Atypical Brain Asymmetry in Autism – A Candidate for Clinically Meaningful Stratification

Supplementary Information

Participants, study design and exclusion criteria

All participants with autism had an existing clinical diagnosis of autism according to DSM-IV (1), DSM-IV-TR (2), DSM-5 (3) or ICD-10 (4) criteria. Participants underwent comprehensive clinical, cognitive and MRI assessment at one of six collaborating sites: the Institute of Psychiatry, Psychology and Neuroscience, King's College London (KCL), London, United Kingdom; Autism Research Centre at the University of Cambridge, Cambridge, United Kingdom; Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands; University Medical Centre Utrecht, Utrecht, the Netherlands; Central Institute of Mental Health, Mannheim, Germany; and University Campus Bio-Medico, Rome, Italy. For a distribution of participants by diagnostic group and sex across the sites, see Figure S1. Exclusion criteria included the presence of any MRI contraindications (e.g., metal implants, braces, claustrophobia) or failure to give informed written consent to MRI scanning, as well as significant hearing or visual impairments not corrected by glasses or hearing aids. In addition, we excluded participants with missing T1-weighted MRI scans, low image quality (i.e., structural brain abnormality, excessive head motion, insufficient coverage), and failed image preprocessing. Due to low number of participants from the Rome site after quality control, we restricted analyses to the five remaining sites. Neurotypical controls scoring positively for attention-deficit hyper-activity disorder (ADHD) as assessed by the DSM-5 ADHD rating scale were also excluded. The study was approved by the local ethical committees of the participating centers and written informed consent was obtained from all participants or their legal guardians.

Clinical, cognitive and demographic measures

General intellectual abilities were assessed using the Wechsler Abbreviated Scales of Intelligence-Second Edition (WASI-II (5)), or if unavailable the Wechsler Intelligence Scale for Children-III/IV (WISC-III/IV (6,7)) for children or Wechsler Adult Intelligence Scale for Adults-III/IV (WAIS-III/IV (8,9)) for adults. Standardized estimates of verbal IQ (VIQ), performance IQ (PIQ), and full-scale IQ (FIQ) were derived using IQ norms with mean=100 and SD=±15.

The Autism Diagnostic Observation Schedule (ADOS-G (10)) was used to measure the impact of current, clinically observed core symptoms of autism. Based on ADOS-2 algorithm totals, we report ADOS-2 Calibrated Severity Score (CSS) for 'Social Affect' indexing social-communication difficulties and 'RRBs' indexing restricted and repetitive behaviors. CSS Total serves an overall indicator of ASD severity. The CSS ranges from 1 to 10, with higher scores indicating more severe ASD symptom severity. For follow-up analyses (see relative importance section) we median-split the CSS measure into a group of individuals with high CSS and one with low CSS.

The Autism Diagnostic Interview-Revised (ADI-R) (11) is a structured parent interview completed by parents or caregivers of participants with autism. Algorithm scores were derived from current and historical symptom information for the domains of Reciprocal Social Interaction, Communication, and Restricted, Repetitive and Stereotyped Behaviors and Interests. The ADI-R also assessed history of language development.

Language delay was defined as having onset of first words later than 24 months and/or having onset of first phrases later than 33 months. ADHD symptoms were assessed with the DSM-5 ADHD rating scale, covering both inattention and hyper-activity/impulsivity symptoms based on either self-or parent-report (12). Self-report scores were only used when parent-report scores were unavailable (N=83). A categorical variable was computed based on the DSM-5 criteria (i.e., at least five positive responses in children and six in adults on either or both scales) which was used in the follow-up analyses (relative importance section). Handedness was assessed with the short version of the Edinburgh Handedness Inventory (13). Scores ranged between +500 (right-handed) and -500 (left-handed). For sample characterization in Table 1, a categorical variable was computed comprising right-handed (+500 to +150), ambidextrous (-149 to +149) and left-handed (-150 to -500). For follow-up analyses, shown in Table S3, we used a categorical variable comprising right-handed (+500 to +150) and non-right-handed (-500 to +149) individuals. Socioeconomic status (SES) was approximated by using information on annual household income. Annual household income was measured on an 8-point-scale ranging from <£25,000 to >£150,000, with the median annual household income at £30,000–£39,999. For detailed information on clinical characteristics, see (14).

MRI data acquisition

All participants were scanned with a contemporary MRI scanner operating at 3T at 5 different sites (University of Cambridge: Siemens Verio; King's College London: GE Medical Systems Discovery MR 750; Mannheim University: Siemens TimTrio; Radboud University: Siemens Skyra; Rome University: GE Medical Systems Signa HDxTt; Utrecht University: Philips Medical Systems Achieva/Ingenia CX). High-resolution structural T1-weighted volumetric images were acquired with full head coverage, at 1.2-mm thickness with 1.2x1.2-mm in-plane resolution. For all other scanning parameters, please see Table S2. Consistent image quality was ensured by a semi-automated quality control procedure.

Gaussian process regression

Gaussian process regression (15) was used to estimate separate normative models of grey matter (GM) laterality at each voxel. Other methods are also suited to this purpose (e.g. Bayesian polynomial regression), however, Gaussian process regression provides superior estimation of the mean and the ability to map the variation across the cohort through centiles of predictive confidence. For a full treatment of Gaussian processes see (16,17).

A Gaussian process specifies a distribution over functions, such that any finite number of elements has a joint Gaussian distribution. They are excellent tools for Bayesian regression: given a dataset specified by $\mathcal{D} = \{\mathbf{x}_i, y_i \in\}_{i=1}^{N}$ – where $\mathbf{x}_i$ are D-dimensional vectors of covariates, N is the total sample size and $y_i \in \mathbb{R}$ are response variables – the response variables are predicted using a potentially nonlinear regression model with additive Gaussian noise, i.e.: $y_i = f_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_n^2)$. Inference then proceeds by placing a Gaussian process prior over this function then computing the posterior distribution using the canonical Gaussian process regression predictive equations (17). This prior is uniquely specified by a mean ($m(x)$) and covariance ($k(x, x')$) function. Here, without loss of generality we choose a mean function equal to zero and a generic covariance function combining linear and non-linear terms, i.e.:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j + \sigma_f \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\Lambda}(\mathbf{x}_i - \mathbf{x}_j)\right)$$

Where σ_f is a signal amplitude parameter for the nonlinear component and $\boldsymbol{\Lambda}$ is a diagonal matrix with ℓ_d^{-2} along the leading diagonal. These are 'automatic relevance determination' parameters (17) that can down-weight irrelevant dimensions in the input space or emphasize important dimensions. Training a Gaussian process model refers to finding the optimal values for the model parameters which are: $\ell_1, \dots, \ell_D, \sigma_n$ and σ_f. This is conveniently achieved by maximizing the logarithm of the model evidence (i.e. the denominator of Bayes rule). Finally, we compute a single subject Z-statistic image for each subject (i) and at each brain location (j) by computing:

$$\frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\sigma_{ij}^2 + \sigma_{nj}^2}}$$

Here, $\hat{y}_{ij}$ is the predicted mean and predicted variance, σ_{ij}, which is combined with the true response (y_{ij}) and variance learned from the TD distribution (σ_{nj}). Because we estimate a separate noise parameter for each voxel, this should accommodate regional differences in population variation (for example, the estimated variance parameter will be higher in the regions where there is greater variation across individuals).

The Bayesian statistical model takes various sources of uncertainty into account, automatically making inferences more conservative in regions where data are sparse. Thus, the normative model can be estimated solely based on the neurotypical cohort (N=233), and avoids enrichment for autism.

Cross validation

To assess generalizability, we used 10-fold cross validation partitioning the data into 10 folds and repeatedly trained the model on 90% of the data, withholding the remaining 10% for estimating generalization performance. This was done 10 times so that each partition was excluded once. This is standard procedure in machine learning and provides unbiased estimates of the true generalizability.

Structural and functional ROIs

Structural ROIs were based on the Harvard-Oxford atlas (18). The Harvard-Oxford atlas was first coregistered using the nearest-neighbor method to the symmetrical study-specific template in MNI space and constrained to voxels in the study-specific template. This resulted in 48 cortical and 8 subcortical right-hemisphere ROIs. Functional ROIs were based on association test maps from the online meta-analytic database neurosynth (http://neurosynth.org, accessed June 2019) (19) using the search terms 'language', 'motor' (left-lateralized), 'visuospatial', 'attention' (right-lateralized) and 'monitoring', 'mentalizing' (no lateral bias). Maps were resliced to match the voxel resolution of the data, binarized, reflected along the *x* axis and the conjunction of right and left ROIs was used for the analyses to ensure symmetrical ROIs

Results based on alternative thresholding approaches

FDR-correction of p-maps

To apply a stringent thresholding approach, the p-map associated with the map showing the correlation between true and predicted values depicted in Supplementary Figure S3 was FDR-corrected, binarized and multiplied with the normative probability maps. All our original findings also remain valid using this approach. Results were computed in line with the primary analyses and revealed that overall, males and females with autism showed both significantly more extreme rightward deviations ($F_{(1,580)}$=12.3, p<0.001, η_p^2=0.026) and leftward deviations ($F_{(1,580)}$=8.4, p=0.004, η_p^2=0.014) compared to neurotypical males and females. The main effects for sex differences were not significant (right: $F_{(1,580)}$=1.1, p=0.3, η_p^2<0.01; left: $F_{(1,580)}$=0.1, p=0.75, η_p^2<0.01). The sex-by-diagnosis interactions were not significant for extreme left- ($F_{(1,580)}$<0.01, p=0.9, η_p^2<0.01) and for extreme rightward deviations ($F_{(1,580)}$=3.6, p=0.06, η_p^2<0.01). For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,468)}$=8.9, p<0.001, η_p^2=0.037; left: $F_{(2,468)}$=6.6, p=0.001, η_p^2=0.028). The main effects of sex (right: $F_{(1,580)}$=0.1, p=0.81, η_p^2<0.01; left: $F_{(1,580)}$=0.2, p=0.64, η_p^2<0.01) and for the sex-by-LD interaction (right: $F_{(2,468)}$=2.3, p=0.1, η_p^2<0.01, left: $F_{(2,468)}$=1.3, p=0.27, η_p^2<0.01) were not significant. We acknowledge that the p-value map based on the true and predicted correlation values might be biased due to cross-validation, however obtaining permutation-based results would be computationally too expensive in this case.

