Supplementary Materials

Table of contents

Supplementary methods

1. Generation and comparison of FSQC scores

1.1 Comparison of FSQC images with 10 versus 20 slices

For comparison and to confirm that 10 slices is sufficient to get a good representation of the quality of the whole image, we generated FSQC images for two participants of 20 slices instead of 10 (supplementary methods section 1.1). These images were created in exactly the same way as the original images with 10 slices, but additional slices were added in between existing slices, resulting in intervals of 10 instead of 20 slices, spanning the whole brain.

1.2 Comparison of rater maps

To ensure results did not differ by rater, the same linear mixed effects models were conducted on cortical thickness using FSQC scores by each main rater (SB, RB), as well as the average score between both raters, and maps displaying the results across the cortex were visually compared across the three analyses.

1.3 Comparison with FreeSurfer 7.1

A subset of 50 participants was run with FreeSurfer 7.1, for comparison with newer, more up to date FreeSurfer tools and methods. FSQC images were generated and rated by the main rater (SB). Spearman correlations were run to examine the consistency between FSQC and Euler number in FreeSurfer version 6.0.1 and 7.1.

1.4 Correlation of all metrics and IQMs

We assessed correlations between all pairs of quality metrics, including all image quality metrics (IQMs) from MRIQC and the QAP. Note that the MRIQC metrics released with ABIDE are only available for ABIDE II, and the QAP metrics are only available for ABIDE I. The Quality Assessment Protocol (Zarrar Shehzad, Steven Giavasis, Qingyang Li, Yassine Benhajali, Chaogan Yan, Zhen Yang, Michael Milham, Pierre Bellec and Cameron Craddock, 2015), released with ABIDE I, includes contrast to noise ratio (CNR), entropy focus criterion (EFC), foreground to background energy ratio (FBER), smoothness of voxels (FWHM), percent of artefact in voxels (Qi1) and signal to noise ratio (SNR). MRIQC was based on the QAP and includes similar measures. IQMs released with ABIDE II include contrast to noise ratio (CNR), signal to noise ratio (SNR), first quality index (a measure of artefact levels; Qi1), entropy focused criterion (EFC), smoothness of voxels, foreground-background energy ratio, and cortical contrast. For details and specifications of each measure please see (Esteban et al., 2017).

2. Additional main analyses

2.1 Desikan-Killiany atlas results

For comparison with previous work and our replication analysis, linear mixed effects models were run for each Desikan-Killiany parcellation, using the same models as the main analyses, for CT, SA and CV, and partial correlations extracted and plotted. Results were corrected for multiple comparisons using the false discovery rate (FDR) across parcellations in all analyses.

2.2 Meta-analysis results

The FSQC and Euler analyses for cortical thickness were also conducted using a meta-analytic approach, as an alternative method to mixed effects models for controlling for site differences. This method also provides information about intersite heterogeneity. For these analyses, linear models were run separately per site, including QC metric, age, age^2, and sex, and then pooled across sites in a random effects meta-analysis based on partial r correlation coefficients, using the metafor package in R.

2.3 Replication in multiple datasets

2.3.1 Euler number replication in Lifespan sample

To validate and attempt to replicate our results in a larger, more representative dataset covering the whole human lifespan, we ran the same main analyses for Euler using one of the largest MRI datasets currently available (see Bethlehem et al, 2022). To avoid confounds of various diagnoses, we used controls only from this dataset, and excluded ABIDE so as not to have overlap with our original analysis, resulting in a sample of 74,647 individuals. We then ran the same linear mixed effects models, including Euler, age, age^2 and sex, with site as a random variable, for all three cortical phenotypes separately. This was done for Euler number, and using DK parcellations, as this was the data available for this dataset.

2.3.2 FSQC and Euler replication in multiple neurodevelopmental datasets

So as to have a comparison of FSQC as well as Euler number, and to be able to show the reliability of this measure across multiple developmental datasets, we also generated FSQC ratings and performed the same analyses in three other datasets including individuals with neurodevelopmental conditions: the Child Mind Institute's (CMI) Healthy Brain Network; the ADHD200 dataset; and the Provide of Ontario Neurodevelopmental (POND) Network. Again, to avoid confounds of different diagnoses, we performed these analyses in controls only. We ran the same linear mixed effects models, separately for FSQC and Euler number, including, age, age^2 and sex as covariates, with site as a random variable, for cortical thickness. These analyses were conducted using Glasser parcellations, for comparison with the main results of this manuscript.

2.3.3 Dataset demographics

The Lifespan Sample

The Lifespan sample was compiled by (Bethlehem et al., 2022) and comprises neuroimaging data over 100 studies internationally, spanning 0-100 years of age. The total dataset comprises 101,457 individuals, with 74,647 controls who were analysed for the purposes of this manuscript. For more details see (Bethlehem et al., 2022)

The Province of Ontario Neurodevelopmental Disorders (POND) Network

POND is an integrated discovery program based in Ontario, Canada. POND includes neuroimaging, clinical, behavioural and genetic data across multiple time points of children with neurodevelopmental disorders, including ADHD, autism, intellectual disability, obsessive compulsive disorder and Tourette syndrome, amongst others, from five different sites in Ontario (Holland Bloorview Kids Rehabilitation Hospital, Toronto; The Hospital for Sick Children, Toronto; McMaster Children's Hospital, Hamilton; Queen's University, and Lawson Health Research Institute, London). The number of control participants analysed here was 105. For more details please see: https://pond-network.ca/.

The Healthy Brain Network (HBN) at the Child Mind Institute (CMI)

The HBN(Alexander et al., 2017) is an initiative based at the Child Mind Institute, in New York City, USA, aiming to create a biobank of data from children and adolescents, with the goal of advancing our understanding of childhood and adolescent psychiatric and neurodevelopmental disorders. Data collected includes clinical, behavioural and cognitive measures, as well as genetic information and neuroimaging, collected at three sites in and around New York City. The number of control participants analysed here was 145.

