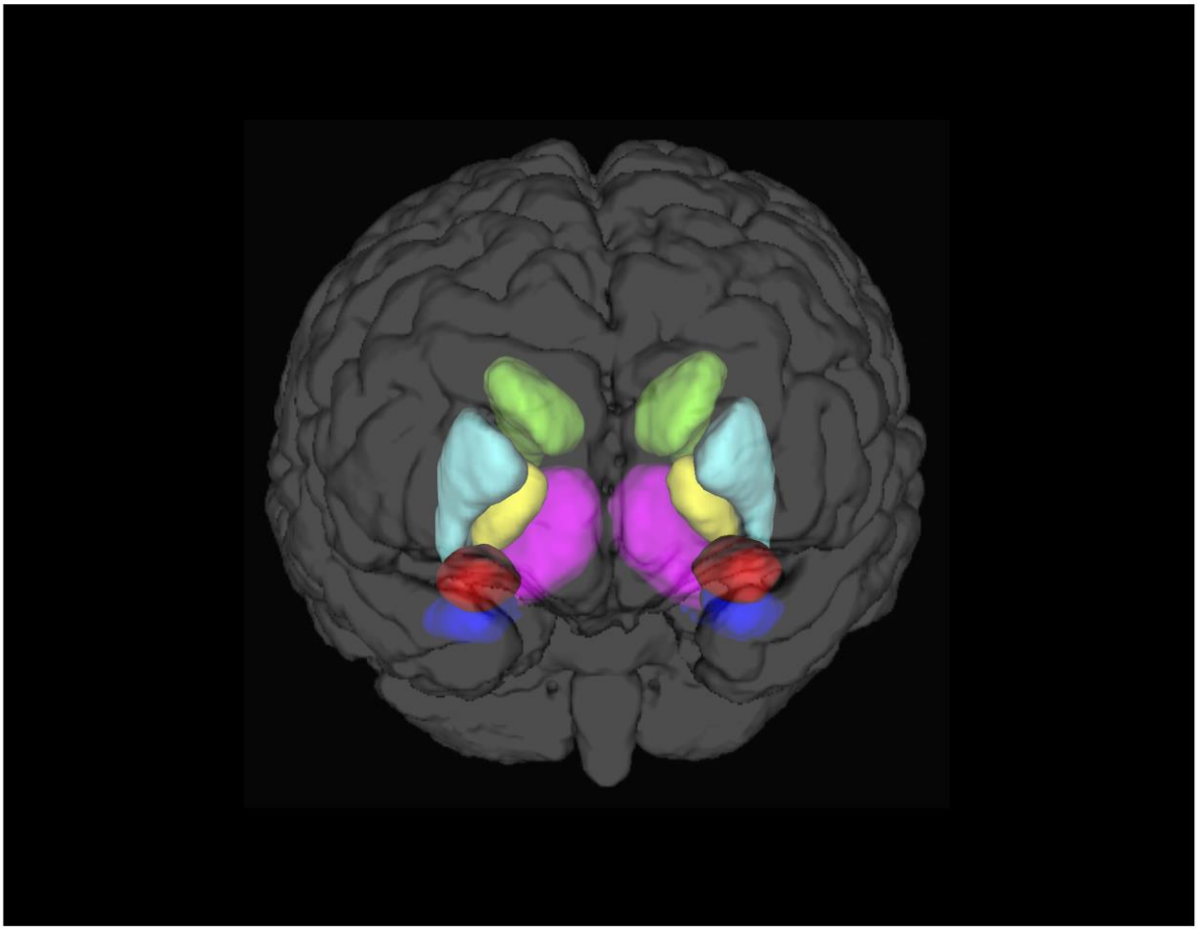

Figure S1. Subcortical segmentation. Green = caudate, Red = amygdala, Dark blue = hippocampus, Pink = thalamus, Yellow = globus pallidus, Light blue = putamen.