FDR correction of normative probability maps

Normative probability maps in the main analysis were thresholded at an absolute value of Z>|2.6| (32,51) based on the following rationale: 1) a fixed threshold simplifies the comparison across individuals, which is complicated when controlling the false discovery rate (FDR) separately for each normative probability map; 2) FDR-correction is insensitive to an overall shift in deviations from the normative model in each subject, meaning if one subject has small deviations across the entire cortex, they may seem to have a typical pattern when using FDR-thresholding, because the overall distribution of deviations is shifted. Nevertheless, we repeated the analyses using FDR-thresholding as outlined here:

Overall, males and females with autism showed both significantly more extreme rightward deviations ($F_{(1,580)}$=27.1, p<0.001, η_p^2=0.045) and leftward deviations ($F_{(1,580)}$=15.1, p<0.001, η_p^2=0.025) compared to neurotypical males and females. The main effects for sex differences were significant (right: $F_{(1,580)}$=8.3, p=0.004, η_p^2=0.014; left: $F_{(1,580)}$=6.0, p=0.01, η_p^2=0.01). The sex-by-diagnosis interaction was not significant for extreme left- ($F_{(1,580)}$=2.8, p=0.1, η_p^2<0.01), but for extreme rightward deviations ($F_{(1,580)}$=6.6, p=0.01, η_p^2=0.011), with females with autism showing the strongest rightward deviations. For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,468)}$=13.9, p<0.001, η_p^2=0.066; left: $F_{(2,468)}$=10.6, p<0.001, η_p^2=0.043). This was not the case for sex for both rightward deviations ($F_{(1,468)}$=2.9, p=0.09, η_p^2<0.01) and for leftward deviations ($F_{(1,468)}$=2.3, p=0.13, η_p^2<0.01). The sex-by-LD interaction was not significant for left- ($F_{(2,468)}$=2.2, p=0.11, η_p^2<0.01), however for rightward deviations ($F_{(2,468)}$=3.5, p=0.03, η_p^2=0.015), showing that the stepwise pattern was more pronounced in males with autism than in females with autism. For extreme rightward deviations, follow-up analyses showed that individuals with autism and LD were not significantly different from each other ($t_{(227)}$=0.8, p=0.43), however both individuals with autism with and without LD were significantly different from neurotypical individuals (neurotypicals vs. autism-LD: $t_{(108)}$=4.8, p<0.001), neurotypicals vs. autism-noLD: $t_{(169)}$=3.8, p<0.001). For extreme leftward deviations, individuals with autism and LD were not different from individuals with autism without LD ($t_{(205)}$=0.9, p=0.39), however both individuals with autism with and without LD were significantly different from neurotypical individuals (neurotypicals vs. autism-LD: $t_{(136)}$=4.6, p<0.001), neurotypicals vs. autism-noLD: $t_{(161)}$=3.3, p=0.001).

Detailed language delay (LD) results

To assess the impact of LD, statistical second level-analyses were re-run by including LD (i.e., autism with LD, autism without LD, neurotypicals) as independent variable (instead of diagnostic group (i.e., autism and neurotypicals)). For follow-up analysis, we only selected those ROIs functionally related to language. For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,468)}$=10.5, p<0.001, η_p^2=0.043; left: $F_{(2,468)}$=10.0, p<0.001, η_p^2=0.04), but not for sex (right: $F_{(1,468)}$=0.1, p=0.8, η_p^2<0.01; left: $F_{(1,468)}$=0.001, p=0.97, η_p^2<0.01), nor the sex-by-LD interactions (right: $F_{(2,468)}$=1.5, p=0.22, η_p^2<0.01; left: $F_{(2,468)}$=1.5, p=0.26, η_p^2<0.01) (see Figure 4). These results remained unchanged when controlling for FIQ, handedness and SES (see next section). For extreme rightward deviations, follow-up analyses showed that individuals with autism and LD were significantly different from both individuals with autism without LD ($t_{(213)}$=2.5, p=0.01, d=0.32) and neurotypical individuals ($t_{(152)}$=4.6, p<0.001, d=0.58), while individuals with autism without LD did not reach a significant difference from neurotypical individuals ($t_{(256)}$=1.9, p=0.06, d=0.21). This stepwise pattern was overall more pronounced in males with autism than in females with autism. In contrast, for extreme leftward deviations, individuals with autism and LD were not different from individuals with autism without LD ($t_{(229)}$=1.2, p=0.22, d=0.16), however both individuals with autism with and without LD were significantly different from neurotypical individuals (neurotypicals vs. autism-LD: $t_{(182)}$=4.4, p<0.001, d=0.53, neurotypicals vs. autism-noLD: $t_{(271)}$=3.0, p=0.003, d=0.32). When matching individuals with autism with and without LD on symptom severity and age, results remained stable (see section ‘Matched sub-sample in autism’ and see Figure S9).

For further details see Supplementary Table 3.

Controlling for confounds

We repeated second-level statistical analyses controlling for FIQ, handedness (dimensionally based on the Edinburgh Handedness Inventory) and SES (based on income) in the same model. Due to missing values for handedness and SES information, this analysis was conducted in a smaller dataset (males with autism=178, females with autism=73, neurotypical males=99, neurotypical females=51). Results remained unchanged despite the smaller sample size. Overall, males and females with autism showed both significantly more extreme rightward deviations ($F_{(1,387)}$=7.7, p=0.006, η_p^2=0.019) and leftward deviations ($F_{(1,387)}$=4.0, p=0.05, η_p^2=0.01) compared to neurotypical males and females. There were no sex differences (right: $F_{(1,387)}$=0.4, p=0.53, η_p^2<0.01; left: $F_{(1,387)}$=0.54, p=0.5, η_p^2<0.01) and no sex-by-diagnosis interactions (right: $F_{(1,387)}$=3.5, p=0.06, η_p^2<0.01; left: $F_{(1,387)}$=0.4, p=0.5, η_p^2<0.01). For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,313)}$=4.6, p=0.01, η_p^2=0.029; left: $F_{(2,313)}$=4.2, p=0.01, η_p^2=0.026), however not for sex (right: $F_{(1,313)}$=0.01, p=0.9, η_p^2<0.01; left: $F_{(1,313)}$=1.5, p=0.21, η_p^2<0.01), or the sex-by-LD interactions (right: $F_{(2,313)}$=2.6, p=0.08, η_p^2=0.016; left: $F_{(2,313)}$=2.0, p=0.13, η_p^2=0.013).

Exclusion of individuals with autism and intellectual disability (ID)

We excluded individuals with autism and ID, resulting in a sample of 230 males with autism and 77 females with autism. Overall, males and females with autism showed both significantly more extreme rightward deviations ($F_{(1,522)}$=8.8, p=0.003, η_p^2=0.017) and leftward deviations ($F_{(1,522)}$=9.3, p=0.002, η_p^2=0.017) compared to neurotypical males and females. There were no sex differences (right: $F_{(1,522)}$=1.8, p=0.18, η_p^2<0.01; left $F_{(1,522)}$=1.9, p=0.17, η_p^2<0.01) and no sex-by-diagnosis interactions (right:

$F_{(1,522)}$=0.6, p=0.42, η_p^2<0.01; left: $F_{(1,522)}$<0.001, p=0.99, η_p^2<0.01). For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,417)}$=7.4, p=0.01, η_p^2=0.034; left: $F_{(2,417)}$=9.0, p<0.001, η_p^2=0.041), however not for sex (right: $F_{(1,417)}$=2.2, p=0.14, η_p^2<0.01; left: $F_{(1,417)}$=0.8, p=0.36, η_p^2<0.01), or the sex-by-LD interactions (right: $F_{(2,417)}$=0.3, p=0.73, η_p^2<0.01; left: $F_{(2,417)}$=0.5, p=0.62, η_p^2<0.01).

Handedness

To check for the influence of handedness, two additional approaches were applied that confirmed primary analyses. Both control analyses converged to show that despite reduced sample size, results remained unchanged. First, we created another normative model in a reduced subsample that had handedness information available for all included subjects in the analysis (males with autism=208, females with autism=84, NT males=117, NT females=66). All analyses steps as in the primary analysis were applied to obtain information on extreme right- and leftward deviations. Results showed that there was a significant group effect for rightward deviations ($F_{(1,470}$=7.04, p=0.008, η_p^2=0.015) and a marginal effect for leftward deviations ($F_{(1,470)}$=3.9, p=0.05, η_p^2=0.01). The effects of sex (right: $F_{(1,470)}$=0.03, p=0.86, η_p^2<0.01; left: $F_{(1,470)}$=1.02, p=0.31, η_p^2<0.01) and the sex-by-diagnosis interaction (right: $F_{(1,470)}$=1.1, p=0.3; left: $F_{(1,470)}$=0.6, p=0.45, η_p^2<0.01) were not significant. When considering LD, for both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,376)}$=7.9, p<0.001, η_p^2=0.04; left: $F_{(2,376)}$=5.7, p=0.004, η_p^2=0.029), however not for sex (right: $F_{(1,376)}$=0.001, p=0.98, η_p^2<0.01; left: $F_{(1,376)}$=0.4, p=0.51, η_p^2<0.01), nor for the sex-by-LD interaction (right: $F_{(2, 376)}$=0.81, p=0.45, η_p^2<0.01; left: $F_{(2,376)}$=1.2, p=0.31, η_p^2<0.01).