The ADHD200 Consortium

The publicly available ADHD200 Consortium consists of neuroimaging and phenotypic data from individuals with ADHD and controls across 8 sites internationally. For details see: http://fcon_1000.projects.nitrc.org/indi/adhd200/. The number of control participants analysed here was 493.

2.4 Variance partitioning

We also conducted a variance partitioning analysis to quantify the proportion of variance accounted for by scan quality (FSQC and Euler), relative to other covariates on global brain measures (cortical GMV, WMV, subcortical GMV, ventricular volume, total GMV, estimated TIV, and mean cortical thickness), and for cortical thickness. This analysis was conducted using the variancePartition package in R (http://bioconductor.org/packages/variancePartition (Hoffman & Schadt, 2016)), often used for, and adapted from, gene expression analysis. Linear mixed models are used to quantify variance in (here) cortical thickness parcellations contributed by FSQC, Euler, and other main variables: diagnosis, sex, age, and site. Percent of variance explained is then plotted for each neuroanatomical phenotype of interest (cortical and subcortical total grey matter volume, white matter volume, and ventricle volume).

3. Thresholding analyses

3.1 Euler thresholding per site

Due to the significant variation (in scanner, scanning parameters, demographics, and quality) between sites in the ABIDE dataset, we repeated the Euler threshold analyses applying the cut-off point based on MADs to each site individually, rather than the whole sample, for cortical thickness. MAD was calculated individually for each site, and used to exclude participants above that value for each site, at MADs of 1, 2 and 3. The same linear models as above were then run at each cut off point, assessing the relationship between CT and Euler.

3.2 FSQC median split analysis

We also compared CT between participants with the best and worst scan quality, based on FSQC scores: participants were split into two groups ("high" and "low" scores) based on a median split of 1.1. A mixed linear effects model was then run, including group (high/low), age, age^2, and sex as fixed effects, and site as a random effect.

3.3 Percent exclusion analysis

Next, we wanted to assess the impact of excluding participants with the worst scan quality, based on the top percentage rather than predefined cut off points. We conducted this analysis based on Euler number, for CT, at thresholds in increments of 5%, from 5-50%, both of the whole dataset, and per individual site. As in previous analyses, regional relationships for CT were assessed at each threshold by running the same linear mixed effects models. The relationship between the number of significant regions remaining and the percentage of participants excluded was plotted, as was the relationship between the strength of the partial correlation of the strongest regions and the percent of excluded participants, to examine how excluding different proportions of the dataset impacts or attenuates the impact of scan quality on morphometric estimates.

4. Diagnosis

4.1 Replication in CMI and POND datasets

We attempted to replicate the cortical thickness analyses in the CMI and POND datasets, comparing autistic to non-autistic individuals with and without QC of different methods, in the same manner as for ABIDE: First, we examined group differences in CT without accounting for quality (using linear mixed effects models with diagnosis, age, age^2 and sex in the model and site as a random factor). Next, the same models were run with the addition of FSQC or Euler number as a covariate to assess the impact of controlling for quality, as well as thresholding by both FSQC (at 2.5) and Euler (at 2 MAD). Both CMI and POND include individuals with various neurodevelopmental conditions; however, we included only autistic and neurotypical individuals in these analyses. CMI had a total of 105 autistic participants and 145 controls, and POND had 396 autistic participants and 105 controls.

4.2 Cut off points

To further explore the impact of thresholding by scan quality, we reran the same analysis exploring the impact of diagnosis on CT at cut-off points of 3, 2.5, 2, 1.5 for FSQC, and 1, 2 and 3 MADs for Euler number. At each threshold, the model included diagnosis, age, age^2, and sex as fixed variables and site as a random variable.

4.3 Thresholding by FSQC but controlling for Euler

Finally, we examined the effect of combining FSQC and Euler as QC methods, by examining the effect of diagnosis while controlling for Euler at various FSQC thresholds (3, 2.5, 2, and 1.5 as above). Here, the model included diagnosis, age, age^2, Euler number and sex as fixed variables and site as a random variable.

4.4 Diagnosis Interaction

Finally, the interaction between diagnosis and FSQC or Euler was assessed on each cortical phenotype. For significant regions, the effect of FSQC and Euler on cortical estimates were then examined separately in the autistic and control groups.

4.5 SA and CV diagnosis analysis

The main diagnosis analyses were also run for SA and CV, including examining the impact of diagnosis alone, diagnosis controlling for FSQC and Euler, and the impact of diagnosis while thresholding by each FSQC (at 2.5) and Euler (at 2 MADs).

Supplementary results

1. Generation and comparison of FSQC scores

1.1 Comparison of FSQC images with 10 versus 20 slices

To demonstrate both ends of the spectrum, we chose one participant with an overall score of 1 and one of 3.3 by rater SB. In the FSQC images of 20 slices, we do not see any artefacts or reconstruction errors that were not already noted in the 10 slices, demonstrating that the scores are unlikely to be altered by adding additional slices, and that 10 slices seems to be sufficient to represent the quality of the whole brain.

