Normative age modelling of cortical thickness in autistic males

Authors: *Richard A. I Bethlehem[1,2,±], Jakob Seidlitz[1,3], Rafael Romero-Garcia[1],Guillaume Dumas[4,5,6] & Michael V. Lombardo [2,7]*

Affiliations:

[1]Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Cambridge CB2 0SZ, United Kingdom

[2]Autism Research Centre, Department of Psychiatry, University of Cambridge, Cambridge CB2 8AH, United Kingdom

[3] Developmental Neurogenomics Unit, National Institute of Mental Health, Bethesda, MD 20892, USA

[4] Institut Pasteur, Human Genetics and Cognitive Functions Unit, Paris, France

[5] CNRS UMR3571 Genes, Synapses and Cognition, Institut Pasteur, Paris, France

[6] University Paris Diderot, Sorbonne Paris Cité, Human Genetics and Cognitive Functions, Paris, France

[7] Laboratory for Autism and Neurodevelopmental Disorders, Center for Neuroscience and Cognitive Systems @UniTn, Istituto Italiano di Tecnologia, Rovereto, Italy

± Corresponding author:

Dr. Richard A.I. Bethlehem

Department of Psychiatry

Douglas House

18B Trumpington Road

CB5 8AH Cambridge

United Kingdom

rb643@medschl.cam.ac.uk

Abstract

Understanding heterogeneity in neural phenotypes is an important goal on the path to precision medicine for autism spectrum disorders (ASD). Age is a critically important variable in normal structural brain development and examining structural features with respect to age-related norms could help to explain ASD heterogeneity in neural phenotypes. Here we examined how cortical thickness (CT) in ASD can be parameterized as an individualized metric of deviance relative to typically-developing (TD) age-related norms. Across a large sample (n=870 per group) and wide age range (5-40 years), we applied a normative modelling approach that provides individualized whole-brain maps of age-related CT deviance in ASD. This approach isolates a subgroup of ASD individuals with highly age-deviant CT. The median prevalence of this ASD subgroup across all brain regions is 7.6%, and can reach as high as 10% for some brain regions. This work showcases an individualized approach for understanding ASD heterogeneity that could potentially further prioritize work on a subset of individuals with significant cortical pathophysiology represented in age-related CT deviance. Rather than cortical thickness pathology being a widespread characteristic of most ASD patients, only a small subset of ASD individuals are actually highly deviant relative to age-norms. These individuals drive small on-average effects from case-control comparisons. Rather than sticking to the diagnostic label of autism, future research should pivot to focus on isolating subsets of autism patients with highly deviant phenotypes and better understand the underlying mechanisms that drive those phenotypes.

Introduction

Autism spectrum disorder (ASD) is a clinical behavioural consensus label we give to a diverse collection of patients with social-communication difficulties and pronounced repetitive, restricted, and stereotyped behaviours (Lai et al., 2014). Beyond the single label of ASD, patients are in fact widely heterogeneous in phenotype, but also with regard to the diversity of different aetiologies (Lombardo et al., 2019). Even within mesoscopic levels of analysis such as examining brain endophenotypes, heterogeneity is the rule rather than the exception (Ecker, 2017). At the level of structural brain variation, neuroimaging studies have identified various neuroanatomical features that might help identify individuals with autism or reveal elements of a common underlying biology (Ecker, 2017). However, the vast neuroimaging literature is also considerably inconsistent, with reports of hypo- or hyper-connectivity, cortical thinning versus increased grey or white matter, brain overgrowth, arrested growth, etc., leaving stunted progress towards understanding mechanisms driving cortical pathophysiology in ASD.

Multiple explanations could be behind this inconsistency across the literature. Methodology widely differs across studies (e.g., low statistical power, different ways of estimating morphology or volume) and is likely a very important contributing factor (Haar et al., 2016; Vissers et al., 2012). Initiatives such as the Autism Brain Imaging Data Exchange (ABIDE; (Di Martino et al., 2017)) have made it possible to significantly boost sample size by pooling together data from several different studies. However, within-group heterogeneity in the autism population also immediately stands out as another factor obscuring consistency in the literature, particularly when the dominant approach of case-control models largely ignores heterogeneity within the ASD population. In particular, some autism-related heterogeneity reported in literature might be explained by factors such as age (Georgiades et al., 2017; Lord et al., 2015). Indeed, with regards to structural brain features of interest for study in ASD (e.g., volume, cortical thickness, surface area), these features change markedly over development (Mills et al., 2016; Raznahan et al., 2011a, 2011b; Smith et al., 2016). However, typical approaches towards dealing with age revolve around group statistical modelling of age as the variable of interest or removing age as a covariate and then parametrically modelling on-average differences between cases versus controls. While these are common approaches in the literature, they do not provide any individualized estimates of age-related deviance. In contrast, normative models of age-related variation may likely be an important alternative to these approaches and may mesh better with some conceptual views of deviance in ASD as being an extreme of typical population norms (Marquand et al., 2019, 2016). In contrast to the canonical case-control model, normative age modelling allows for computation of individualized metrics that can hone in on the precision information we are interested in – that is, deviance expressed in specific ASD individuals

relative to non-ASD norms. Such an approach may be a fruitful way forward in isolating individuals whom are ‘statistical outliers’. The reasons behind why these individuals are outliers relative to non-ASD norms may be of potential clinical and/or mechanistic importance. Indeed, if we are to move forward towards stratified psychiatry and precision medicine for ASD (Kapur et al., 2012), we must go beyond case-control approaches and employ dimensional approaches that can tell us information about which individuals are atypical and how or why they express such atypicality. Thus, this approach aims to provide more than a mere statistical advance, it aims to better conceptualize and capture personalized inferences that may ultimately result in more meaningful and targeted clinical inference.

In the present study, we employ normative modelling on age-related variability as a means to individualize our approach to isolate specific subsets of patients with very different neural features. Here we focus specifically on a neural feature of cortical morphology known as cortical thickness (CT). CT is a well-studied neuroanatomical feature thought to be differentially affected in autism and has received increasing attention in recent years (Jiao et al., 2011; Khundrakpam et al., 2017; Moradi et al., 2017; Smith et al., 2016; van Rooij et al., 2018; Zielinski et al., 2014). Recent work from our group also identified a genetic correlate for autism specific CT variation despite considerable heterogeneity in group specific CT in children with autism (Romero-Garcia et al., 2018). A study examining ABIDE I cohort data discovered case-control differences in CT, albeit very small in effect size (Haar et al., 2016). Similarly, the most recent and largest study to date, a mega-analysis combining data from ABIDE and the ENIGMA consortium, also indicated very small on-average case-control differences in cortical thickness restricted predominantly to areas of frontal and temporal cortices, indicating very subtle age-related between-group differences and substantial within-group age-related variability (van Rooij et al., 2018). Overall, these studies emphasize two general points of importance. First, age or developmental trajectory is extremely important (Courchesne et al., 2011, 2007; Georgiades et al., 2017; Schumann et al., 2010). Second, given the considerable within-group age-related variability and the presence of a large majority of null and/or very small between-group effects, rather than attempting to find on-average differences between all cases versus all controls, we should shift our focus to capitalize on this dimension of large age-related variability and isolate autism cases that are at the extremes of this dimension of normative variability.