	Met criteria at 24 and 36 mo	Met criteria at 24 mo only (not assessed at 36 mo)	Met criteria at 36 but not 24 mo	Met criteria at 24 but not 36 mo	Group comparison
Total N	37	16	8	10	--
Sex	31 M, 6 F	15 M, 1 F	6 M, 2 F	7 M, 3 F	$X^2(3) = 2.9$, $p = .4$
ADOS Restricted/ Repetitive 24 mo	6.4 (2.7)	6.7 (2.1)	2.6 (2.3)	6.2 (1.0)	$F(3,67) = 6.54$, $p = .0006$
ADOS Social Affect 24 mo	6.2 (2.0)	6.3 (1.4)	2.4 (1.7)	5.3 (1.4)	$F(3,67) = 10.61$, $p < .0001$
Mullen ELC 24 mo	79.1 (16.2)	77.1 (17.9)	94.6 (8.6)	87.4 (16.1)	$F(3,65) = 2.9$, $p = .04$

Table S1. Autism symptoms and intellectual ability scores in subgroups who were diagnosed with ASD at either 24 or 36 months or both.

Associations between ISOP score and behavior

	ISOP 6 mo & 6 mo behavior $N \sim 352$	ISOP 12 mo & 12 mo behavior $N \sim 384$	ISOP avg (6-12 mo) & 24 mo behavior $N \sim 422$
ADOS (24 mo only)			
Social Affect	--	--	.14 **
Restricted/Repetitive	--	--	.07
Mullen			
Visual reception	-.03	.07	-.08
Fine motor	.04	.07	-.03
Gross motor	.06	.07	-.08
Receptive language	-.05	-.05	-.08
Expressive language	-.14 *	-.11 *	-.11 *
Vineland			
Communication	-.05	-.03	-.13 *
Daily living skills	.03	.01	-.08
Motor skills	.06	.04	-.06
Socialization	-.02	-.08	-0.15 **

Table S2. Pearson correlation coefficients are presented for associations between ISOP score and behavioral assessment measures. The left column shows correlations between 6-mo ISOP scores and 6-mo behavioral scores. The center column shows correlations between 12-mo ISOP scores and 12-mo behavioral scores. The right column shows correlations between average ISOP score (from 6-12 mo) and behavioral scores at 24 months.
* $p < .05$, ** $p < .01$.

Table S3. Full model results for all six subcortical structures. Indicated significance levels have been corrected for multiple comparisons ($^{+}$ p < .10, * p < .05, ** p < .01, *** p < .001).