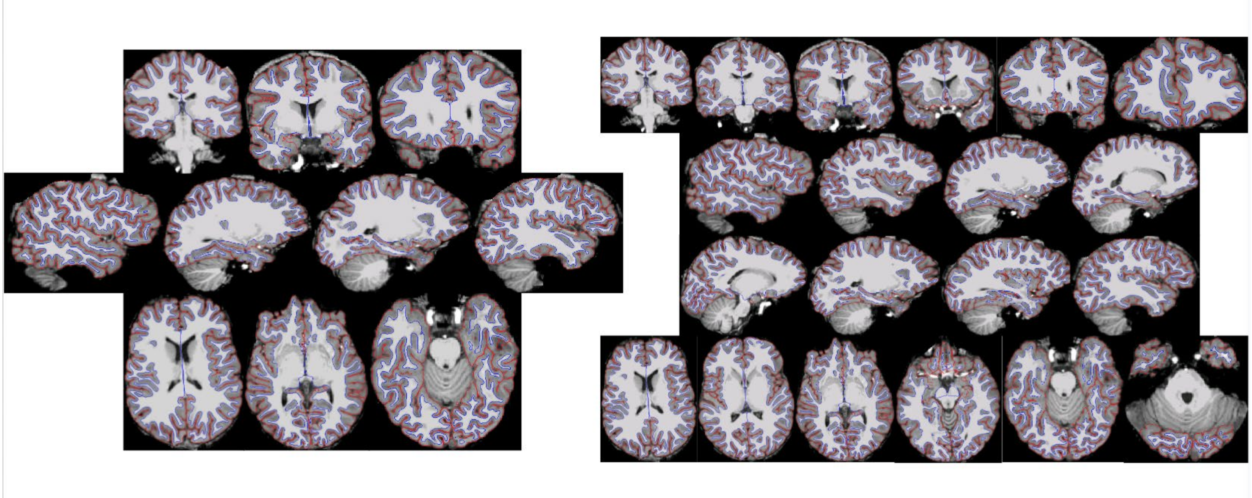

Figure S1.1.1. A participant with a score of 1 (no visible artefact or reconstruction error). We demonstrate that no additional errors or artefacts are apparent when viewing additional slices.

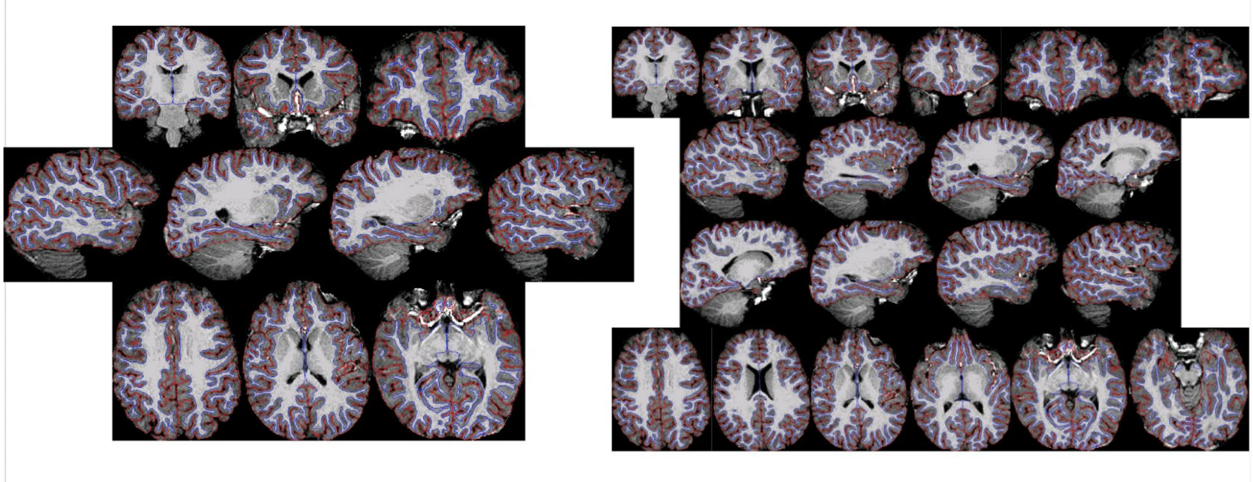

Figure S1.1.2 A participant with a score of 3.3 (multiple slices with visible motion and/or missing areas of cortex). We demonstrate that no additional errors or artefacts are apparent when viewing additional slices.

1.2 SB, RB and average scores

Regional results depicting the relationship between FSQC and cortical thickness were almost identical for both raters and average ratings across the cortex.

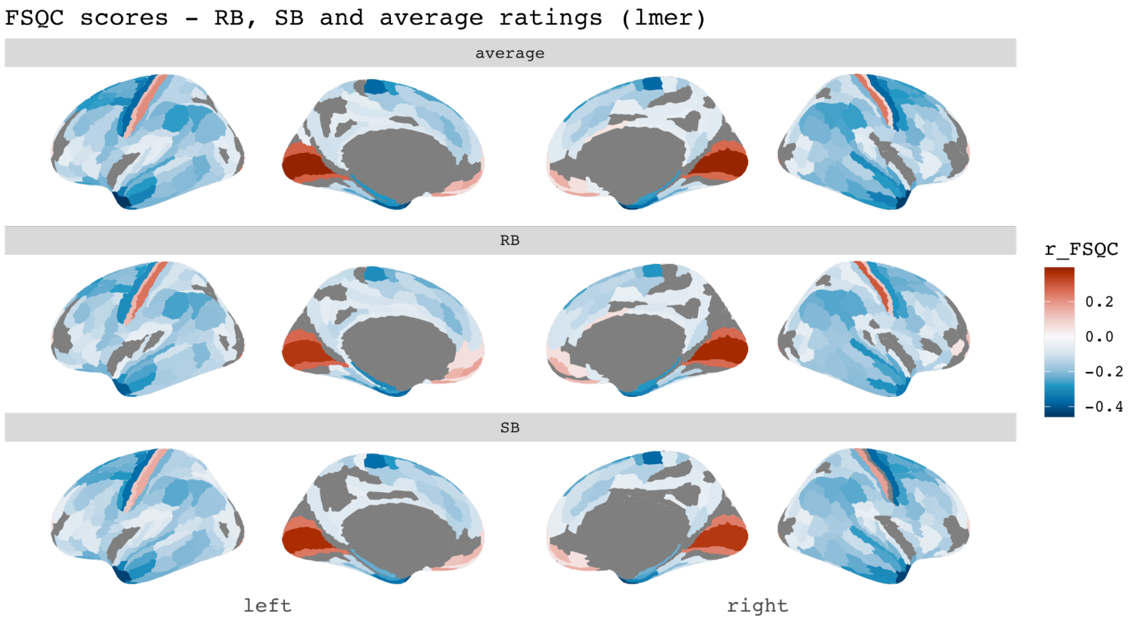

Figure S1.2 Associations between FSQC and cortical thickness for average scores between SB and RB (top), RB only (middle) and SB only (bottom row). Spatial maps were consistent across all three sets of scores.