Given our approach of age-related normative CT modelling, we first compare the utility of age-related normative modelling directly against more traditional case-control models. We then describe the prevalence of ASD cases that show significant age-related deviance in CT (i.e. > 2 SD from age-related norms) and show how a metric of continuous variability in age-related deviance in CT is expressed across the cortex in autism. Finally, we identify age-deviant CT-

behaviour associations and assess whether such dimensional analyses associated with behaviour identify similar or different regions than typical case-control analyses. To show applicability of this approach we also applied the same method to other measures of neuroanatomy; gyrification, volume and surface area. Results and analyses of these metrics can be found in the supplementary materials and all code and data used are available on GitHub (Bethlehem et al., 2018).

Methods

Participants

In this study, we sought to leverage large neuroimaging datasets to yield greater statistical power for identifying subtle effects. To achieve this, we utilized the ABIDE datasets (ABIDE I and II; (Di Martino et al., 2017)) (see Supplementary Table S1 and S2 for full list of sites used in the current analyses). Given that the normalized modelling approach gives us individual level measures we chose to also include sites with limited numbers of subjects. Groups were subsequently matched on age using the non-parametric nearest neighbour matching procedure implemented in the Matchit package in R (https://cran.r-project.org/web/packages/MatchIt/index.html) (Ho et al., 2011). After matching case and control groups and excluding scans of poorer quality (see supplementary materials) we were left with a sample size N=870 per group (Table 1 and 2). Because of power limitations in past work with small samples, we conducted an a priori statistical power analysis indicating that a minimum case-control effect size of d = 0.1752 could be detected at this sample size with 80% power at a conservative alpha set to 0.005 (Benjamin et al., 2018). For correlational analyses looking at brain-behaviour associations, we examined a subset of patients with the data from the SRS (N_{autism_male} = 421) and ADOS total scores (N_{autism_male} = 505). With same power and alpha levels the minimum effect for SRS is r = 0.1765 and r = 0.1651 for the ADOS.

Table 1: Sample characteristics of Age

Dx	Sex	Mean	SD	N	Median	Min	Max
Autism	Male	16.32	9.09	754	13.75	5.13	64
Autism	Female	15.06	8.43	116	12.57	5.22	54
NT	Male	16.64	8.98	660	13.69	5.89	64
NT	Female	13.25	5.33	210	11.09	5.91	32

Table 2: Sample characteristics

Measure	Dx	Sex	Mean	SD	N	Median
IQ	Autism	Male	106.13	16.51	754	107

	Autism	Female	105.88	16.21	116	106.5
	NT	Male	111.28	12.13	660	111
	NT	Female	112.07	13.21	210	112
ADOS	Autism	Male	11.15	3.86	505	11
	Autism	Female	11.41	3.9	63	11
	Control	Male	1.55	1.58	38	1
	Control	Female	3	1.05	10	3
SRS	Autism	Male	80.42	21.41	421	77
	Autism	Female	85.95	22.07	61	88
	Control	Male	38.43	15.25	337	41
	Control	Female	39.93	12.13	120	42

Imaging processing and quantification

Cortical surface reconstruction was performed using the MPRAGE (T1) image of each participant with FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) version (v5.3.0, to ensure comparability with previous ABIDE publications). The reconstruction pipeline performed by FreeSurfer "recon-all" involved intensity normalization, registration to Talairach space, skull stripping, WM segmentation, tessellation of the WM boundary, and automatic correction of topological defects. Briefly, non-uniformity intensity correction algorithms were applied before skull stripping (Ségonne et al., 2004), resulting in resampled isotropic images of 1mm. An initial segmentation of the white matter tissue was performed to generate a tessellated representation of the WM/GM boundary. The resulting surface was deformed outwards to the volume that maximize the intensity contrast between GM and cerebrospinal fluid, generating the pial surface (Dale et al., 1999). Resulting surfaces were constrained to a spherical topology and corrected for geometrical and topological abnormalities. Cortical thickness of each vertex was defined as the shortest distance between vertices of the GM/WM boundary and the pial surface (Fischl and Dale, 2000). We chose to not conduct manual segmentations and excluded failed subjects from any subsequent analysis (and these subjects were removed prior to the matching and QC procedures). To assess quality of Freesurfer reconstructions we computed the Euler index (Rosen et al., 2018). The Euler number is a quantitative proxy index of segmentation quality and has shown high overlap with manual quality control labelling (Rosen et al., 2018). The index counts the number of times the freesurfer has had to interpolate surface gaps during the reconstruction to ensure a continuous outcome surface. As such the index is effectively a measure for the reliability of the surface reconstruction and the resulting CT estimates. In the full sample we found a small but significant difference in both hemispheres (Figure S2) with the

autism group having overall slightly worse scan quality (d = 0.176 and d = 0.187 for left and right hemisphere respectively). Therefore, we chose to exclude subjects with an extreme Euler index of 300 or higher in either hemisphere (corresponding to approximately the top 10%) and included the index itself as a confound variable in all models to account for potential differences in the reliability of the freesurfer reconstruction across groups and sites.

Across both ABIDE I and ABIDE II cortical thickness was extracted for each subject using two different parcellations schemes: an approximately equally-sized parcellation of 308 regions (~500mm^2 each parcel) (Romero-garcia et al., 2012) and a parcellation of 360 regions derived from multi-modal features extracted from the Human Connectome Project (HCP) dataset (Glasser et al., 2016). The 308-region parcellation was constructed in the FreeSurfer fsaverage template by subdividing the 68 regions defined in the Desikan-Killiany atlas (Desikan et al., 2006). Thus, each of the 68 regions was sequentially sub-parcellated by a backtracking algorithm into regions of ~500mm^2, resulting in a high resolution parcellation that preserved the original anatomical boundaries defined in the original atlas (Romero-garcia et al., 2012). Surface reconstructions of each individual were co-registered to the fsaverage subject. The inverse transformation was used to map both parcellation schemes into the native space of each participant.

Statistical analyses

There are likely many variables that contribute to variability in CT between individuals and across the brain. In order to visually assess the contribution of some prominent sources of variance we adopted a visualization framework derived from gene expression analysis (http://bioconductor.org/packages/variancePartition) (Hoffman and Schadt, 2016) and included the most commonly available covariates in the ABIDE dataset: Age, Sex, Diagnosis, Scanner Site, Full-scale IQ, Verbal IQ, Handedness and SRS. Given that ABIDE was not designed as an integrated dataset from the outset, it seems plausible that scanner site might be related to autism or autism-related variables (e.g., some sites might have different case-control ratios or only recruited specific subgroups). Figure 1 shows the ranked contribution of those covariates. Perhaps unsurprisingly, scanner site and age proved to be the most dominant sources of variance (each explaining on average around 15% of the total variance). Our initial conventional analysis was aimed to delineate potential broad case-control differences, as has been done in previous studies (Haar et al., 2016; van Rooij et al., 2018). We used a linear mixed effects model with scanner site as random effect. Given the potentially strong contribution of age we chose to include this as fixed effects covariates in the model. Multiple comparison correction was implemented with Benjamini-Hochberg FDR at $q<0.05$ (Benjamini and Hochberg, 1995).