Next, a subsample was created by excluding left-handed and ambidextrous subjects (specified as a handedness score below 150). This resulted in a reduced sample of 106 males with autism, 47 females with autism, 69 NT males and 37 NT females. Results in this sample indicated that for both extreme left- and rightward deviations, there was a significant main effect of group (right: $F_{(1,254)}$=4.4, p=0.03, η_p^2=0.017; left: $F_{(1,254)}$=17.1, p<0.001, η_p^2=0.063), however not for sex (right: $F_{(1,254)}$=1.5, p=0.2, η_p^2<0.01; left: $F_{(1,254)}$=2.8, p=0.09, η_p^2=0.011) or the group-by-sex interaction (right: $F_{(1,254)}$=2.3, p=0.13, η_p^2<0.01; left: $F_{(1,254)}$=0.01, p=0.9, η_p^2<0.01). When considering LD, for both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,252)}$=5.1, p<0.001, η_p^2=0.039; left: $F_{(2,252)}$=10.3, p<0.001, η_p^2=0.074), however not for sex (right: $F_{(1,252)}$=2.0, p=0.16, η_p^2<0.01; left: $F_{(1,252)}$=3.3, p=0.07, η_p^2=0.013), nor for the sex-by-LD interaction (right: $F_{(2,252)}$=1.6, p=0.2, η_p^2=0.013; left: $F_{(2,252)}$=1.7, p=0.18, η_p^2=0.014).

Matched sub-sample in autism (language-delayed vs. not-language-delayed)

We used the python package ‘pymatch’ (https://github.com/benmiroglio/pymatch) to create age- and symptom-severity matched sub-samples of individuals with autism with and without language delay (LD). Matching resulted in a sample of 93 individuals with autism with LD (males=74; females=19) and 62 individuals with autism without LD (males=46; females=16). There were no significant differences between them on age ($t_{(126)}$=0.1, p=0.91) and symptom severity as assessed by the CSS on the ADOS ($t_{(130)}$=0.9, p=0.38). For both extreme left- and rightward deviations, there was a significant main effect of LD (right: $F_{(2,381)}$=8.3, p<0.001, η_p^2=0.042; left: $F_{(2,381)}$=10.5, p<0.001, η_p^2=0.052), however not for sex (right: $F_{(1,381)}$=1.7, p=0.19; left: $F_{(1,381)}$=0.04, p=0.83, η_p^2<0.01). The sex-by-LD interaction was not significant for left- ($F_{(2,381)}$=1.6,

p=0.2, η_p^2<0.01), however trending for rightward deviations ($F_{(2,381)}$=3.0, p=0.05, η_p^2=0.015), showing that the stepwise pattern was more pronounced in males with autism than in females with autism (see Figure S9).

Replication sample ABIDE I and II

To assess replicability, we selected a sample from the publicly available Autism Brain Imaging Data Exchange (ABIDE) I and II (20,21). Sites using the same scanner and acquisition parameters across ABIDE I and II were merged into one single site (KKI, NYU, OHSU, SDSU and UCLA 1) resulting in 15 sites in total. Selection criteria were as follows: a) subjects in the same age range as the primary analysis sample between 6-30 years of age, b) images with low image quality (due to excessive motion, insufficient coverage or brain abnormalities) and failed preprocessing were excluded; c) neurotypical individuals with an ADHD diagnosis were excluded. This resulted in a sample of 418 males with autism, 95 females with autism, 473 neurotypical males and 218 neurotypical females. No information on language delay was available. For further details on demographic and clinical information see Table S2. T1 images were preprocessed with the same pipeline as described in the methods of the main manuscript and as shown in Figure S2. No study-specific DARTEL template was created, but images were registered to the DARTEL template obtained from primary image analysis as described in the main text. Next, normative modeling and the creation of normative probability maps were done in accordance with the primary imaging analysis in the main text. Finally, we also tested for the spatial overlap between overlap maps in EU-AIMS LEAP (thresholded at 2%) and overlap maps in ABIDE (thresholded at 2%) to identify most strongly implicated regions across the datasets.

Replicability results – EU-AIMS LEAP and ABIDE overlap

Regions with highest overlap in EU-AIMS LEAP and ABIDE for extreme rightward deviations were in the middle and superior temporal gyrus, precentral gyrus, hippocampus, putamen and intracalcerine cortex in males with autism and in the thalamus, parahippocampal gyrus, precuenus, intra- and supracalcerine cortex, superior frontal gyrus, frontal pole, orbitofrontal cortex and middle temporal gyrus in females with autism (see Figure 11A). Regions with highest overlap in EU-AIMS LEAP and ABIDE for extreme leftward deviations were in the thalamus and cerebellum in males with autism and in the caudate, anterior cingulate cortex, frontal pole and medial frontal cortex in females with autism (see Figure 11B).

Supplemental References

1. American Psychiatric Association; (1994): Diagnostic and Statistical Manual of Mental Disorders Source Information. *Am Psychiatr Assoc*.

2. American Psychiatric Association (2000): Diagnostic and statistical manual of mental disorders: DSM-IV-TR (text revision). *American Journal of Psychiatry*.

3. American Psychiatric Association (2013): Diagnostic and Statistical Manual of Mental Disorders 1. American Psychiatric Association. *Arlington*. https://doi.org/10.1176/appi.books.9780890425596.893619

4. World Health Organization (1993): The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*.

5. Wechsler D (2011): *Wechsler Abbreviated Scale of Intelligence–Second Edition (WASI-II).* San Antonio, TX: NCS Pearson.

6. Wechsler D (1991): *Wechsler Intelligence Scale for Children (3rd Ed.)*. San Antonio, TX: Psychological Corporation.

7. Wechsler D (2003): *Wechsler Intelligence Scale for Children (4th Ed.)*. San Antonio, TX: Psychological Corporation.

8. Wechsler D (1997): *Wechsler Adult Intelligence Scale (3rd Ed.)*. San Antonio, TX: The Psychological Corporation.

9. Wechsler D (2008): *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio, TX: Pearson.

10. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, *et al.* (2000): The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* 30: 205–223.

11. Lord C, Rutter M, Le Couteur A (1994): Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 24: 659–685.

12. DuPaul GJ, Power TJ, Anastopoulos AD, Reid R (2016): ADHD Rating Scale-5 for children and adolescents: Checklists, norms, and clinical interpretation. *ADHD Rating Scale-5 for Children and Adolescents: Checklists, Norms, and Clinical Interpretation.*

13. Oldfield RC (1971): The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9: 97–113.

14. Charman T, Loth E, Tillmann J, Crawley D, Wooldridge C, Goyard D, *et al.* (2017): The EU-AIMS Longitudinal European Autism Project (LEAP): Clinical characterisation. *Mol Autism*. https://doi.org/10.1186/s13229-017-0145-9

15. Rasmussen CE, Williams CKI (2018): Model Selection and Adaptation of Hyperparameters. *Gaussian Processes for Machine Learning*. https://doi.org/10.7551/mitpress/3206.003.0008

16. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF (2016): Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biol Psychiatry Cogn Neurosci Neuroimaging* 1: 433–447.

17. Rasmussen CE, Williams CKI (2006): Gaussian Processes for Machine Learning. *Adaptive Computation and Machine Learning*. https://doi.org/10.1142/S0129065704001899

18. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, *et al.* (2004): Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2004.07.051

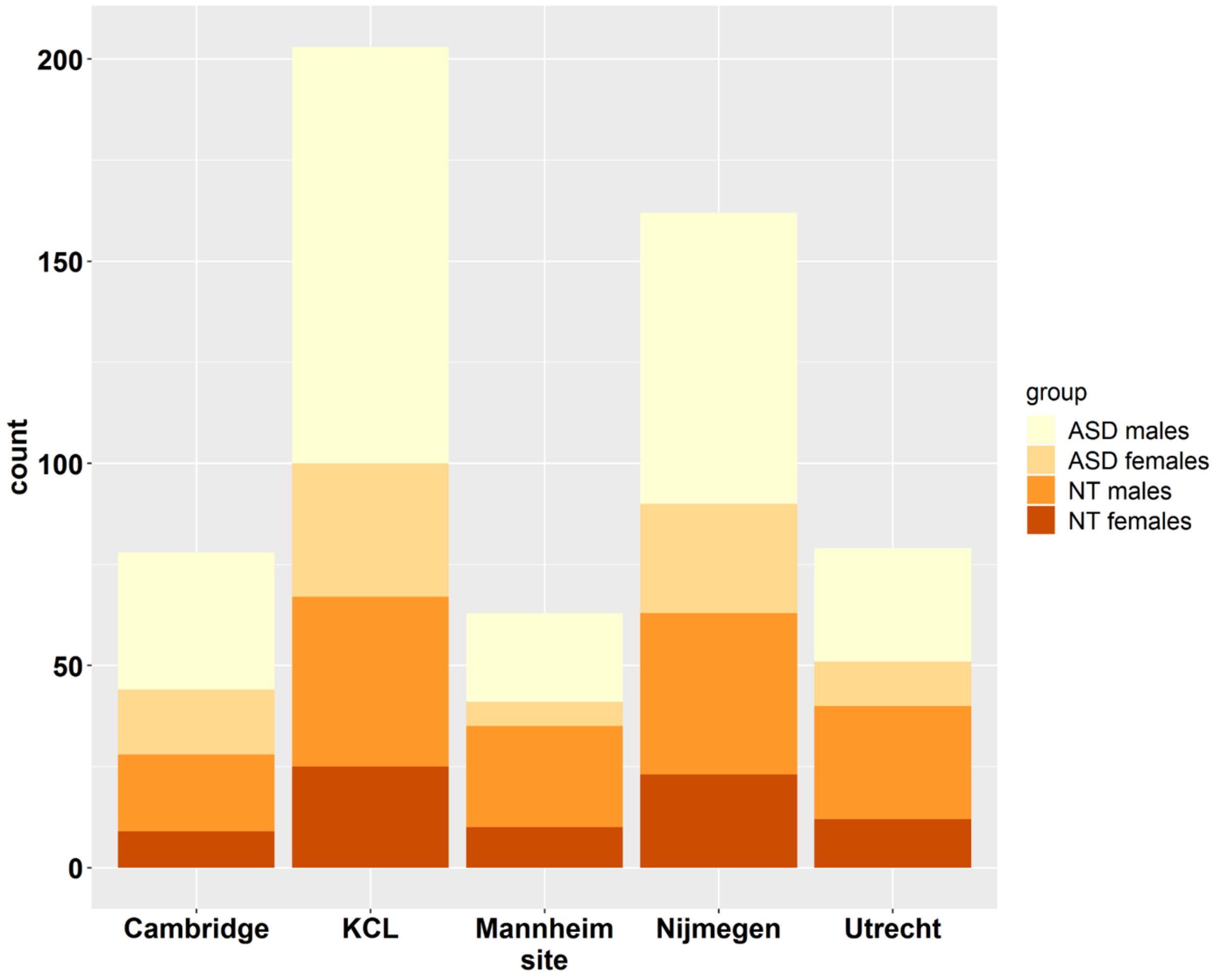
19. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011): Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*. https://doi.org/10.1038/nmeth.1635

20. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, *et al.* (2014): The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*. https://doi.org/10.1038/mp.2013.78

21. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, *et al.* (2017): Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data* 4: 170010.