	HIPPOCAMPUS			AMYGDALA			CAUDATE		
	Estimate	Std. Error	t	Estimate	Std. Error	t		Std. Error	t
ISOP	100.9	31.4	3.2**	22.7	11.6	2.0	-2.0	52.6	0.0
HR-NonASD:ISOP	-109.0	37.0	-2.9*	-27.3	13.6	-2.0	-10.5	61.9	-0.2
LR:ISOP	-108.7	43.2	-2.5^{+}	-13.3	15.9	-0.8	29.7	72.0	0.4
HR-NonASD	537.8	169.0	3.2**	265.8	68.1	3.9**	-153.0	276.1	-0.6
LR	549.4	182.9	3.0*	178.3	73.6	2.4+	-149.6	298.4	-0.5
age	-14.2	12.4	-1.1	40.7	6.1	6.7***	83.6	19.4	4.3***
age.2	0.9	0.3	2.5*	-1.2	0.2	-6.7***	-2.4	0.5	-4.4***
Male	178.9	93.4	1.9	61.2	34.6	1.8	-198.7	153.0	-1.3
Total cerebral volume	498.0	24.0	20.8***	235.2	9.6	24.4***	775.6	38.3	20.2***
siteSEA	-13.7	40.2	-0.3	-5.2	14.8	-0.4	-104.2	66.9	-1.6
siteSTL	-58.4	39.3	-1.5	6.6	14.4	0.5	-8.0	65.1	-0.1
siteUNC	1.8	39.8	0.0	-2.2	14.6	-0.1	-102.0	66.6	-1.5
HR-NonASD:age	-7.3	12.0	-0.6	-21.7	6.2	-3.5**	12.0	18.5	0.6
LR:age	-14.0	13.0	-1.1	-14.5	6.7	-2.2	8.5	19.9	0.4
HR-NonASD:age.2	0.1	0.4	0.3	0.6	0.2	3.3**	-0.1	0.6	-0.3
LR:age.2	0.3	0.4	0.7	0.4	0.2	2.0	0.0	0.6	0.0
HR-NonASD:Male	-209.6	99.8	-2.1	-50.7	36.8	-1.4	90.2	163.9	0.6
LR:Male	-108.0	104.9	-1.0	-49.9	38.7	-1.3	-19.4	172.4	-0.1
	GLOBUS PALLIDUS			PUTAMEN			THALAMUS		
	Estimate	Std. Error	t	Estimate	Std. Error	t	Estimate	Std. Error	t
ISOP	3.0	13.8	0.2	10.8	58.3	0.2	38.6	58.4	0.7
HR-NonASD:ISOP	2.8	16.3	0.2	4.4	68.7	0.1	-71.7	68.7	-1.0
LR:ISOP	-3.2	19.0	-0.2	21.7	80.4	0.3	-35.7	80.4	-0.4
HR-NonASD	46.9	73.9	0.6	65.1	295.6	0.2	248.4	310.4	0.8
LR	43.3	80.0	0.5	182.8	320.5	0.6	147.3	336.0	0.4
age	58.9	5.3	11.0***	139.0	16.7	8.3***	-121.0	21.9	-5.5***
age.2	-1.6	0.1	-11.0***	-3.3	0.4	-7.5***	3.1	0.6	5.2***
Male	23.8	41.1	0.6	-31.5	172.5	-0.2	342.9	173.4	2.0
Total cerebral volume	210.1	10.4	20.1***	832.9	37.5	22.2***	1052.2	43.6	24.1***
siteSEA	-24.2	17.7	-1.4	-188.2	74.8	-2.5	-21.4	74.7	-0.3
siteSTL	-4.5	17.3	-0.3	-30.5	73.1	-0.4	0.1	73.0	0.0
siteUNC	-17.8	17.5	-1.0	-117.7	74.2	-1.6	33.9	74.1	0.5
HR-NonASD:age	-5.9	5.1	-1.1	-15.7	15.0	-1.0	1.9	20.9	0.1
LR:age	-1.4	5.6	-0.2	-18.8	16.3	-1.2	6.4	22.7	0.3
HR-NonASD:age.2	0.2	0.2	1.4	0.6	0.5	1.3	0.1	0.6	0.2
LR:age.2	0.1	0.2	0.7	0.7	0.5	1.4	0.0	0.7	0.0
HR-NonASD:Male	3.3	43.9	0.1	184.8	185.2	1.0	-239.3	185.5	-1.3
LR:Male	-24.6	46.2	-0.5	40.6	194.9	0.2	-209.2	195.0	-1.1

Hippocampal volume predicted by IBQ subscale scores

	Estimate	Std. Error	t	sig level
Activity level (15 items)	-13.4	48.3	-0.3	*n.s*
Approach (12 items)	-47.2	49.0	-1.0	*n.s*
Cuddliness (17 items)	-22.7	45.1	-0.5	*n.s*
Distress to Limitations (16 items)	45.5	46.0	1.0	*n.s*
Duration of Orienting (12 items)	46.2	42.5	1.1	*n.s*
Falling Reactivity/Rate of Recovery from Distress (13 items)*	-84.3	39.5	-2.1	*n.s*
Fear (16 items)	-5.8	46.5	-0.1	*n.s*
High Pleasure (11 items)	-112.2	60.1	-1.9	*n.s*
Low Pleasure (13 items)	-119.0	42.9	-2.8	*n.s*
Perceptual Sensitivity (12 items)	27.1	34.7	0.8	*n.s*
Sadness (14 items)	35.5	43.7	0.8	*n.s*
Smiling and Laughter (10 items)	-31.7	36.5	-0.9	*n.s*
Soothability (18 items)	-56.0	48.4	-1.2	*n.s*
Vocal Reactivity (12 items)	-27.5	39.1	-0.7	*n.s*

Table S4. Subset of results relating IBQ-R subscale scores to hippocampal volume trajectories. For definitions of IBQ-R subscales, see Garstein & Rothbart, 2003.
n.s. = not significant.