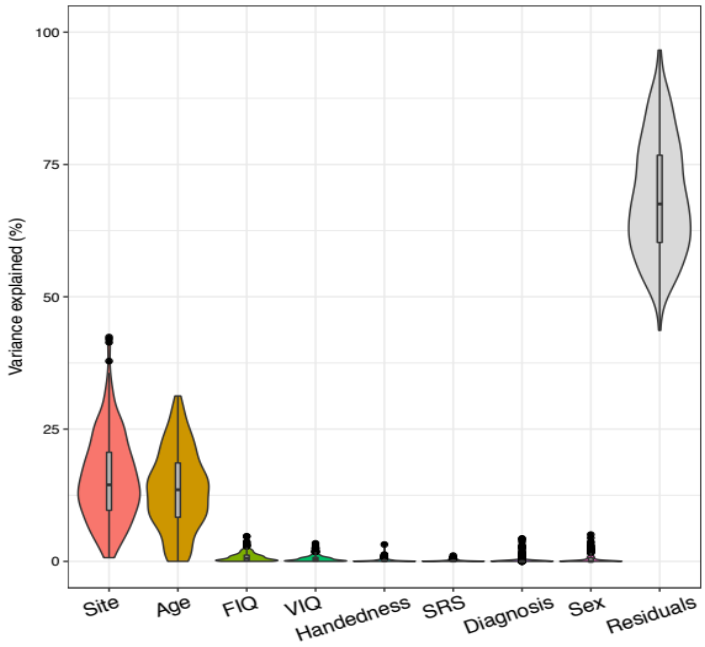

Figure 1: Explained variance in cortical thickness for each covariate. (Age: Age at the time of scanning; FIQ: Functional Intelligent Quotient; VIQ: Verbal Intelligent Quotient ; SRS: total score of the Social Responsive Scale ; Diagnosis : diagnostic group, i.e. ASD or NT).

Age-related normative modelling

Normative modelling of age-related CT effects was done utilizing data from the typically-developing group (TD) (see Figure 2 for a schematic overview). We used a local polynomial regression fitting procedure (LOESS) (Cleveland et al., 1988; Lefebvre et al., 2018), where the local width or smoothing kernel of the regression was determined by the model that provided the overall smallest sum of squared errors. Though we also assessed consistency of our output using centiles scoring (see Supplementary Materials). To align the TD and ASD groups, both were binned into one year age bins. For each age bin and every brain region we computed a normative mean and standard deviation from the TD group. This was done separately for each sex, given known sex differential developmental trajectories. These statistical norms were then used to compute a w-score (analogous to a z-score) for every individual with autism. The w-score for an individual reflects how far away their CT is from TD norms in units of standard deviation. Because w-scores are computed for every brain region, we get a w-score map for each ASD participant showing how each brain region for that individual is atypical relative to TD norms. Age bins that contained fewer than 2 data-points in the TD group were excluded from subsequent analysis as the standard deviations for these bins would essentially be zero (and thus the w-score could not be computed). The characteristics of the final autism sample are listed in table 3.

Table 3: Sample characteristics of autism group after normative modelling selection

Dx	Mean	SD	N	Median	Min	Max
Autism	14.93	6.03	714	13.37	5.53	39.2
NT	15.43	6.46	636	13.37	5.89	39.4

To assess the reliability of this w-score we bootstrapped the normative sample (1000 bootstraps, with replacement) and computed 1000 bootstrapped w-scores for each individual and each brain region. To subsequently quantify the reliability of the w-score we computed an FDR corrected analogous p-value for each subject by computing the absolute position of the real w-score in the distribution of bootstrapped w-scores. The rationale being that if a real w-score would be in the top 5% of the bootstrapped distribution it would likely not be a reliable score (e.g. the score would be influenced by only a small subset of the normative data). The median number of brain regions per subject with a significant p-value was 1 (out of 308), indicating that the w-score provides a robust measure of atypicality. More details on the bootstrapping procedure are provided in the supplementary material (SI: boostrapping and SI figure S3).

Because w-score maps are computed for each individual, we ran hypothesis tests at each brain region to identify regions that show on-average non-zero w-scores stratified by sex (FDR corrected at $q<0.05$). To assess the effect of age-related individual outliers on the global case-control differences we re-ran the hypotheses tests on w-scores after removing region-wise individual outliers (based on a 2SD cut-off). Unfortunately, despite a significant female sub-group, the age-wise binning greatly reduced the number of bins with enough data-points in the female group. Given the reduced sample size in the female group and the known interaction between autism and biological sex (Lai et al., 2015a, 2015b), we conducted normative modelling on the male group only.

To assess the distribution in the normative group we also conducted one-sample linear mixed effects modelling in the normative group only to determine if any of all brain regions would show outlier consistency. There were no brain regions for which the w-score showed a deviation significant from zero in the normative group (even without correcting for multiple comparisons across all brain regions).