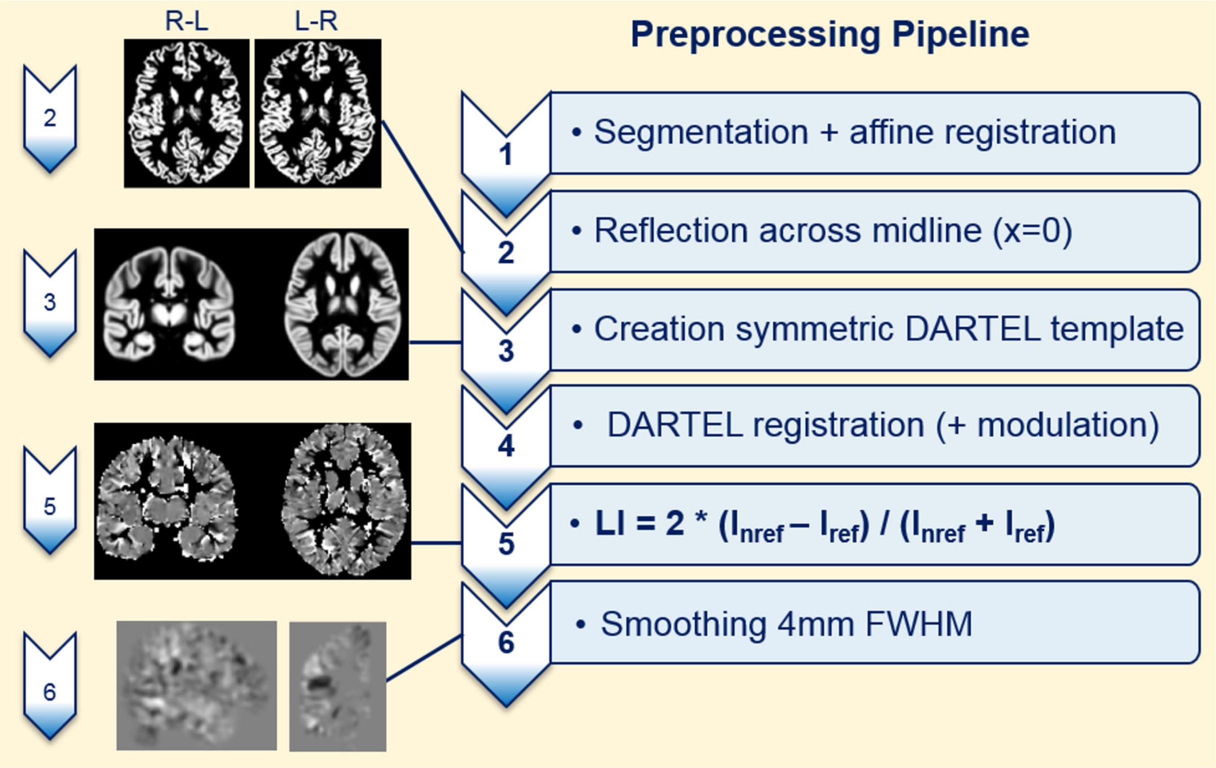
Supplementary Figures and Tables

Figure S1 – Distribution of subjects across acquisition sites



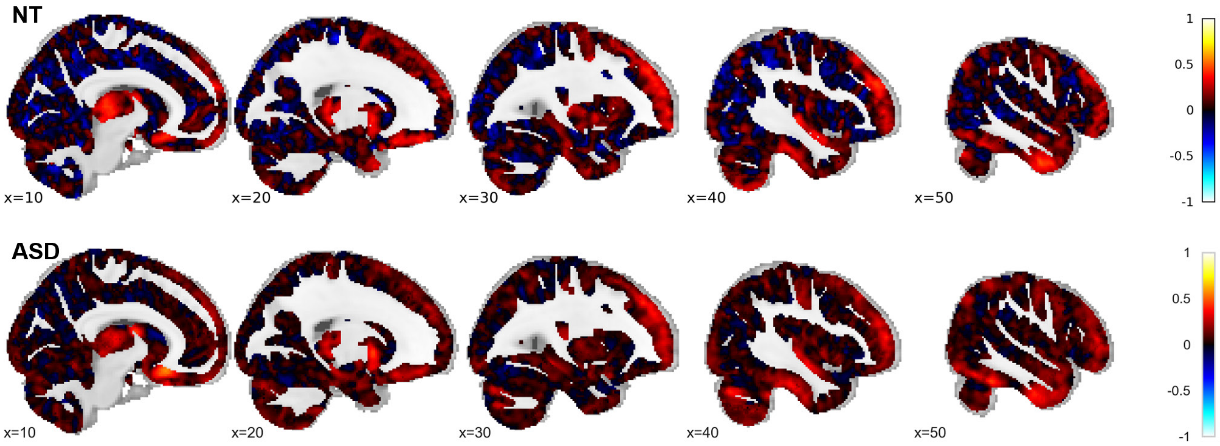
Abbreviations: ASD=autism spectrum disorder, NT=neurotypicals. The bar plots show the distribution of subjects by diagnostic group and sex across the five acquisition sites. Cambridge: males with autism=34, females with autism=16, NT males=19, NT females=9; KCL: males with autism=103, females with autism=33, NT males=42, NT females=25; Mannheim: males with autism=22, females with autism=6, NT males=25, NT females=10; Nijmegen: males with autism=72, females with autism=27, NT males=40, NT females=23; Utrecht: males with autism=28, females with autism=11, NT males=28, NT females=12. The sixth Rome site was excluded due to low number of subjects.

Figure S2 – Summary of grey matter VBM laterality pre-processing pipeline


R-L
L-R
Preprocessing Pipeline
1
• Segmentation + affine registration
2
• Reflection across midline (x=0)
3
• Creation symmetric DARTEL template
4
• DARTEL registration (+ modulation)
5
• LI = 2 * (I_{nref} – I_{ref}) / (I_{nref} + I_{ref})
6
• Smoothing 4mm FWHM

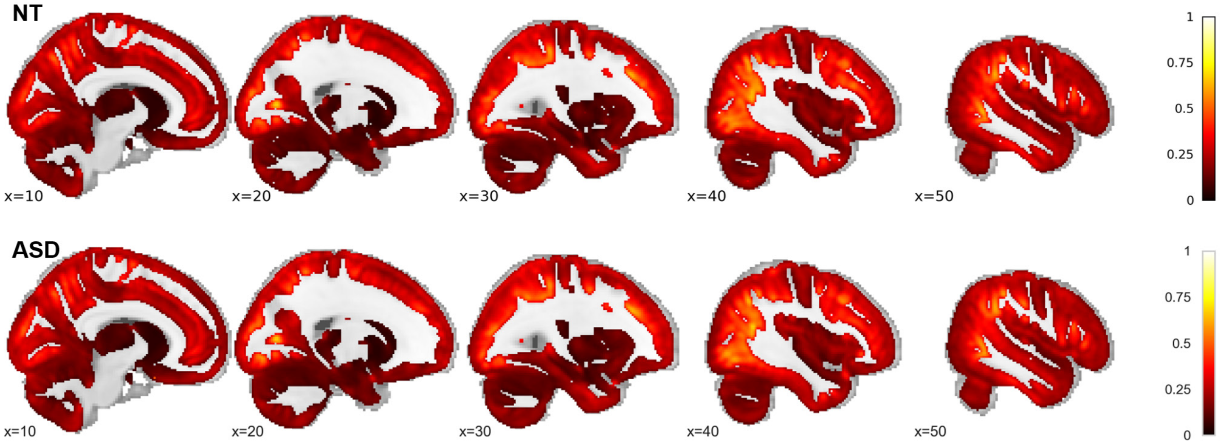
Abbreviations: GM=grey matter, VBM=voxel-based morphometry, DARTEL= Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (51), LI=laterality index, I_{nref} =non-reflected GM images, I_{ref}= reflected GM images. The figure depicts the pre-processing pipeline for structural T1-weighted images. Note that the standard VBM pipeline is adjusted to meet the needs for laterality analyses, i.e., using symmetric registration.