References

1. Sparrow SS, Cicchetti DV, Balla DA (2005): Vineland adaptive behavior scales:(Vineland II), survey interview form/caregiver rating form. *Livonia, MN: Pearson Assessments*.
2. Mullen EM (1995): Mullen scales of early learning.
3. American Psychiatric Association (2000): *Diagnostic and statistical manual of mental disorders*, 4 ed. Washington, DC: Author.
4. Rutter M, Le Couteur A, Lord C (2003): Autism diagnostic interview-revised. *Los Angeles, CA: Western Psychological Services*. 29: 30.
5. Lord C, Rutter M, DiLavore PC, Risi S (1999): Autism diagnostic observation schedule-WPS (ADOS-WPS). *Los Angeles, CA: Western Psychological Services*.
6. Lord C, Rutter M, DiLavore PC, Risi S, Gotham K, Bishop S (2012): Autism diagnostic observation schedule: ADOS-2.
7. Hus V, Gotham K, Lord C (2014): Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. *J Autism Dev Disord*. 44: 2400–2412.
8. Elliott CD, Murray GJ, Pearson LS (1990): Differential ability scales. *San Antonio, Texas*.
9. Gouttard S, Styner M, Prastawa M, Piven J, Gerig G (2008): Assessment of reliability of multi-site neuroimaging via traveling phantom study. *Med Image Comput Comput Assist Interv*. 11: 263–270.
10. Fonov VS, Janke A, Caramanos Z, Arnold DL, Narayanan S, Pike GB, Collins DL (2010): Improved Precision in the Measurement of Longitudinal Global and Regional Volumetric Changes via a Novel MRI Gradient Distortion Characterization and Correction Technique. In:. *Medical Imaging and Augmented Reality*, Lecture Notes in Computer Science. (Vol. 6326), Berlin, Heidelberg: Springer, Berlin, Heidelberg, pp 324–333.
11. Sled JG, Zijdenbos AP, Evans AC (1998): A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging*. 17: 87–97.
12. Collins DL, Neelin P, Peters TM, Evans AC (1994): Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr*. 18: 192–205.
13. Smith SM (2002): Fast robust automated brain extraction. *Hum Brain Mapp*. 17: 143–155.
14. Wang J, Vachet C, Rumple A, Gouttard S, Ouziel C, Perrot E, *et al.* (2014): Multi-atlas segmentation of subcortical brain structures via the AutoSeg software pipeline. *Front Neuroinform*. 8: 7.
15. Gouttard S, Styner M, Joshi S, Smith RG, Hazlett HC, Gerig G (2007): Subcortical structure segmentation using probabilistic atlas priors. In: Pluim JPW, Reinhardt JM, editors. *Medical Imaging 2007: Image Processing*, SPIE Proceedings. (Vol. 6512), International Society for Optics and Photonics, p 65122J.
16. Avants B, Epstein C, Grossman M, Gee J (2008): Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*. 12: 26–41.
17. Shen MD, Kim SH, McKinstry RC, Gu H, Hazlett HC, Nordahl CW, *et al.* (2017): Increased Extra-axial Cerebrospinal Fluid in High-Risk Infants Who Later Develop Autism. *Biol Psychiatry*. 82: 186–193.
18. Benjamini Y, Hochberg Y (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 57: 289–300.

19. Swanson MR, Shen MD, Wolff JJ, Elison JT, Emerson RW, Styner MA, *et al.* (2017): Subcortical Brain and Behavior Phenotypes Differentiate Infants With Autism Versus Language Delay. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2: 664–672.
20. Pisch M, Wiesemann F, Karmiloff-Smith A (2019): Infant wake after sleep onset serves as a marker for different trajectories in cognitive development. *Journal of Child Psychology and Psychiatry,* 1st ed. 60: 189–198.