1.3 Comparison with FreeSurfer 7.1

FSQC ratings of FreeSurfer 7.1 outputs were highly and significantly correlated with FSQC ratings of 6.0.1 outputs (rho=0.87, p<0.0001), as was Euler number (rho=0.88, p<0.0001). We note that Euler numbers were overall much lower in FS version 7.1, at a factor of about one half, but this seemed to scale consistently. Despite these differences, FSQC scores were largely unaffected and remained consistent.

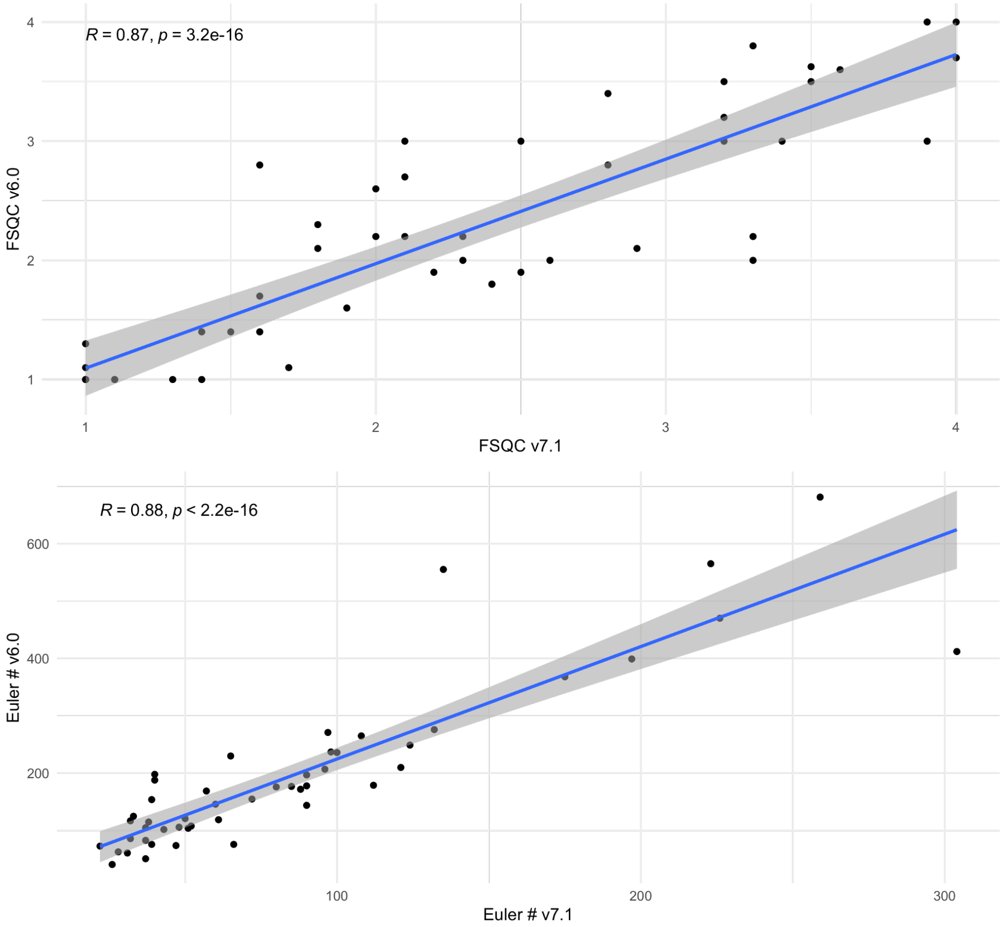

Figure S1.3. Spearman correlations of FSQC and Euler between FreeSurfer versions 6.0.1 and 7.1.

1.4 Correlations with MRIQC and QAP metrics

Correlation matrix showing spearman correlations for all possible pairs of QC metrics, including all image quality metrics (IQMs) from MRIQC and the QAP. Note that MRIQC is only available for ABIDE II, and the QAP is only available for ABIDE I, so correlations between MRIQC and QAP IQMs was not possible. Measures from CNR - cortical contrast are from MRIQC, and anat_cnr to anat_snr are from QAP. Correlations between our main metrics and MRIQC and QAP IQMs were generally low, with the highest correlations seen for Euler number (maximum correlation = -0.38). The highest correlation between any IQM and FSQC was 0.24.