Figure S3 – Mean model accuracy 1



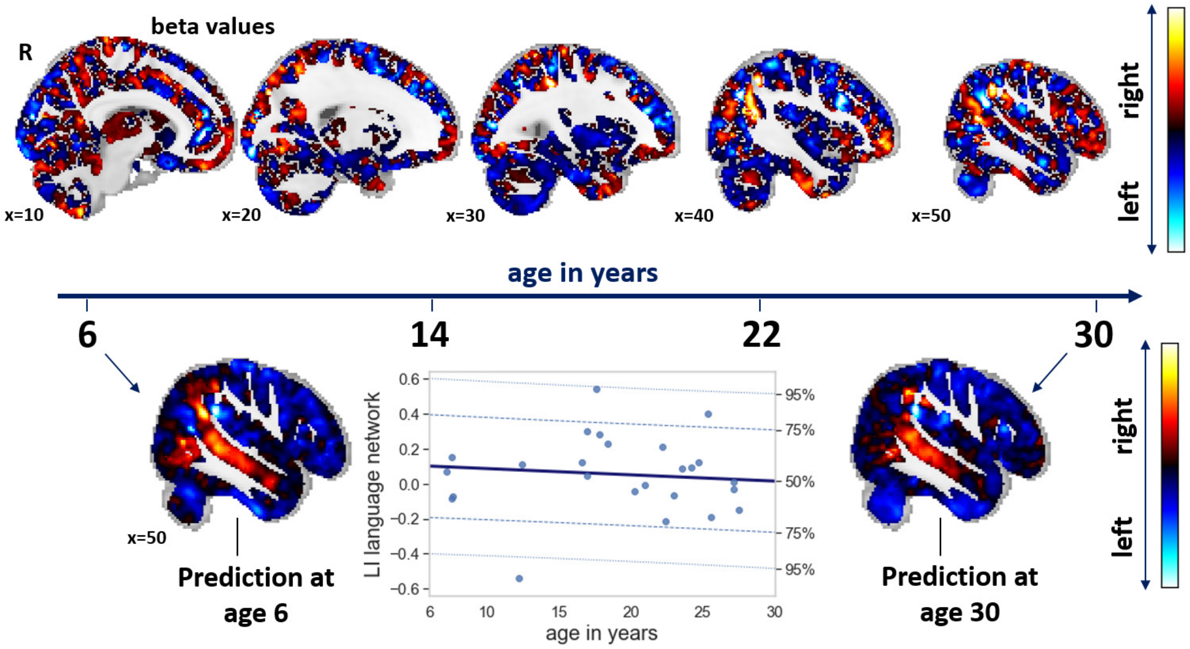
The figure shows the correlation (Rho) between true and predicted grey matter laterality values in NT controls and individuals with autism.

Figure S4 – Mean model accuracy 2



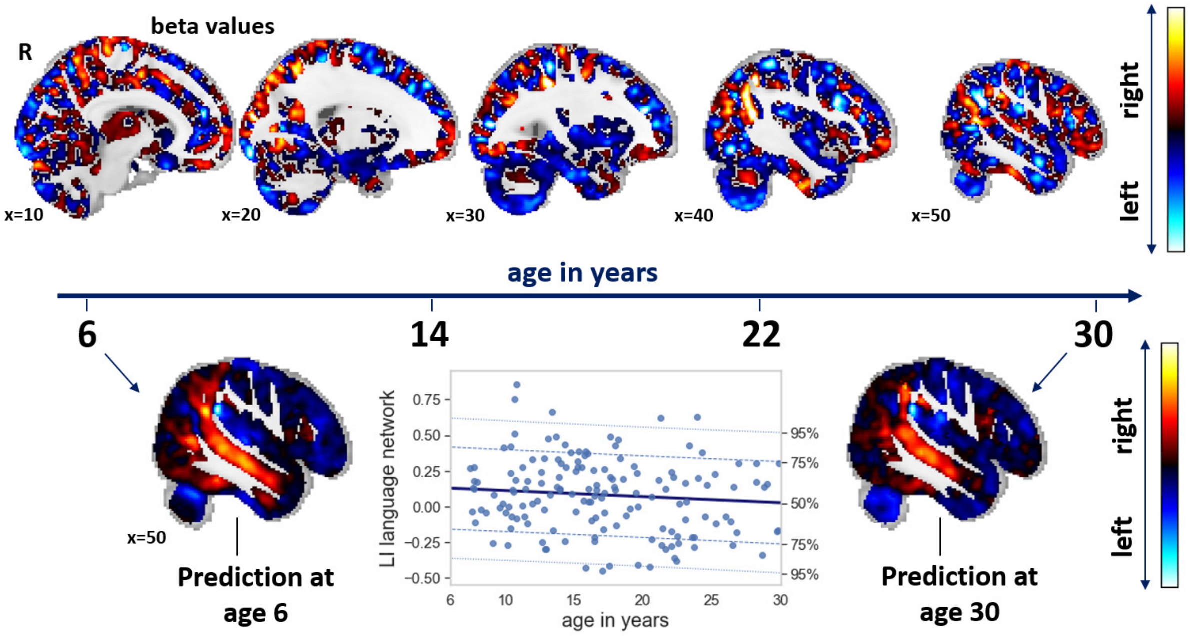
The figure shows the root mean square error of true and predicted mean of grey laterality values in NT controls and individuals with autism.

Figure S5 - Normative developmental changes for grey matter laterality in neurotypical females



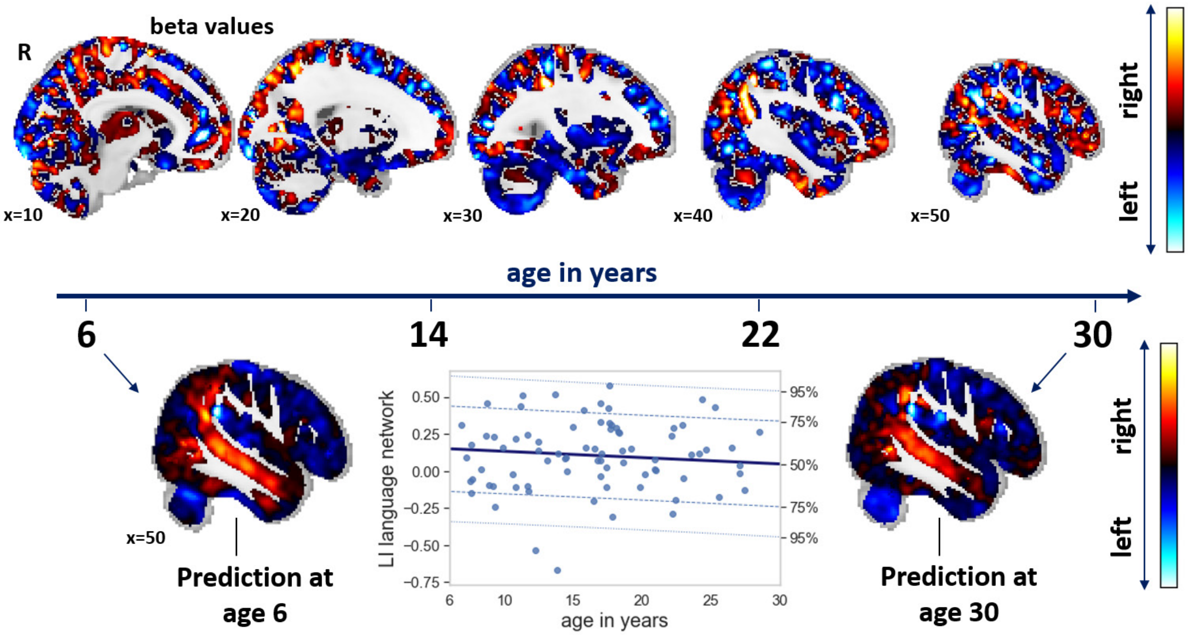
Supplementary Figure 5 captures the spatial representation of the voxelwise normative model in neurotypical (NT) females from the KCL site. The upper panel shows the beta values of laterality change across 6 to 30 years of age. The lower panel shows the actual prediction of grey matter laterality at ages 6 and 30. Blue colors indicate a shift towards leftward asymmetry, whereas red colors indicate a shift towards rightward asymmetry. The regression line depicts the predicted laterality values extracted from the peak voxel of the language network based on neurosynth (x=24, y=54, z=51) between 6 and 30 years of age along with centiles of confidence. These are based on the normative model maps thresholded with the positive correlation map between true and predicted values. Blue dots are the true values for NT females in KCL.

Figure S6A – Normative developmental changes for grey matter laterality in neurotypical males using ComBat