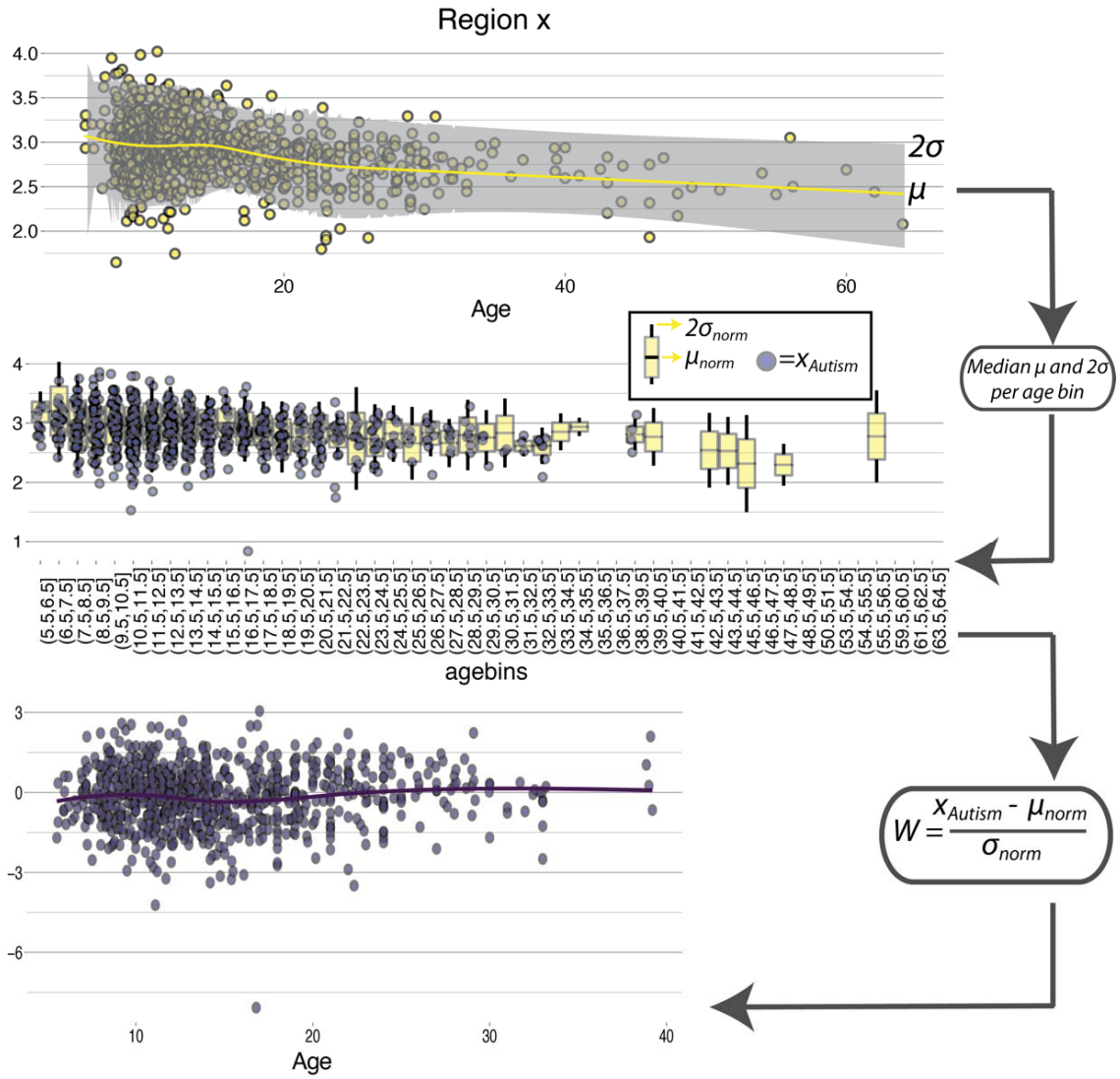

Figure 2: Schematic overview of normative modelling. In first instance LOESS regression is used to estimate the developmental trajectory on CT for every individual brain region to obtain an age specific mean and standard deviation. Then we computed median for each one-year age-bin for these mean and median neurotypical estimates to align them with the ASD group. Next, for each individual with autism and each brain region the normative mean and standard deviation are used to compute a w-score relative to their neurotypical age-bin. Contrary to conventional boxplots, the second panel shows mean, 1sd and 2sd for the neurotypical group (in yellow) and individuals with an autism diagnosis in purple.

To isolate subsets of individuals with significant age-related CT deviance, we used a cut-off score of 2 standard deviations (i.e. w >= 2 or w<=2). This cut-off allows us to isolate specific ASD patients with markedly abnormal CT relative to age-norms for each individual brain region. We then calculated sample prevalence (percentage of all ASD patients with atypical w-scores), in order to describe how frequent such individuals are in the ASD population and for each brain region individually. A sample prevalence map can then be computed to show the frequency of these patients across each brain region. We also wanted to assess how many patients have

markedly atypical w-scores (beyond 2SD) across a majority of brain regions. This was achieved by computing an individual global w-score ratio as follows:

$$gW = \frac{\sum |(w)| > 2}{\sum |(w)| < 2}$$

We also computed global w-score ratios for positive and negative w regions separately.

Behavioural analyses

In addition to assessing the effect of normative outlier on conventional case-control analyses we also conducted some exploratory analysis on the normative w-scores. First, to explore whether the w-scores reflect a potentially meaningful phenotypic feature we also computed Spearman correlations for each brain region between the most commonly shared phenotypic features in ABIDE: ADOS, SRS, SCQ, AQ, FIQ and Age. Resulting p-values matrices were corrected for multiple comparisons using Benjamini-Hochberg FDR correction and only regions surviving and FDR corrected p-value of $p < 0.05$ are reported.

Finally, we explored whether the raw CT values could be used in a multivariate fashion to separate groups by diagnosis or illuminate stratification within ASD into subgroups. Here we used k-medoid clustering on t-Distributed Stochastic Neighbour Embedding (tSNE) (van der Maaten, 2014). Barnes-Hut tSNE was used to construct a 2-dimensional embedding for all parcels in order to be able to run k-medoid clustering in a 2D representation and in order to visually assess the most likely scenario within the framework suggested by Marquand and colleagues (Marquand et al., 2016). Next, we performed partitioning around medoids (PAM), estimating the optimum number of clusters using the optimum average silhouette width (Hennig and Liao, 2013). Details of this exploratory analysis are reported in the supplementary materials.

Data and code availability

Data and code are available on GitHub (Bethlehem et al., 2018), Cohen's d were computed using: https://github.com/mvlombardo/utils/blob/master/cohens_d.R and the centiles cross-validation code can be found on https://github.com/deep-introspection/PyNM.

Results

Case-control differences versus age-related normative modelling

Our first analysis examined conventional case-control differences. As expected from prior papers utilizing large-scale datasets for case-control analysis (e.g., (Haar et al., 2016; van Rooij et al., 2018)), a small subset of regions (12%, 38/308 regions) pass FDR correction. Of these regions, most are of small effect size, with 34 of the detected 38 regions showing an effect less than 0.2 standard deviations of difference (Figure 3A). We suspected that such small effects could be largely driven by a few ASD patients (Byrge et al., 2015) with highly age-deviant CT. Because we also had computed w-scores from our normative age-modelling approach, we identified specific 'statistical outlier' patients for each individual region with w-scores > 2 standard deviations from typical norms and excluded them from the case-control analysis. This analysis guards against the influence of these extreme outliers, and if there are true on-average differences in ASD, the removal of these outlier patient*regions should have little effect on our ability to detect case-control differences. However, removal of outlier patients now revealed only 16 regions passing FDR correction instead of 38 regions with small case-control differences - a 2.3-fold decrease in the number of regions detected. Indeed, the majority of case-control differences identifying small on-average effects were primarily driven by this small subset of highly-deviant patients (Figure 3B). These remaining 16 regions with small on-average effects were restricted to areas near posterior cingulate cortex, temporo-parietal cortex and areas of visual cortex.

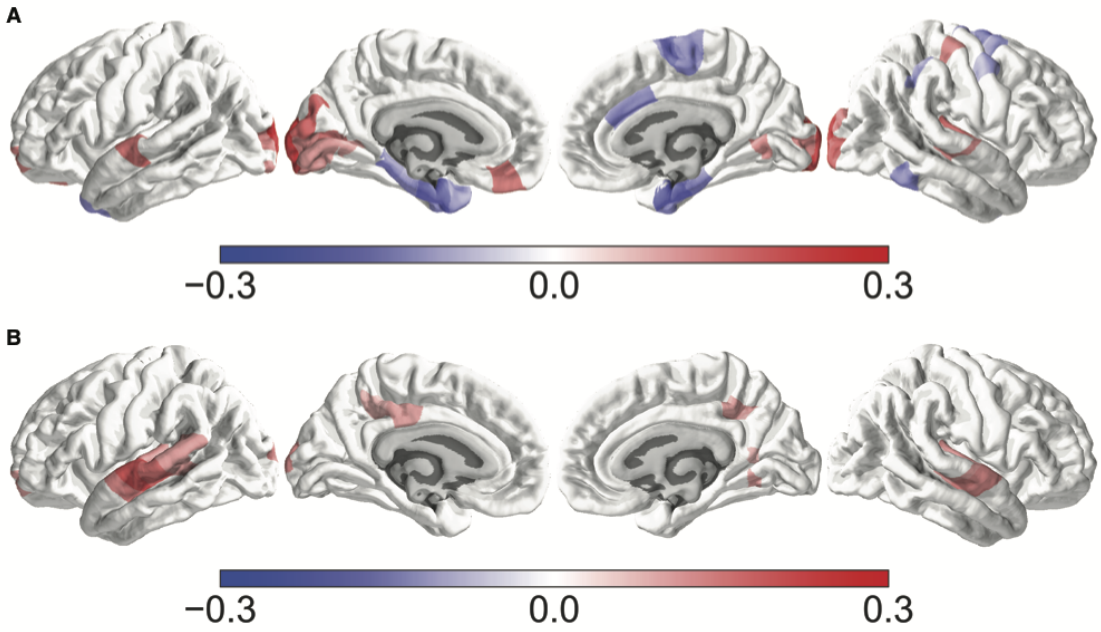

Figure 3: Case control difference analysis with linear mixed effect model. Panel A shows effect sizes for regions passing FDR correction for linear mixed effect modelling of conventional case control difference analysis. Cohen's d values represent ASD – Control, thus blue denotes ASD<Control and red denotes ASD>Control. Panel B shows effect sizes for regions passing FDR correction after outlier removal for the same linear mixed effect modelling of conventional case control difference analysis.