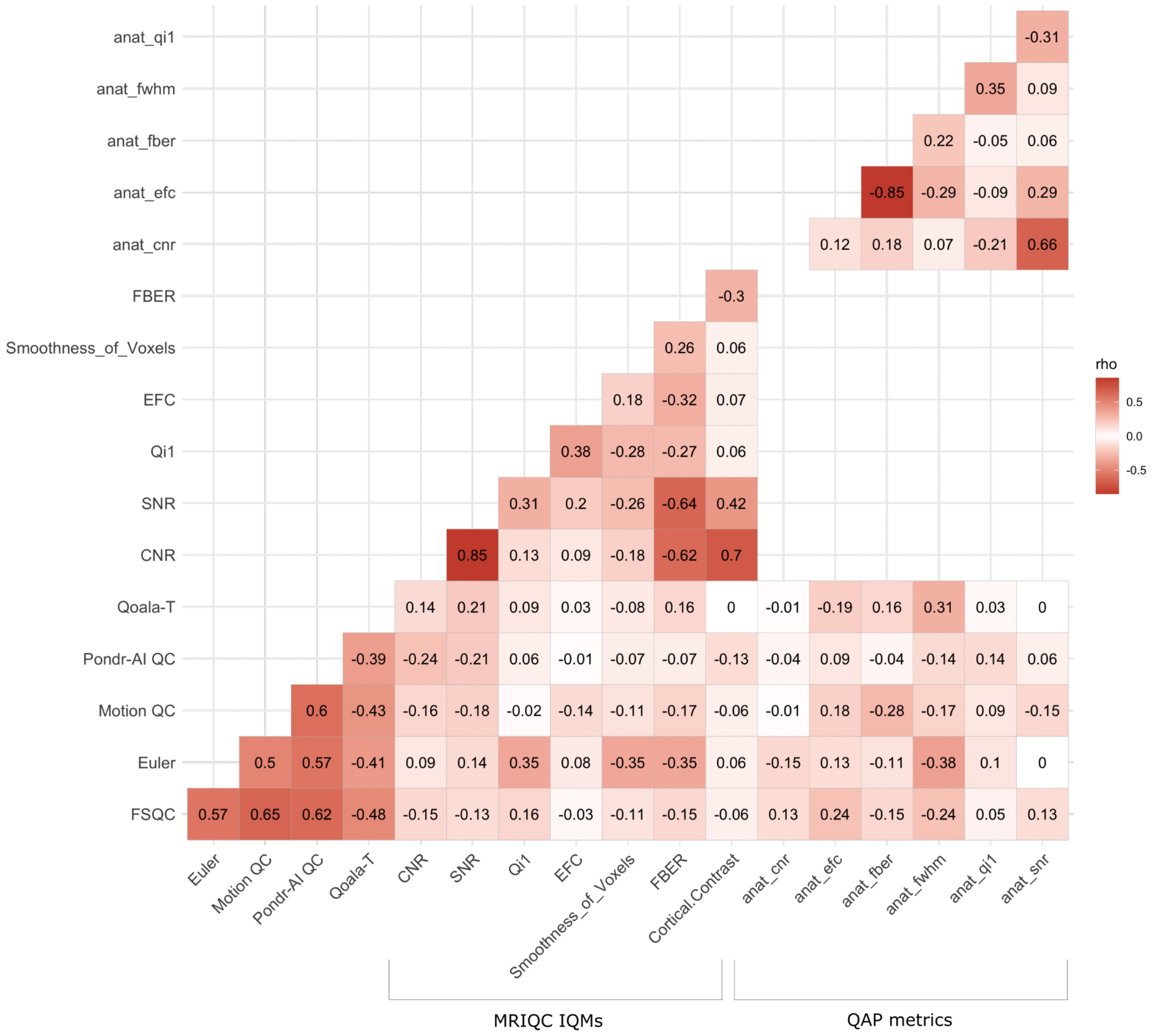

Figure S1.4. Correlation between all possible pairs of QC metrics. Measures from MRIQC (from CNR to cortical contrast) are for ABIDE II only; measures from QAP (from anat_cnr to anat_snr) are for ABIDE I only; thus these correlations only include those participants from the respective ABIDE release. The blank space in the correlation matrix reflects the fact that correlation between MRIQC and QAP metrics are not possible because they only exist for either ABIDE I or II. The colour map is non-divergent to reflect the fact that metrics differ in whether higher or lower quality is associated with higher scores.