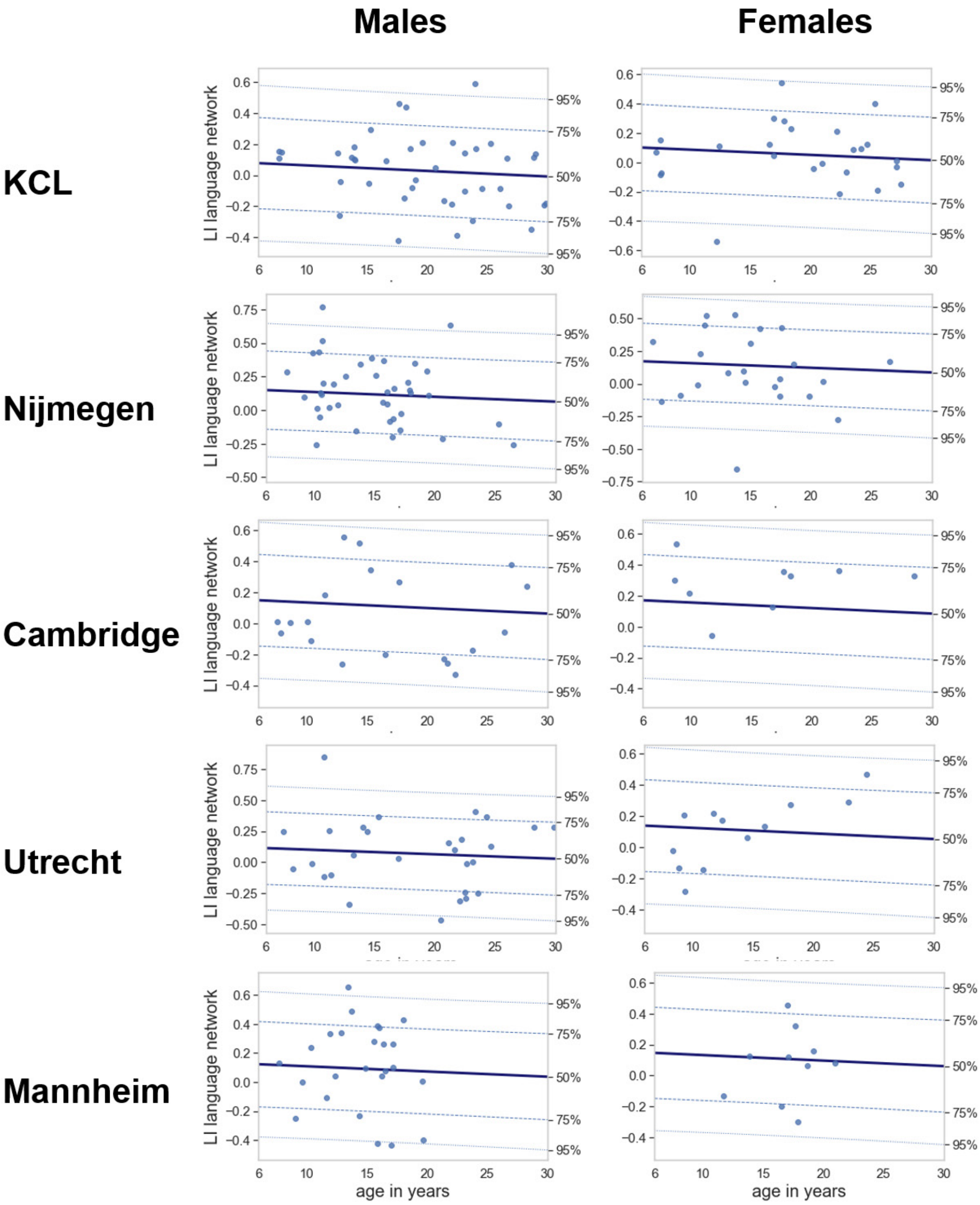
Supplementary Figure 6A captures the spatial representation of the voxelwise normative model in neurotypical (NT) males across all sites. The upper panel shows the beta values of laterality change across 6 to 30 years of age. The lower panel shows the actual prediction of grey matter laterality at ages 6 and 30. Blue colors indicate a shift towards leftward asymmetry, whereas red colors indicate a shift towards rightward asymmetry. The regression line depicts the predicted laterality values extracted from the peak of the language network based on neurosynth (x=24, y=54, z=51) between 6 and 30 years of age along with centiles of confidence. These are based on the normative model maps thresholded with the positive correlation map between true and predicted values. Blue dots are the true values for NT males across all sites.

Figure S6B – Normative developmental changes for grey matter laterality in neurotypical females using ComBat



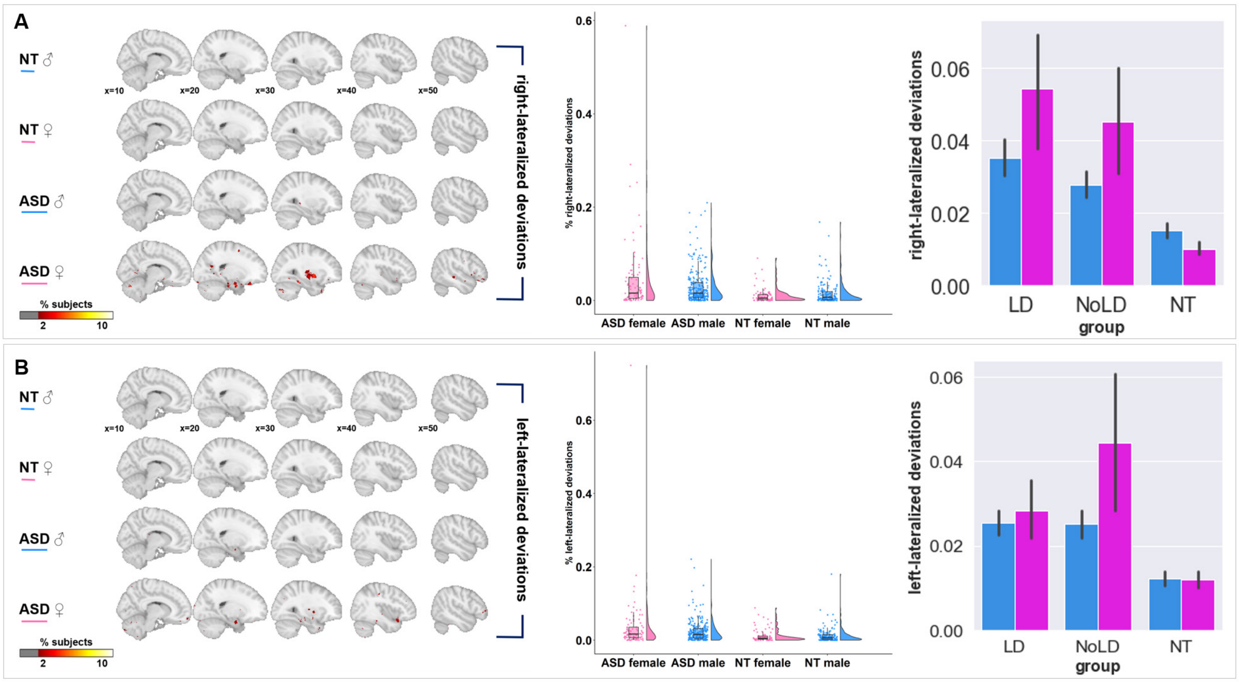
Supplementary Figure 6B captures the spatial representation of the voxelwise normative model in neurotypical (NT) females across all sites. The upper panel shows the beta values of laterality change across 6 to 30 years of age. The lower panel shows the actual prediction of grey matter laterality at ages 6 and 30. Blue colors indicate a shift towards leftward asymmetry, whereas red colors indicate a shift towards rightward asymmetry. The regression line depicts the predicted laterality values extracted from the peak of the language network based on neurosynth (x=24, y=54, z=51) between 6 and 30 years of age along with centiles of confidence. These are based on the normative model maps thresholded with the positive correlation map between true and predicted values. Blue dots are the true values for NT females across all sites.

Figure S7 – Normative model by sex and site in NT individuals for the language network



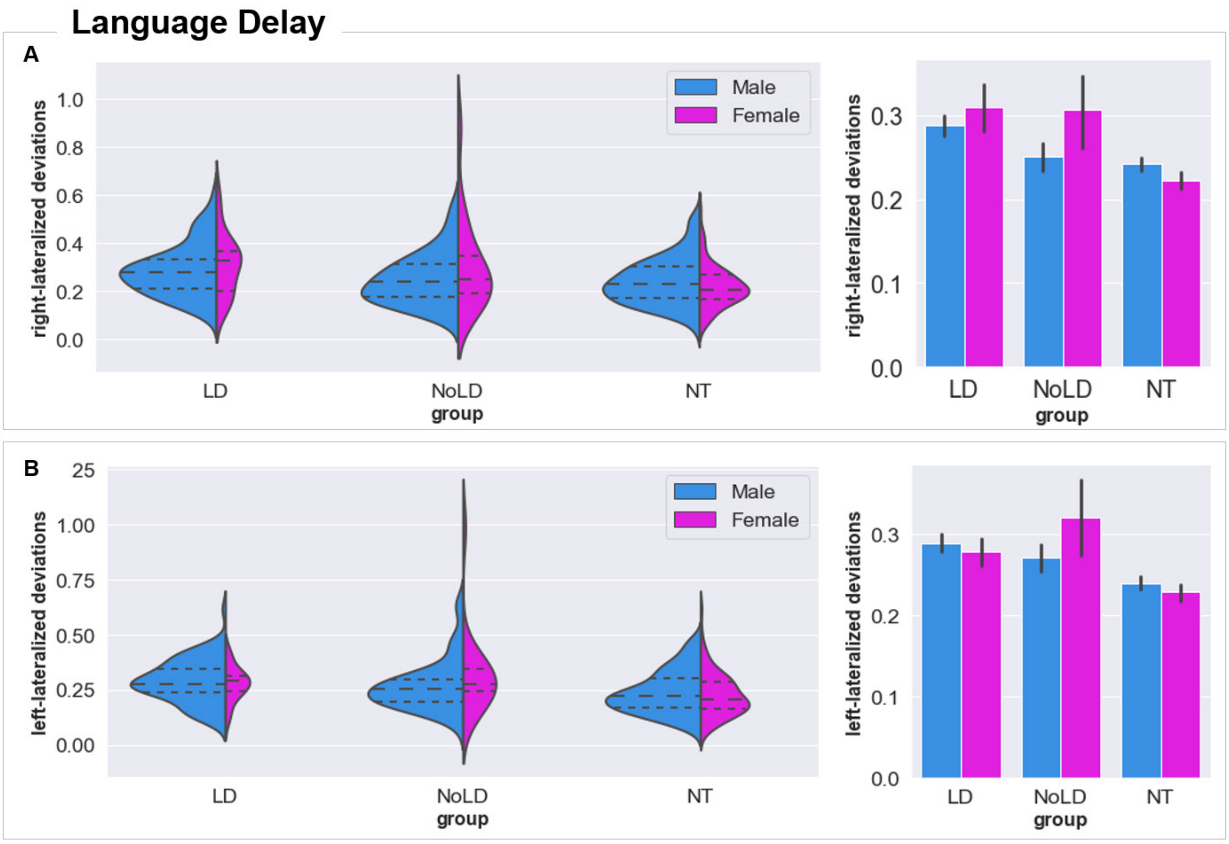
The regression lines depict the predicted laterality values extracted from the peak of the language network based on neurosynth (x=24, y=54, z=51) between 6 and 30 years of age along with centiles of confidence. These are based on the normative model maps thresholded with the positive correlation map between true and predicted values. Blue dots are the true values for NT males on the left and NT females on the right categorized by sites.