In contrast to a canonical case-control model, we computed normative models of age which resulted in individualized w-scores that indicate how deviant CT is for an individual compared to typical norms for that age. This modelling approach allows for computation of w-scores for every region and in every patient, thus resulting in a w-score map that can then itself be tested for differences from a null hypothesis of w-score = 0, indicating no significant on-average ASD deviance in age-normed CT. These hypothesis tests on normative w-score maps revealed no regions surviving FDR correction.

Isolating ASD individuals with significant age-related CT deviance

While the normative modelling approach can be sensitive to different pathology than traditional case-control models, another strength of the approach is the ability to isolate individuals expressing highly significant CT-deviance. We operationalized 'significant' deviance in statistical terms as w-scores greater than 2SD away from TD norms. By applying this cut-off, we can then describe what proportion of the ASD population falls into this CT subgroup category for each individual brain region. Over all brain regions the median prevalence of these patients is around 7.6% (Figure 4). This prevalence estimation is much higher than the expected 4.55%

prevalence one would expect by chance for greater than 2 standard deviations of difference. The distribution of prevalence across brain regions also has a positive tail indicating that for a small number of brain regions the prevalence can jump up to more than 10%. Expressed back into sample size numbers, if 10% of the ASD population had significant CT abnormalities, with a sample size of n=754, this means that n=75 patients possess such significant issues. Underscoring the prevalence of these significant cases is important since as shown earlier, it is likely that primarily these 'statistical outlier' patients drive most of the tiny case-control differences observed.

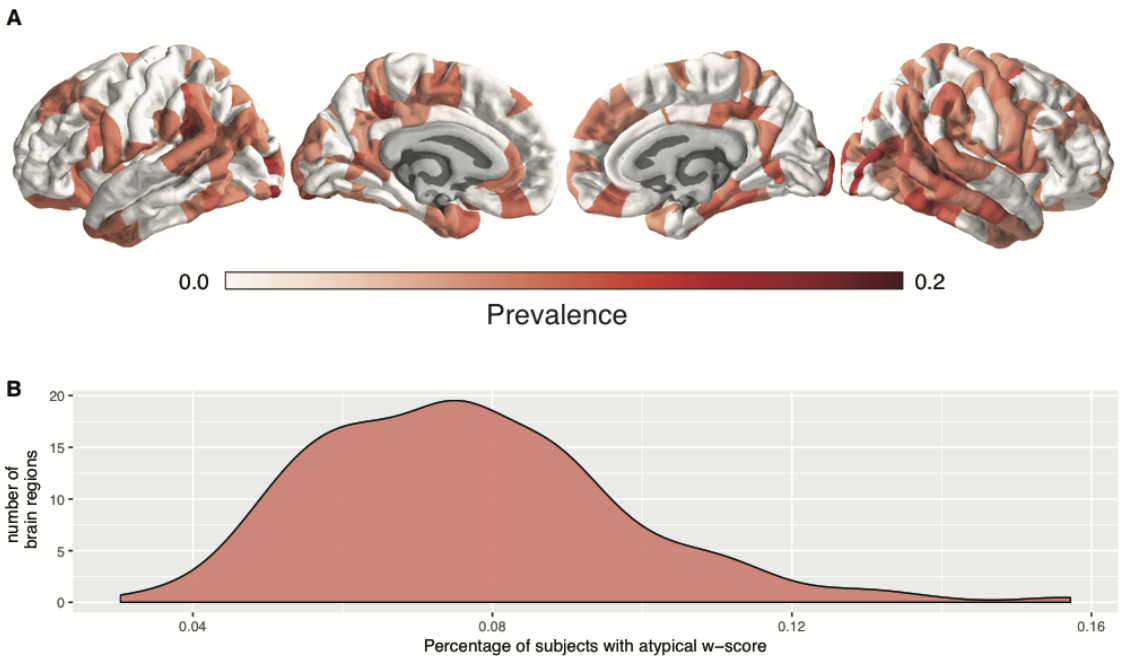

Figure 4: Region specific prevalence of atypical w-scores. Panel A shows the by region prevalence of individuals with a w-score of greater than +/-2SD. For visualization purposes these images are thresholded at the median prevalence of 0.076. Panel B shows the overall distribution of prevalence across all brain regions.

There are other interesting attributes about this subset of brain regions. With regard to age, these patients were almost always in the age range of 6-20, and were much less prevalent beyond age 20 (S5). The median age of outliers across brain regions ranged from [10.6 – 20.2] years old, with an overall skewed distribution towards the younger end of the spectrum (supplementary Figure S6), showing that CT deviance potentially normalizes with increasing age in ASD, though it should be noted that this may partially be explained by the overall skewed age distribution in the dataset.

Patients with significant CT deviance were also largely those that expressed such deviance within specific brain regions and were not primarily subjects with globally deviant CT. To show this we computed a w-score ratio across brain regions that helps us isolate patients that show

globally atypical CT deviance across most brain regions. The small number of patients with a ratio indicating a global difference (ratio > 0.5, n=14) were those that had globally thinner cortices. This small subset of individuals was much smaller than the number of region-wise ourtiers as shown in Figure 4. Upon visual inspection of the raw data for these participants, it is clear that the global thinning effect is not likely a true biological difference but rather one driven by the quality of the raw images, even though the Euler index did not indicate failure in reconstruction. Unfortunately, we did not have enough complete phenotypic data on these subjects to warrant further in-depth phenotypic analysis.

Exploratory analysis of brain-behaviour relationships

An additional advantage of the use of normative modelling over the traditional case-control modelling is that we can use the individualized deviation as a novel metric for finding associations with phenotypic features. Here we used w-scores to compute Spearman correlations for the most commonly shared phenotypic features in the ABIDE dataset: ADOS, SRS, SCQ, AQ, FIQ and Age. After correcting for multiple comparisons across phenotype and region (6 phenotypic measures * 308 regions = 1848 tests) we identified a number of brain regions that survive multiple comparison corrections for the SRS and ADOS scores (figure 5). SRS is associated with w-scores primarily in areas of lateral frontal and parietal cortex, while ADOS is associated with w-scores primarily in lateral and inferior temporal cortex. Notably, these regions are largely different from regions that appear to show on-average differentiation in case-control and w-score analyses.