2. Sensitivity analyses (main results)

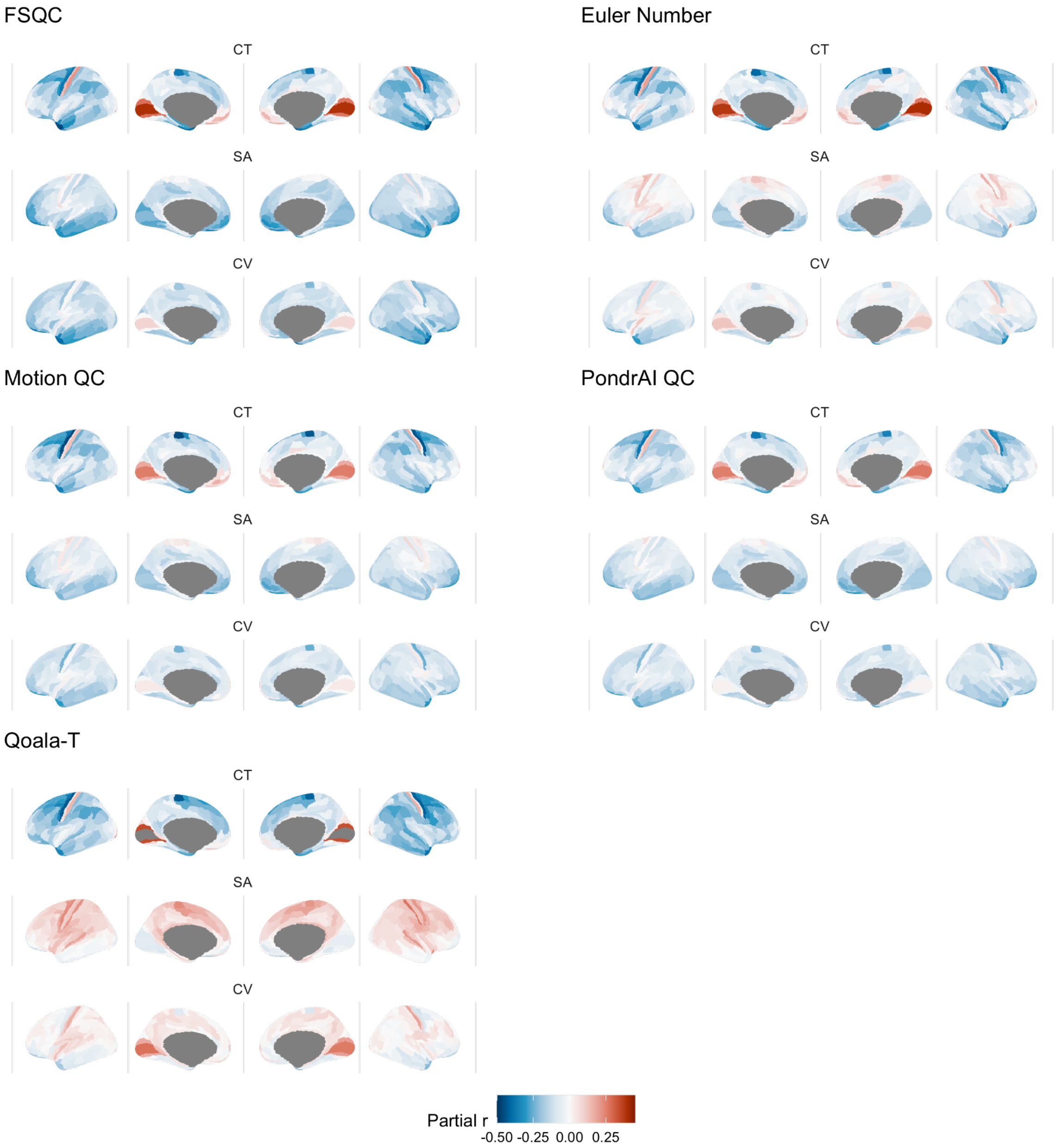

Figure S2. Results of regional analyses (main figure 2): effect of all QC metrics on each cortical phenotype, here unthresholded for significance.

FSQC correlations with global brain measures

	Spearman rho	p-value
cGMV	-0.066	0.004
WMV	-0.21	< 0.0001
sGMV	-0.065	0.005
Ventricles	-0.055	0.018
TBV	-0.157	< 0.0001
meanCT	-0.044	0.059

Supplementary table S1

2.1 DK results

Results using the Desikan-Killiany parcellations were consistent with the main analyses using HCP parcellations, and again consistent across QC metrics with the exception of Euler for SA and CV. Strongest effects were seen for cortical thickness across all metrics. Associations were largely negative, with positive correlations observed in the occipital and ventromedial prefrontal cortices for CT, and motor cortex for SA.

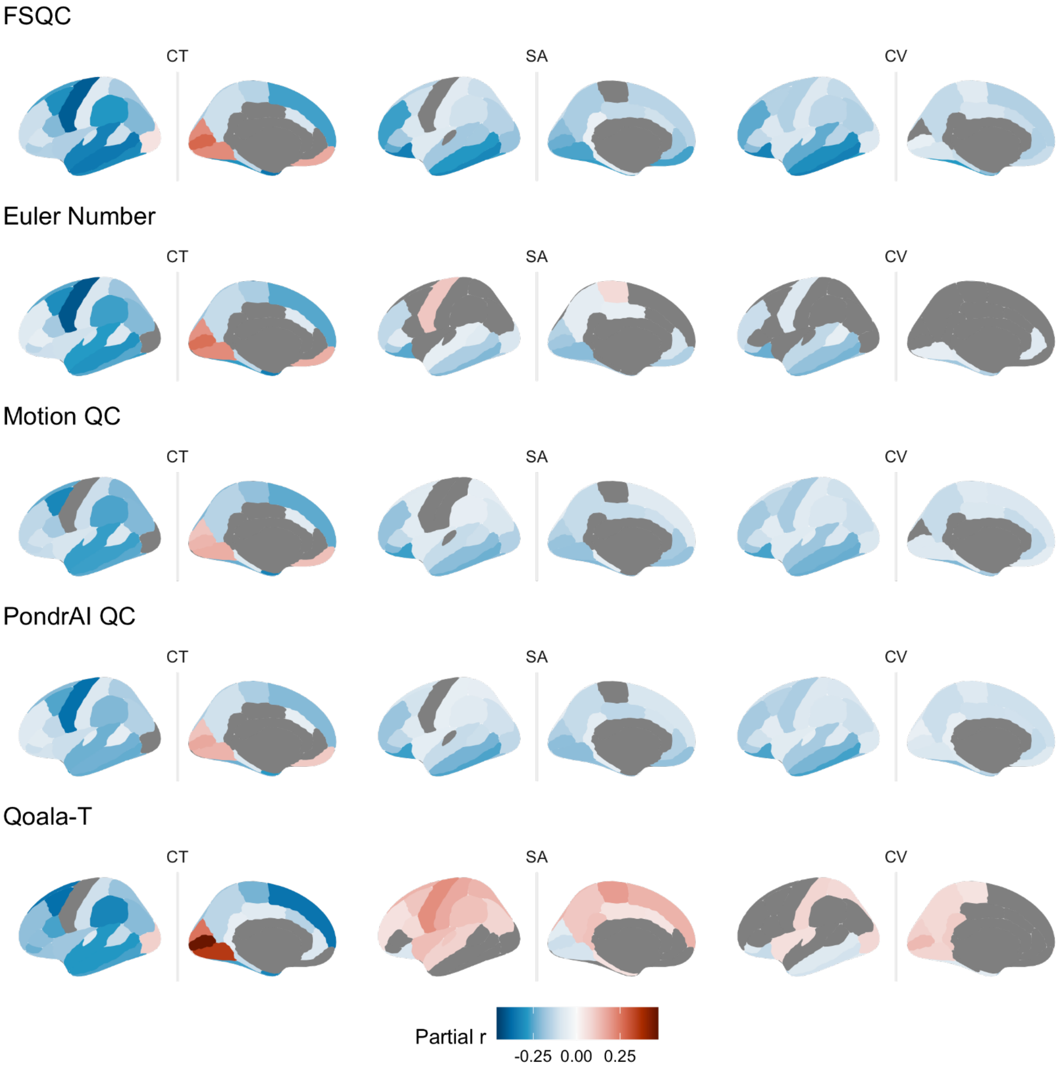

Figure S2.1. Associations between QC metrics and regional cortical morphometry using the Desikan-Killiany atlas parcellations. Results are consistent with those derived using the HCP-Glasser parcellation.

2.2 Meta-analysis results

Results from the meta-analytic technique for CT also yielded consistent, though slightly weaker, results to those using linear mixed effects models. Forest plots demonstrate largely consistent results across sites, with some variability.