Figure S8 – Characterization of extreme laterality deviations using FDR-correction



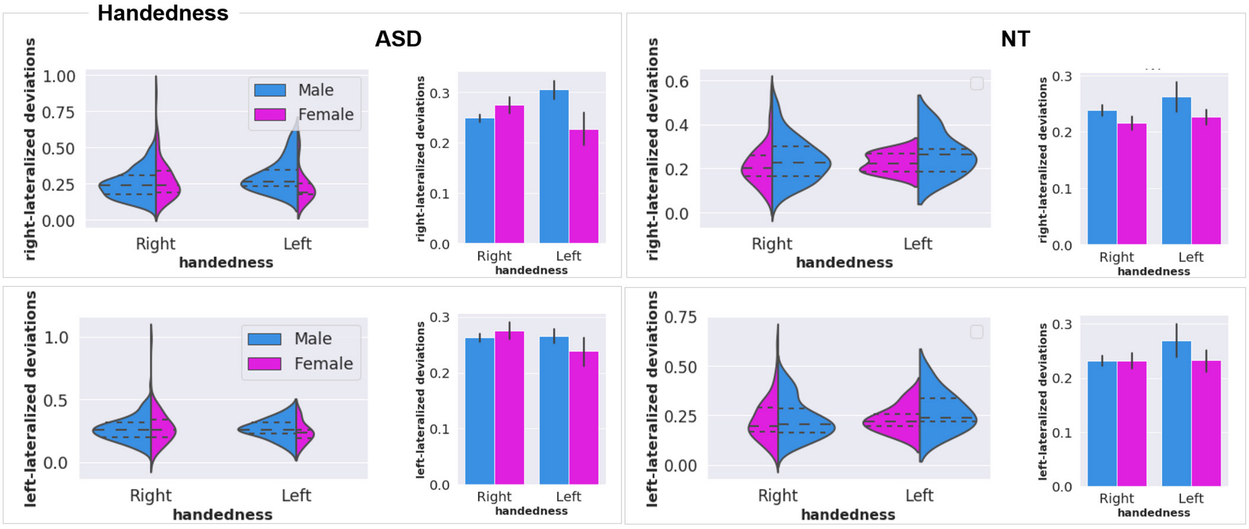
Abbreviations: ASD=autism spectrum disorder, NT=neurotypicals, LD=individuals with autism with language delay, noLD=individuals with autism without language delay. Blue=males, pink/purple=females. The figure depicts the replication of results when using FDR-correction as an alternative thresholding method. The upper panel characterizes extreme rightward deviations (A), the lower panel extreme leftward deviations (B). The left panels show the percentage of extreme right-and leftward deviations from the normative model at each brain locus for each diagnostic group and gender separately. We depict loci where at least 2% of the subjects show overlaps. The violin plots in the middle show the extreme deviations for each individual within each diagnostic and gender group. On average, individuals with autism show more extreme deviations than NT controls for both right- and leftward deviations. The bar plots on the right show that individuals with autism with LD have more extreme-rightward deviations.

Figure S9 – Extreme laterality deviations as a function of language delay in matched subsamples



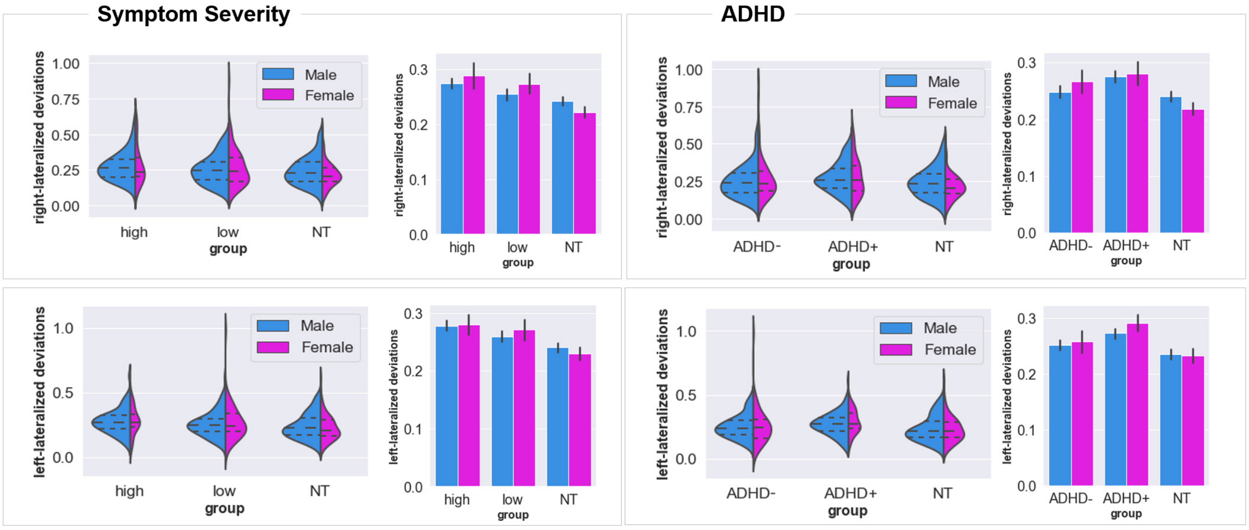
Abbreviations: LD=individuals with autism with language delay, noLD=individuals with autism without language delay, NT=neurotypicals. Language delay was defined as having onset of first words later than 24 months and/or having onset of first phrases later than 33 months. Replicated results are shown in a sample where individuals with autism with and without LD are matched for age and symptom severity. The upper panel characterizes extreme rightward deviations (A), the lower panel extreme leftward deviations (B). The violin and bar plots on the right show that individuals with autism with LD show more extreme rightward deviations than the other two groups.

Figure S10A – Extreme laterality deviations as a function of handedness



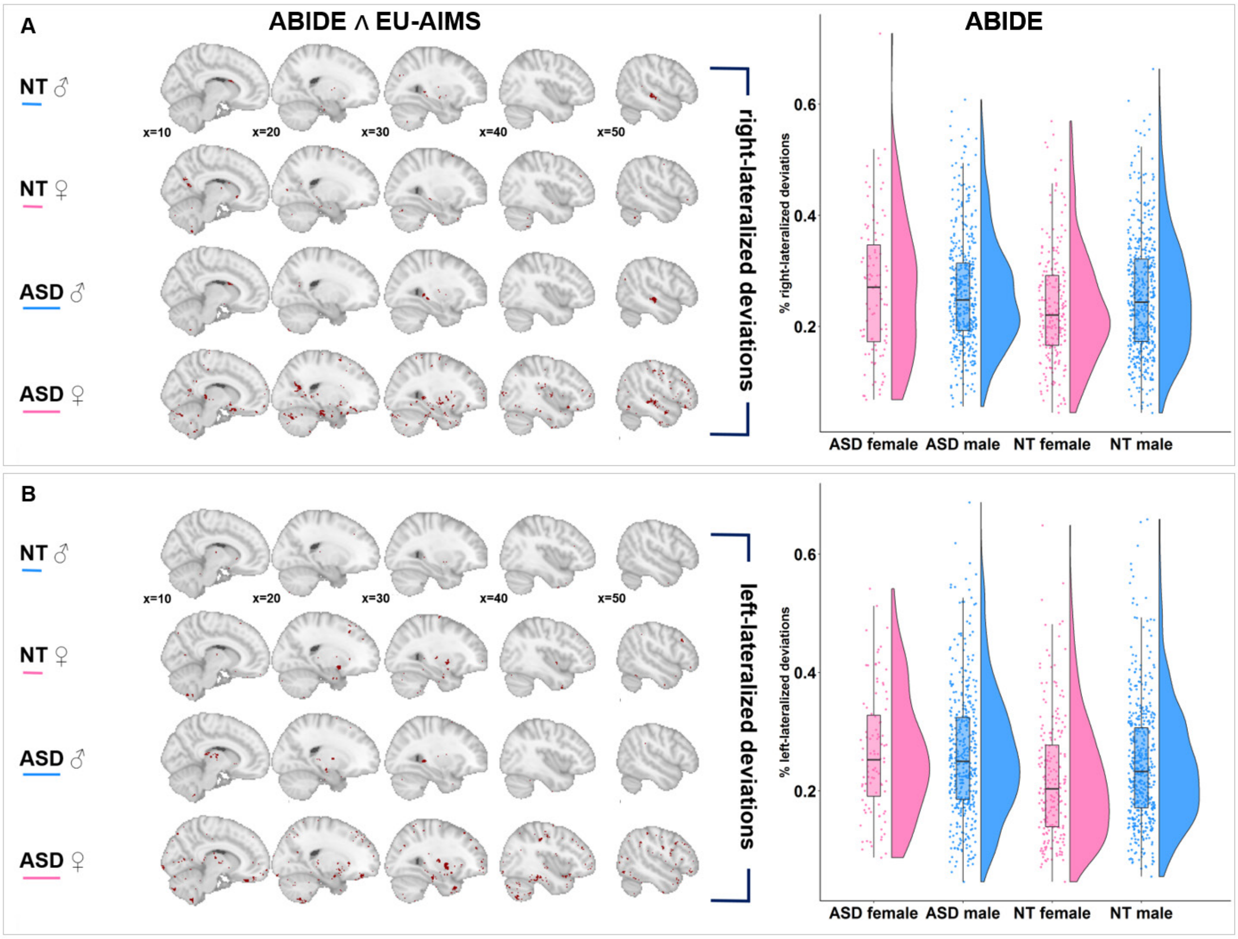
Abbreviations: R=right-handed, NonRight=non-right-handed, ASD=autism spectrum disorder, NT=neurotypicals. Based on the short version of Edinburgh Handedness Inventory, a categorical variable comprising right-handed (+500 to +150) and non-right-handed (-500 to +149) individuals was computed. For group differences, see Table S3.