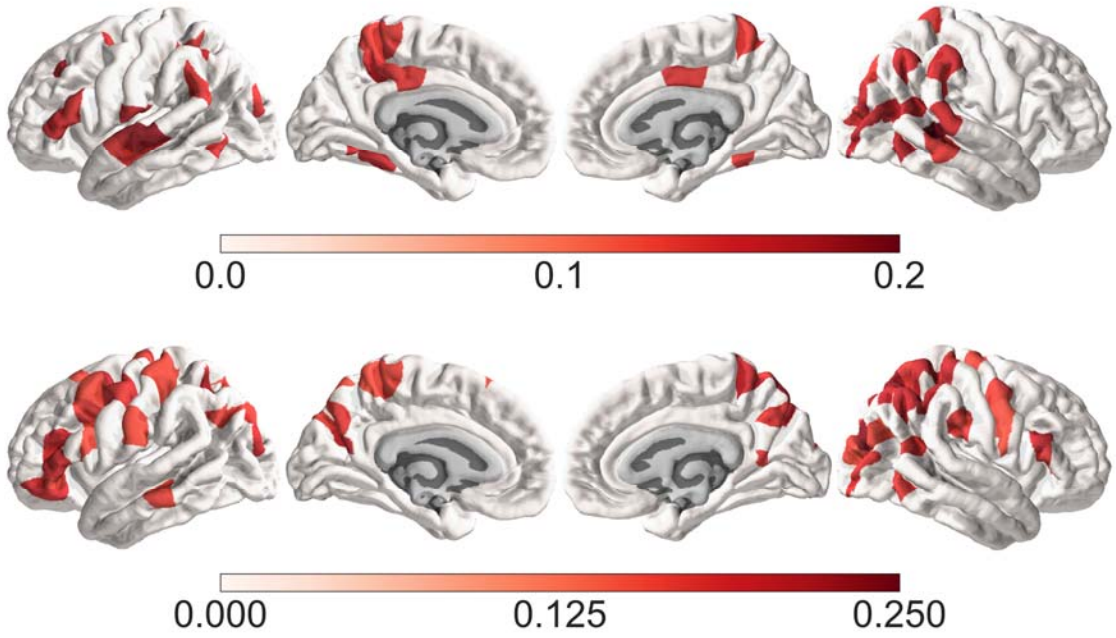

Figure 5: Phenotype – W-Score correlations
Spearman correlations between ADOS and w-score in the top panel. The lower panel shows the same for the SRS.

Discussion

In the present study, we find that with a highly-powered dataset, conventional case-control analyses reveal small differences in cortical thickness in autism and are restricted to a small subset of regions. In general, this idea about subtle effect sizes for case-control comparisons is compatible with other recent papers utilizing partially overlapping data — Haar and colleagues utilized only ABIDE I data (Haar et al., 2016), while van Rooji and colleagues (van Rooij et al., 2018) utilized both ABIDE I and II dataset combined with further data from the ENIGMA consortium. While these statements about small effect sizes are not novel, our findings suggest that even these small effect sizes may be misleading and over-optimistic. Utilizing normative modelling as a way of identifying and removing CT-deviant outlier patients, we find that most small case-control differences are driven by a small subgroup of patients with high CT-deviance for their age. In contrast, we further showed that analysis of CT-normed scores (i.e. w-scores) themselves reveals a completely different set of regions that are on-average atypical in ASD. The directionality of such differences also reverses in some cases. For instance, Haar and colleagues discovered that areas of visual cortex are thicker in ASD compared to TD in ABIDE I (Haar et al., 2016). Our case-control analyses here largely mirror that finding. However, re-analysis after w-score outlier removal totally removes the effects previously reported in visual cortex. Thus, here is a clear case whereby our normative age modelling approach identifies effects that are likely driven by only a small subset of individuals.

The revelation of new insights via normative age modelling, alongside cleaning up interpretations behind case-control models, both highlight the significant utility of such approach. The presence of small region-dependent outlier effects in ASD misleadingly drives on-average inferences from case-control models. Thus, it is important for the field to better understand how prevalent these outlier individuals are for a given brain region (i.e. our analyses did not reveal a consistent brain region or group of individuals with spatially overlapping patterns of extreme w-scores). With our normative modelling approach, we were able to quantify the overall prevalence of this CT outlier ‘subgroup’ at a median prevalence of 7.6%. This means that across the brain, a median level of 7.6% of individuals have an extreme w-score. This estimate much larger than the expected 4.55% for standard deviations greater than 2. However, considerable heterogeneity exists across brain regions, as a small proportion of brain regions are even more enriched for the CT outlier ‘subgroup’ and can contain greater than 10% prevalence.

We also noted that this CT outlier subgroup was predominantly restricted to the childhood to early adult age range. In later adult ages, the prevalence of this subgroup drops off. This could be a potential indicator that highly atypical CT is more prevalent and detectable at earlier ages. However, it should be noted that this observation may partially be explained by the overall skewed age distribution in the overall dataset. It will be important to assess even earlier age ranges such as the first years of life (Courchesne et al., 2011), as well as later adult years when significant aging processes begin to take effect (Happé and Charlton, 2012). Future studies with either more balanced developmental sample or samples that cover the entire lifespan will be better positioned to confirm this age-related skewed profile in 'atypical' brain regions.

In addition, we also identify a very small group of individuals that have atypical patterns in over 50% of brain regions. Unfortunately, not much behavioural or phenotypic information was available for this sub-group. We hope that future studies will obtain more detailed phenotypic information in order to delineate more precisely what the clinical and or more broad behavioural implications might be of this atypicality. Furthermore, it is clear from the present work that this subgroup only covers a very small subset in the autism population and thus future studies will require large sample sizes to be able to identify this subgroup. However, mirroring work in autism genetics, whereby discoveries are continually being made regarding very small proportions of the ASD population being explained by highly penetrant genetic mechanisms (Geschwind and State, 2015), it also may be the case that such individuals with highly age-deviant CT are individuals with specific highly penetrant biological mechanisms underlying them, and possibly related to neurogenesis and other factors that are implicated in CT changes (Romero-Garcia et al., 2018). With animal models of highly penetrant genetic mechanisms linked to autism, it is notable that such mechanisms have heterogeneous effects on brain volume (Ellegood et al., 2014). Thus, the fact that this is only a small subset need not be an obstacle for the discovery of core biological mechanisms. Ideally, future studies will also collect detailed genetic and/or other biological information in order to probe the core biological aetiology underlying the pattern of broad atypical cortical thickness.

We also conducted exploratory analyses to relate the w-scores back to phenotypic information more broadly, in so far as this was available in ABIDE. Here, we find collections of areas that are largely different from regions normally detected with on-average case-control or on-average non-zero w-score differences. Interestingly, the associations with ADOS and SRS show somewhat differential spatial topography which may suggest that the overall scores are related to different underlying neurobiological mechanisms. Overall, these results could suggest that the normative model also picks up signal related to behavioural variability. However, it should be emphasized that ADOS and SRS scores are not available for the full dataset and the reported

effects were small and should thus be considered exploratory. Based on present results however we expect future studies, with more comprehensive phenotypic information such as EU-AIMS2-trials (https://www.aims-2-trials.eu), to be able to confirm this brain-behaviour relationship. It will be interesting to see whether in a larger more comprehensive sample the same topological dissociation becomes apparent as well.

The current results can be contrasted with a recent study on the EU-AIMS LEAP cohort (Zabihi et al., 2019). This study differs from the current work in being based on a completely independent dataset (EU-AIMS LEAP vs ABIDE). The studies also differ in how normative models are estimated - LOESS and centiles vs Gaussian process regression. This study applied normative modeling only to males to reduce sex-related heterogeneity whereas Zahibi et al., utilized both males and females and used sex as a factor in the model. The current study also utilizes a larger sample size (autism n=754, NT n=660; autism n=321, NT n=206 in (Zabihi et al., 2019)). Despite these differences, some important consistencies emerge. In particular, our map of prevalence of the CT outlier group (Figure 4) is somewhat consistent with the spatial topology Zahibi and colleagues report for negative deviations from the normative model (e.g., Figure 4 of (Zabihi et al., 2019)). Furthermore, while our analyses of brain-behavioral relationships is limited, there is some consistency across this study and Zahibi et al., with the correlation between ADOS total scores and left inferior frontal gyrus. Thus, despite the methodological differences, the overall consistency suggests that many of the inferences from these works generalize to the autism population.