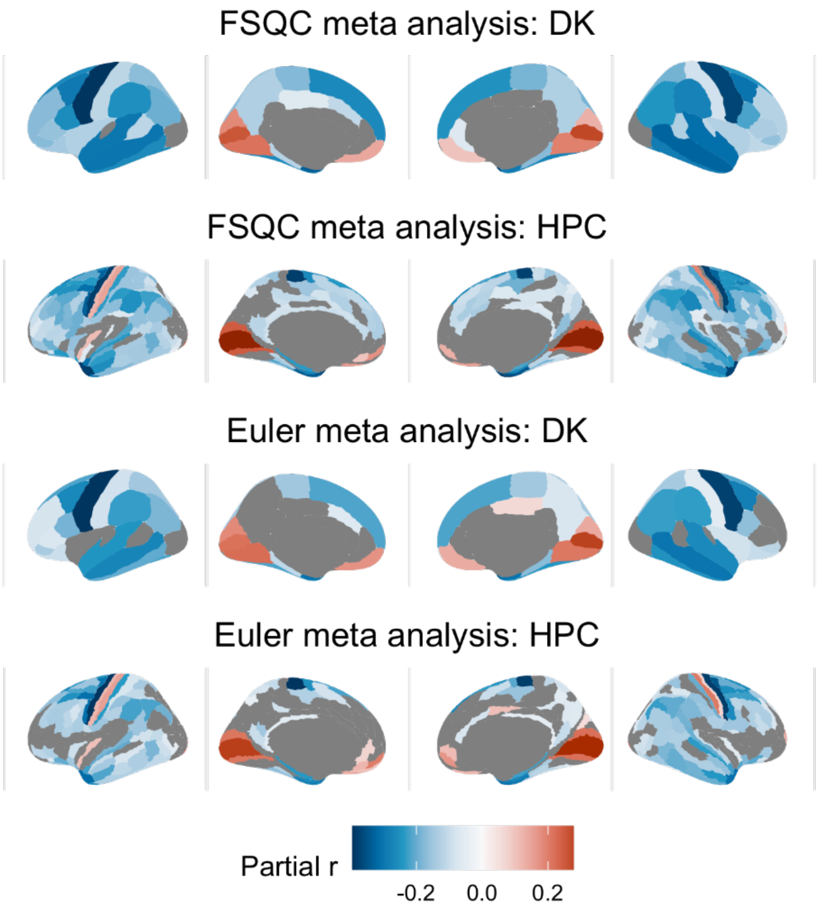

Fig. S2.2.1 Associations between Euler/FSQC and CT using a meta-analytic technique, with DK and HPC parcellations. Results are largely consistent with results from linear mixed effects models.

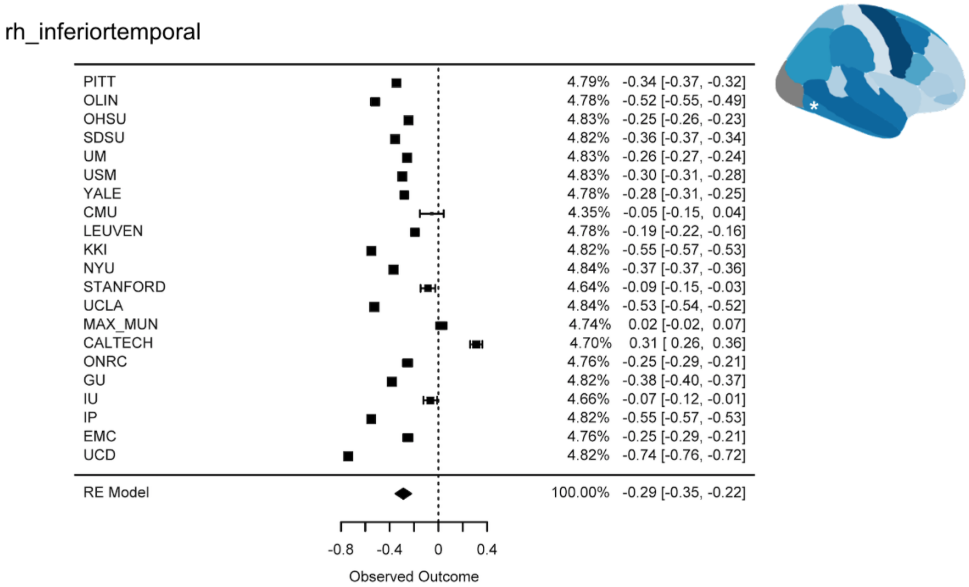

Fig S2.2.2. FSQC: Forest plot for DK inferior temporal gyrus parcellation. Effects are largely consistent across sites: all but two sites show a negative effect size, with some variability.

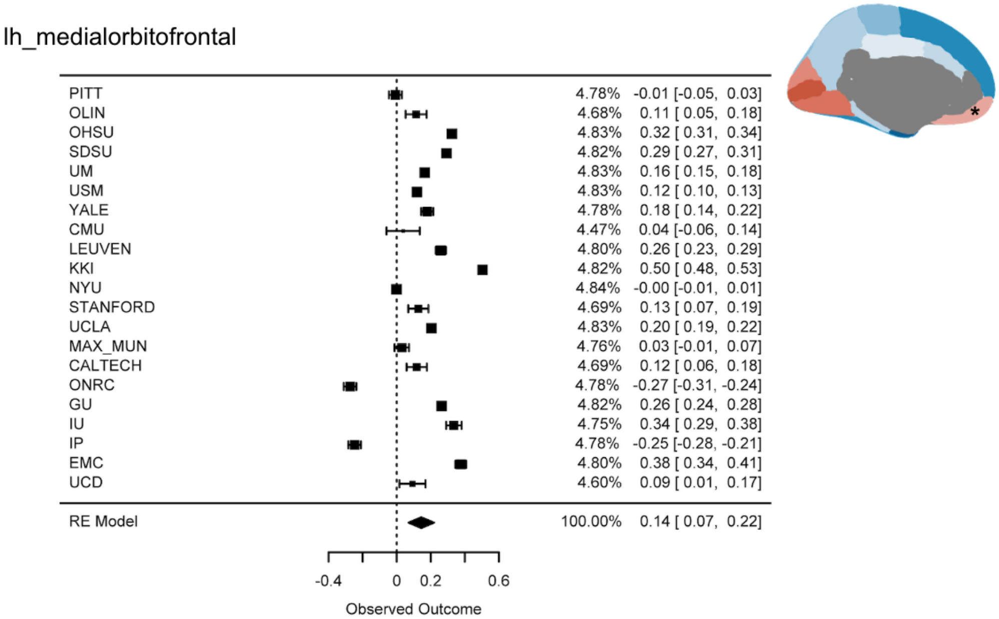

Fig S2.2.3. FSQC: Forest plot for DK ventromedial prefrontal cortex parcellation. Most sites show a positive effect size, with a few exceptions and slightly more variability.