Figure S10B – Extreme laterality deviations as a function of symptom severity and ADHD



Abbreviations: ADHD+ = individuals with autism and attention-deficit hyperactivity disorder, ADHD– = individuals with autism and without attention-deficit hyperactivity disorder, NT=neurotypicals. ADHD symptoms were assessed with the DSM-5 ADHD rating scale, covering both inattention and hyperactivity/impulsivity symptoms based on either self-or parent-report. A categorical variable was computed based on the DSM-5 criteria (i.e., at least five positive responses in children and six in adults on either or both scales). Symptom severity was captured by the total ADOS-2 Calibrated Severity Score (CSS). We median-split the CSS measure into one group of individuals with autism with high CSS and one with low CSS. For group differences, see Table S3.

Figure S11 – Overlap between ABIDE and EU AIMS



Abbreviations: ASD=autism spectrum disorder, NT=neurotypicals, ABIDE=Autism Brain Imaging Data Exchange. The figure depicts the replication of results in an independent dataset. The upper panel characterizes extreme rightward deviations (A), the lower panel extreme leftward deviations (B). The left panels show the percentage of extreme right-and leftward deviations from the normative model at each brain locus for each diagnostic group and gender separately where ABIDE and EU-AIMS LEAP show overlap. The violin plots on the right show the extreme deviations for each individual within each diagnostic and gender group in the ABIDE dataset. On average, individuals with autism show more extreme deviations than NT controls for both right- and leftward deviations.

Table S1 – Summary of acquisition parameters across sites

Site	Manufacturer	Model	Software Version	Acquisition sequence	Coverage	Slices	Thickness [mm]	Resolution [mm^3]	TR [s]	TE [ms]	FA [°]	FOV
Cambridge	Siemens	Verio	Syngo MR B17	Tfl3d1_ns	256*256	176	1.2	1.1*1.1*1.2	2.3	2.95	9	270
London	GE Medical systems	Discovery mr750	LX MR DV23.1_V02_ 1317.c	SAG ADNI GO ACC SPGR	256*256	196	1.2	1.1*1.1*1.2	7.31	3.02	11	270
Mannheim	Siemens	TimTrio	Syngo MR B17	MPRAGE ADNI	256*256	176	1.2	1.1*1.1*1.2	2.3	2.93	9	270
Nijmegen	Siemens	Skyra	Syngo MRD13	Tfl3d1_16ns	256*256	176	1.2	1.1*1.1*1.2	2.3	2.93	9	270
Rome	GE Medical systems	Signa HDxt	24/LX/MR HD16.0_V02_ 1131.a	SAG ADNI GO ACC SPGR	256*256	172	1.2	1.1*1.1*1.2	5.96	1.76	11	270
Utrecht	Philips Medical Systems	Achieva/ Ingenia CX	3.2.3, 3.2.3.1	ADNI GO 2	256*256	170	1.2	1.1*1.1*1.2	6.76	3.1	9	270

Table S2 - Demographic and clinical characterization of the ABIDE sample

	ASD M (N=418)	ASD F (N=95)	NT M (N=473)	NT F (N=218)	
	Mean (SD) [Range]	Mean (SD) [Range]	Mean (SD) [Range]	Mean (SD) [Range]	Post-hoc
Age	12.8 (4.1)	12.6 (4.3)	12.7 (4.2)	12.5 (4.7)	ns
	[6.8-30.0]	[6.9-27.0]	[7.1-29.0]	[7.0-29.9]	
Full-Scale IQ[a]	106 (17.6)	104 (17.8)	113 (12.6)	113 (13.0)	(ASD M=ASD F) < (NT M=NT F)
	[49-149]	[66-147]	[71-148]	[80-149]	
Verbal IQ[b]	107 (18.7)	104 (17.7)	114 (13.2)	113 (14.6)	(ASD M=ASD F) < (NT M=NT F)
	[45-180]	[62-145]	[73-147]	[83-156]	
Performance IQ[c]	106 (17.3)	102 (18.1)	109 (13.4)	109 (13.3)	(ASD M=ASD F) < (NT M=NT F)
	[59-149]	[53-148]	[62-147]	[79-145]	
ADI-R					
Social[d]	19.6 (5.4)	19.0 (6.4)	-	-	ns
	[4-30]	[0-30]			
Communication[e]	15.7 (4.6)	14.8 (5.3)	-	-	ns
	[2-25]	[0-25]			
RRB[e]	6.0 (2.5)	5.7 (2.5)	-	-	ns
	[0-13]	[0-12]			

	ASD M (N=418)	ASD F (N=95)	NT M (N=473)	NT F (N=218)	
	Mean (SD) [Range]	Mean (SD) [Range]	Mean (SD) [Range]	Mean (SD) [Range]	Post-hoc
ADOS-2					
Social-Affect[f]	9.2 (3.8) [2-20]	9.1 (3.5) [4-18]	-	-	ns
RRB[g]	2.9 (1.9) [0-8]	2.6 (1.5) [0-5]	-	-	ns
CSS total[h]	6.9 (2.2) [1-10]	6.8 (1.9) [2-10]	-	-	ns
Handedness	171 / 85.1% (R) 14 / 7% (L) 16 / 7.9% (A)	43 / 82.7% (R) 6 / 11.5% (L) 3 / 5.8% (A)	207 / 88.1% (R) 14 / 5.9% (L) 14 / 5 .9% (A)	128 / 93.4% (R) 5 / 3.6% (L) 4 / 3% (A)	(ASD M=ASD F) < (NT M>NT F)

Abbreviations: ASD=Autism Spectrum Disorder, NT=neurotypical, M=males, F=females, ADI-R= Autism Diagnostic Interview-Revised, ADOS=Autism Diagnostic Observation Schedule, RRB=restricted, repetitive behavior, CSS=calibrated severity score based on ADOS-2. [a]Full-Scale IQ information was available for 384 ASD M, 87 ASD F, 435 NT M and 197 NT F. [b]Verbal IQ information was available for 330 ASD M, 76 ASD F, 382 NT M and 159 NT F. [c]Performance IQ information was available for 335 ASD M, 78 ASD F, 398 NT M and 173 NT F. [d]ADI-R Social Scores were available for 341 ASD M and 74 ASD F. [e]ADI-R Communication and RRB Scores were available for 342 ASD M and 74 ASD F. [f]ADOS-Gotham Social-Affect Scores were available for 132 ASD M and 28 ASD F. [g]ADOS-Gotham RRB Scores were available for 134 ASD M and 29 ASD F. [h]ADOS-Gotham Severity Scores were available for 134 ASD M and 30 ASD F. Posthoc analyses were computed using ANOVA tests (or Chi-square tests in the case of categorical variables).

Table S3 – Differences in extreme deviations as a function of language delay, symptom severity, ADHD and handedness

LD	**LD vs. noLD**	**LD vs. NT**	**noLD vs. NT**
rightward deviations	$t_{(213)}=2.5$, $p=0.01$ sig	$t_{(152)}=4.6$, $p<0.001$ sig	$t_{(256)}=1.9$, $p=0.06$ ns
leftward deviations	$t_{(229)}=1.2$, $p=0.2$ ns	$t_{(182)}=4.4$, $p<0.001$ sig	$t_{(271)}=3.0$, $p=0.003$ sig
CSS	**high vs. low**	**high vs. NT**	**low vs. NT**
rightward deviations	$t_{(307)}=-1.3$, $p=0.2.$ ns	$t_{(303)}=4.0$, $p<0.001$ sig	$t_{(284)}=2.3$, $p=0.02$ sig
leftward deviations	$t_{(301)}=-1.2$, $p=0.2$ ns	$t_{(331)}=4.1$, $p<0.001$ sig	$t_{(293)}=2.4$, $p=0.02$ sig
ADHD	**ADHD + vs. ADHD-**	**ADHD+ vs. NT**	**ADHD- vs. NT**
rightward deviations	$t_{(298)}=1.7$, $p=0.08$ ns	$t_{(266)}=3.8$, $p<0.001$ sig	$t_{(297)}=1.7$, $p=0.09$ ns
leftward deviations	$t_{(297)}=2.0$, $p=0.04$ sig	$t_{(311)}=4.1$, $p<0.001$ sig	$t_{(313)}=1.6$, $p=0.1$ ns
Handedness	**ASD R vs. ASD nonR**	**ASD R vs. NT**	**ASD nonR vs. NT**
rightward deviations	$t_{(80)}=-1.5$, $p=0.14$ ns	$t_{(466)}=3.2$, $p=0.001$ sig	$t_{(72)}=2.9$, $p=0.005$ sig
leftward deviations	$t_{(98)}=0.4$, $p=0.71$ ns	$t_{(457)}=2.3$, $p=0.02$ sig	$t_{(90)}=2.0$,s=0.05 sig

Abbreviations: LD=language delay, noLD=no language delay, NT=neurotypical, CSS=calibrated severity score based on ADOS-2, ADHD=attention-deficit hyper-activity disorder, R=right-handed, nonR=non-right handed, ns=not significant, sig=significant (uncorrected).

Language delay was defined as having onset of first words later than 24 months and/or having onset of first phrases later than 33 months. Symptom severity was captured by the total ADOS-2 Calibrated Severity Score (CSS). We median-split the CSS measure into one group of individuals with autism with high CSS and one with low CSS. ADHD symptoms were assessed with the DSM-5 ADHD rating scale, covering both inattention and hyperactivity/impulsivity symptoms based on either self-or parent-report. A categorical variable was computed based on the DSM-5 criteria. Based on the short version of Edinburgh Handedness Inventory, a categorical variable comprising right-handed (+500 to +150) and non-right-handed (-500 to +149) individuals was computed. Group differences were computed using a t-test.