There are a number of caveats to consider in the present study. First and foremost, the present data are cross-sectional and the normative age modelling approach cannot make claims about trajectories at an individual level. With longitudinal data, this normative modelling approach could be extended. However, at the moment the classifications of highly age-deviant CT individuals are limited to static normative statistics within discrete age-bins rather than based on statistics from robust normative trajectories. The dataset also represents ASD within an age range that misses very early developmental and also very late adulthood periods. The dataset also presents a post-hoc collection of sites accumulated through the ABIDE initiative, whereby scanners, imaging acquisition sequences and parameters, sample ascertainment, etc, are highly heterogeneous. As a result, we observed that site had a large effect on explaining variance in CT and this is compatible with observations made by other studies (Haar et al., 2016). Furthermore, it is likely that there may be systematic interactions between scanner site and some variables of interest such as age (e.g. different scanning sites will likely have recruited specific age cohorts). Finally, there are a number of different approaches to normative modelling that all have pros and cons (see (Marquand et al., 2019) for an excellent review). We

chose to use LOESS estimation as its computationally efficient and the resulting w-scores are easily interpretable. However, since its based on estimation of standard deviation from a normative sample it is potentially sensitive to small samples in a given age-bin (e.g. if there are only 4 data-points for a given age-bin there is likely to be a less reliable sd). Hence in situations where data is spare the LOESS approach may allow for less reliable normative scores. In order to assess the sensitivity of our approach in the present data we implemented the aforementioned bootstrapping procedure to identify robustness of outlier detection. In addition, we also conducted a centiles estimation that is relatively standard in for example epidemiology (Visser et al., 2009), similar to quantile rank maps (Chen et al., 2015) and arguably less sensitive to small sample uncertainty. Both approaches showed high significant correlation in determining whole-brain w-score ratios (r=0.87, p=4e-119 and r=0.66, p=5.7e-39 for ABIDE I and ABIDE II respectively; see supplementary materials).

In conclusion, the present study shows how normative age modelling approach in ASD questions our interpretation of conventional case-control modelling by shedding significant new insight into heterogeneity in ASD. We show that results from case-control analyses, even within large datasets, can be highly susceptible to the influence of 'outlier' subjects. Removing these outlier subjects from analyses can considerably clean up the inferences being made about on-average differences that apply to a majority of the ASD population. Rather than only being nuisances for standard group-level analyses, these outlier patients are significant in their own light, and can be identified with our normative age modelling approach. Normative models may provide an alternative to case-control models that test hypotheses at a group-level, by allowing additional insight to be made at more individualized levels, and thus help further progress towards precision medicine for ASD. Furthermore, the current approach is in line with the original normative modelling approach advocated by Marquand and colleagues (Marquand et al., 2016) which suggested the development of methods to move away from the traditional case-vs-control analyses. Normative modelling was originally proposed as one solution among others like stratification. Here, a clear path forward would be to combine both, for instance by using output of normative model as features used in the participant stratification, thus avoiding trivial clustering caused by confounding factors. In the present work we show that normative modelling is more than a purely statistical advancement to improve robustness. It allows us to identify a small subgroup that we expect to have strong relevance for the discovery of core biological or phenotypic clinical targets. It allowed us to explore brain-behaviour relationships that reveal differential spatial topology for ADOS and SRS scores. More importantly however, it moves us conceptually closer to making personalized inferences that can go beyond group determined diagnostic labels.

Acknowledgments

This work was supported by a European Research Council (ERC) Starting Grant (755816; AUTISMS) awarded to MVL. RRG was funded by the Guarantors of Brain. JS was funded by the National Institutes of Health Oxford-Cambridge Scholars Program. RAIB was funded by a British Academy Post-Doctoral Fellowship and Autism Research Trust.

Author Contributions

RAIB, RRG, JS and MVL designed the experiment. RAIB, RRG, JS, GD and MVL conceived and implemented all analyses. RAIB, RRG, JS, GD and MVL wrote the manuscript.

Financial Disclosures

None of the authors have any financial interests to declare.

References

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Hua Ho T, Hoijtink H, Hruschka DJ, Imai K, Imbens G, Ioannidis JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schönbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J, Johnson VE. 2018. Redefine statistical significance. *Nature Human Behaviour* **2**:6–10.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol* **57**:289–300.

Bethlehem RAI, Seidlitz J, Romero-Garcia R, Dumas G, Lombardo MV. 2018. Normative age modelling of cortical thickness in autistic males. doi:10.5281/ZENODO.1325171

Byrge L, Dubois J, Tyszka JM, Adolphs R, Kennedy DP. 2015. Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *J Neurosci* **35**:5837–5850.

Chen H, Kelly C, Xavier Castellanos F, He Y, Zuo X-N, Reiss PT. 2015. Quantile rank maps: A new tool for understanding individual brain development. *Neuroimage* **111**:454–463.

Cleveland WS, Devlin SJ, Grosse E. 1988. Regression by local fitting: Methods, properties, and computational algorithms. *J Econom* **37**:87–114.

Courchesne E, Campbell K, Solso S. 2011. Brain growth across the life span in autism: Age-specific changes in anatomical pathology. *Brain Res* **1380**:138–145.

Courchesne E, Pierce K, Schumann CM, Redcay E, Buckwalter JA, Kennedy DP, Morgan J. 2007. Mapping early brain development in autism. *Neuron* **56**:399–413.

Dale AM, Fischl B, Sereno MI. 1999. Cortical Surface-Based Analysis. *Neuroimage* **9**:179–194.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**:968–980.

Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, Balsters JH, Baxter L, Beggiato A, Bernaerts S, Blanken LME, Bookheimer SY, Braden BB, Byrge L, Castellanos FX, Dapretto M, Delorme R, Fair DA, Fishman I, Fitzgerald J, Gallagher L, Keehn RJJ, Kennedy DP, Lainhart JE, Luna B, Mostofsky SH, Müller RA, Nebel MB, Nigg JT, O'Hearn K, Solomon M, Toro R, Vaidya CJ, Wenderoth N, White T, Craddock RC, Lord C, Leventhal B, Milham MP. 2017. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data* **4**:170010.

Ecker C. 2017. The neuroanatomy of autism spectrum disorder: An overview of structural neuroimaging findings and their translatability to the clinical setting. *Autism* **21**:18–28.

Ellegood J, Markx S, Lerch JP, Steadman PE, Genç C, Provenzano F, Kushner SA, Henkelman RM, Karayiorgou M, Gogos JA. 2014. A highly specific pattern of volumetric brain changes due to 22q11.2 deletions in both mice and humans. *Mol Psychiatry* **19**:6–6.

Fischl B, Dale AM. 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences* **97**:11050–11055.

Georgiades S, Bishop SL, Frazier T. 2017. Editorial Perspective: Longitudinal research in autism - introducing the concept of "chronogeneity." *J Child Psychol Psychiatry* **58**:634–636.

Geschwind DH, State MW. 2015. Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet Neurol* **14**:1109–1120.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC. 2016. A multi-modal parcellation of human cerebral cortex. *Nature* **536**:171–178.
Haar S, Berman S, Behrmann M, Dinstein I. 2016. Anatomical Abnormalities in Autism? *Cereb Cortex* **26**:1440–1452.
Happé F, Charlton RA. 2012. Aging in autism spectrum disorders: a mini-review. *Gerontology* **58**:70–78.
Hennig C, Liao TF. 2013. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *J R Stat Soc Ser C Appl Stat* **62**:309–369.
Ho DE, Imai K, King G, Stuart EA. 2011. Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Softw* **42**:1–28.
Hoffman GE, Schadt EE. 2016. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**:483.
Jiao Y, Chen R, Ke X, Chu K, Lu Z, Herskovits E. 2011. Predictive models of autism spectrum disorder based on brain regional cortical thickness. *Neuroimage* **50**:589–599.
Kapur S, Phillips AG, Insel TR. 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry* **17**:1174–1179.
Khundrakpam BS, Lewis JD, Kostopoulos P, Carbonell F, Evans AC. 2017. Cortical Thickness Abnormalities in Autism Spectrum Disorders Through Late Childhood, Adolescence, and Adulthood: A Large-Scale MRI Study. *Cereb Cortex* **27**:1721–1731.
Lai M-C, Baron-Cohen S, Buxbaum JD. 2015a. Understanding autism in the light of sex/gender. *Mol Autism* **6**:24.
Lai M-C, Lombardo MV, Auyeung B, Chakrabarti B, Baron-Cohen S. 2015b. Sex/Gender Differences and Autism: Setting the Scene for Future Research. *J Am Acad Child Adolesc Psychiatry* **54**:11–24.
Lai M-C, Lombardo MV, Baron-Cohen S. 2014. Autism. *Lancet* **383**:896–910.
Lefebvre A, Delorme R, Delanoë C, Amsellem F, Beggiato A, Germanaud D, Bourgeron T, Toro R, Dumas G. 2018. Alpha Waves as a Neuromarker of Autism Spectrum Disorder: The Challenge of Reproducibility and Heterogeneity. *Front Neurosci* **12**:662.
Lombardo MV, Lai M-C, Baron-Cohen S. 2019. Big data approaches to decomposing heterogeneity across the autism spectrum. *Mol Psychiatry*. doi:10.1038/s41380-018-0321-0
Lord C, Bishop S, Anderson D. 2015. Developmental trajectories as autism phenotypes. *Am J Med Genet C Semin Med Genet* **169**:198–208.
Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF. 2019. Conceptualizing mental disorders as deviations from normative functioning. *Mol Psychiatry*. doi:10.1038/s41380-019-0441-1
Marquand AF, Rezek I, Buitelaar J, Beckmann CF. 2016. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol Psychiatry* **80**:552–561.
Mills KL, Goddings A-L, Herting MM, Meuwese R, Blakemore S-J, Crone EA, Dahl RE, Güroğlu B, Raznahan A, Sowell ER, Tamnes CK. 2016. Structural brain development between childhood and adulthood: Convergence across four longitudinal samples. *Neuroimage* **141**:273–281.
Moradi E, Khundrakpam B, Lewis JD, Evans AC, Tohka J. 2017. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. *Neuroimage* **144**:128–141.
Raznahan A, Lerch JP, Lee N, Greenstein D, Wallace GL, Stockman M, Clasen L, Shaw PW, Giedd JN. 2011a. Patterns of coordinated anatomical change in human cortical development: a longitudinal neuroimaging study of maturational coupling. *Neuron* **72**:873–884.
Raznahan A, Shaw P, Lalonde F, Stockman M, Wallace GL, Greenstein D, Clasen L, Gogtay N, Giedd JN. 2011b. How Does Your Cortex Grow? *Journal of Neuroscience* **31**:7174–7177.
Romero-garcia R, Atienza M, Clemmensen LH, Cantero JL. 2012. Effects of network resolution on topological properties of human neocortex. *Neuroimage* **59**:3522–3532.
Romero-Garcia R, Warrier V, Bullmore ET, Baron-Cohen S, Bethlehem RAI. 2018. Synaptic and

transcriptionally downregulated genes are associated with cortical thickness differences in autism. *Mol Psychiatry*. doi:10.1038/s41380-018-0023-7

Rosen AFG, Roalf DR, Ruparel K, Blake J, Seelaus K, Villa LP, Ciric R, Cook PA, Davatzikos C, Elliott MA, Garcia de La Garza A, Gennatas ED, Quarmley M, Schmitt JE, Shinohara RT, Tisdall MD, Craddock RC, Gur RE, Gur RC, Satterthwaite TD. 2018. Quantitative assessment of structural image quality. *Neuroimage* **169**:407–418.

Schumann CM, Bloss CS, Barnes CC, Wideman GM, Carper RA, Akshoomoff N, Pierce K, Hagler D, Schork N, Lord C, Courchesne E. 2010. Longitudinal Magnetic Resonance Imaging Study of Cortical Development through Early Childhood in Autism. *Journal of Neuroscience* **30**:4419–4427.

Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B. 2004. A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**:1060–1075.

Smith E, Thurm A, Greenstein D, Farmer C, Swedo S, Giedd J, Raznahan A. 2016. Cortical thickness change in autism during early childhood. *Hum Brain Mapp* **37**:2616–2629.

van der Maaten L. 2014. Accelerating t-SNE using Tree-Based Algorithms. *J Mach Learn Res* **15**:3221–3245.

van Rooij D, Anagnostou E, Arango C, Auzias G, Behrmann M, Busatto GF, Calderoni S, Daly E, Deruelle C, Di Martino A, Dinstein I, Duran FLS, Durston S, Ecker C, Fair D, Fedor J, Fitzgerald J, Freitag CM, Gallagher L, Gori I, Haar S, Hoekstra L, Jahanshad N, Jalbrzikowski M, Janssen J, Lerch J, Luna B, Martinho MM, McGrath J, Muratori F, Murphy CM, Murphy DGM, O'Hearn K, Oranje B, Parellada M, Retico A, Rosa P, Rubia K, Shook D, Taylor M, Thompson PM, Tosetti M, Wallace GL, Zhou F, Buitelaar JK. 2018. Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group. *Am J Psychiatry* **175**:359–369.

Visser GHA, Eilers PHC, Elferink-Stinkens PM, Merkus HMWM, Wit JM. 2009. New Dutch reference curves for birthweight by gestational age. *Early Hum Dev* **85**:737–744.

Vissers ME, Cohen MX, Geurts HM. 2012. Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links. *Neurosci Biobehav Rev* **36**:604–625.

Zabihi M, Oldehinkel M, Wolfers T, Frouin V, Goyard D, Loth E, Charman T, Tillmann J, Banaschewski T, Dumas G, Holt R, Baron-Cohen S, Durston S, Bölte S, Murphy D, Ecker C, Buitelaar JK, Beckmann CF, Marquand AF. 2019. Dissecting the Heterogeneous Cortical Anatomy of Autism Spectrum Disorder Using Normative Models. *Biol Psychiatry Cogn Neurosci Neuroimaging* **4**:567–578.

Zielinski BA, Prigge MBD, Nielsen JA, Froehlich AL, Abildskov TJ, Anderson JS, Fletcher PT, Zygmunt KM, Travers BG, Lange N, Alexander AL, Bigler ED, Lainhart JE. 2014. Longitudinal changes in cortical thickness in autism and typical development. *Brain* **137**:1799–1812.