2.3 Replication in multiple samples

2.3.1 Replication in Lifespan sample

Conducting the same analyses for Euler on all three cortical phenotypes in a much larger, more representative sample yielded largely similar results with some key differences. Results for CT were almost identical, with negative associations across the cortex with the exception of the occipital and ventromedial prefrontal cortices. Results for SA differed the most, with positive associations observed across much of the cortex in this larger sample, but only in the motor cortex in the ABIDE sample. CV results were similar in magnitude and direction, but more regions reached significance in the larger sample than in ABIDE.

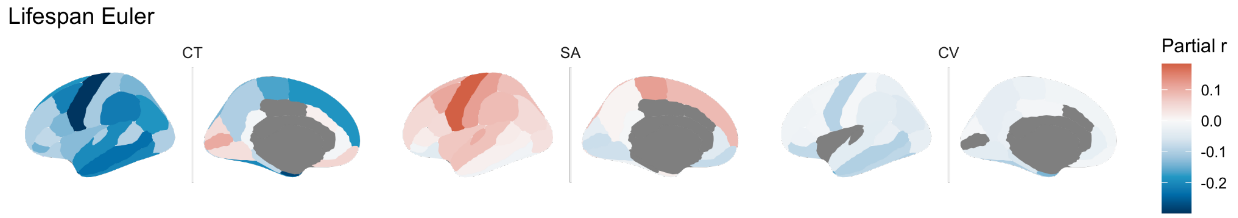

Fig S2.3.1. Relationship between each cortical phenotype and Euler number across the cortex in a large lifespan dataset. Results are largely consistent with ABIDE results, with the exception of more regions showing a positive relationship for SA.

2.3.2 Replication in multiple neurodevelopmental datasets

Analyses comparing the relationship between quality (FSQC/Euler number) and cortical thickness in controls across multiple datasets yielded high similarity in spatial patterning and directionality of effects, with some key differences in strength and significance. Most notably, stronger effects were seen in CMI relative to the other datasets. In ADHD200 some additional regions of positive associations were observed compared to the other datasets in frontal and medial parietal regions, and in POND in superior temporal regions, but relationships were weak. Results are shown below unthresholded (all regions; fig S2.3.2), and only in those passing FDR correction (fig S2.3.3).

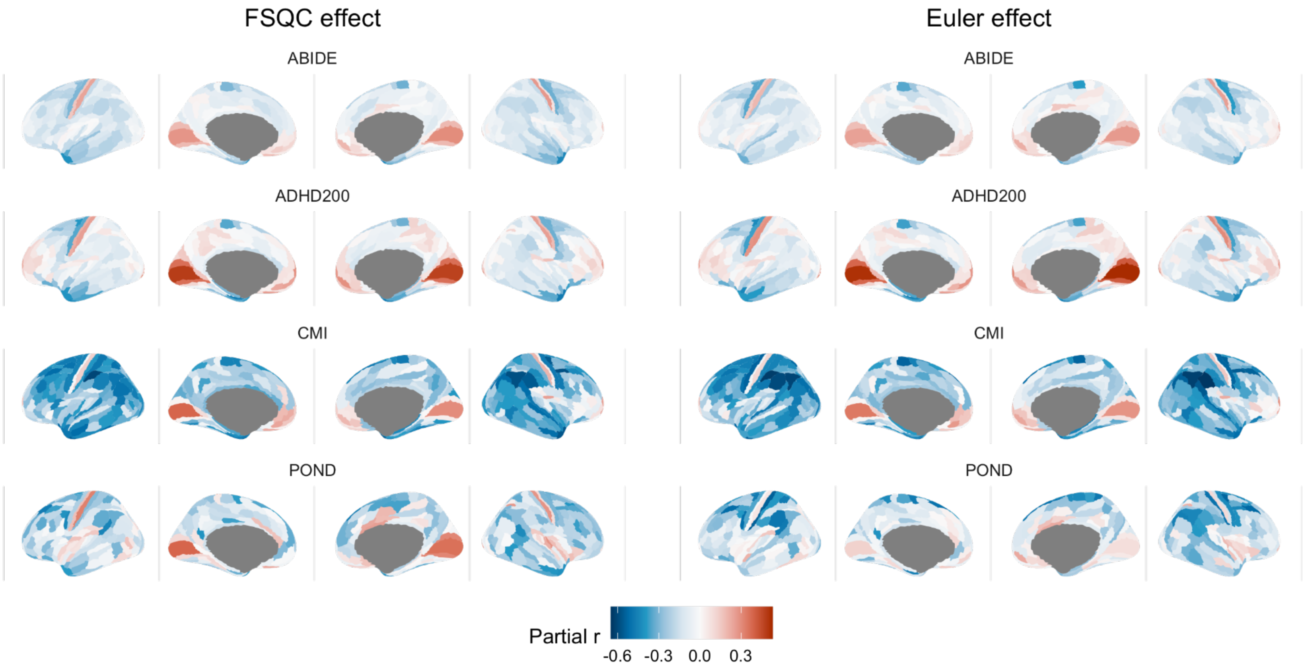

Fig S2.3.2. Relationship between each FSQC (left) and Euler (right) and cortical thickness in multiple datasets (ABIDE, ADHD200, CMI, POND) at all regions. Results are largely consistent in direction of effect, but differ in magnitude.

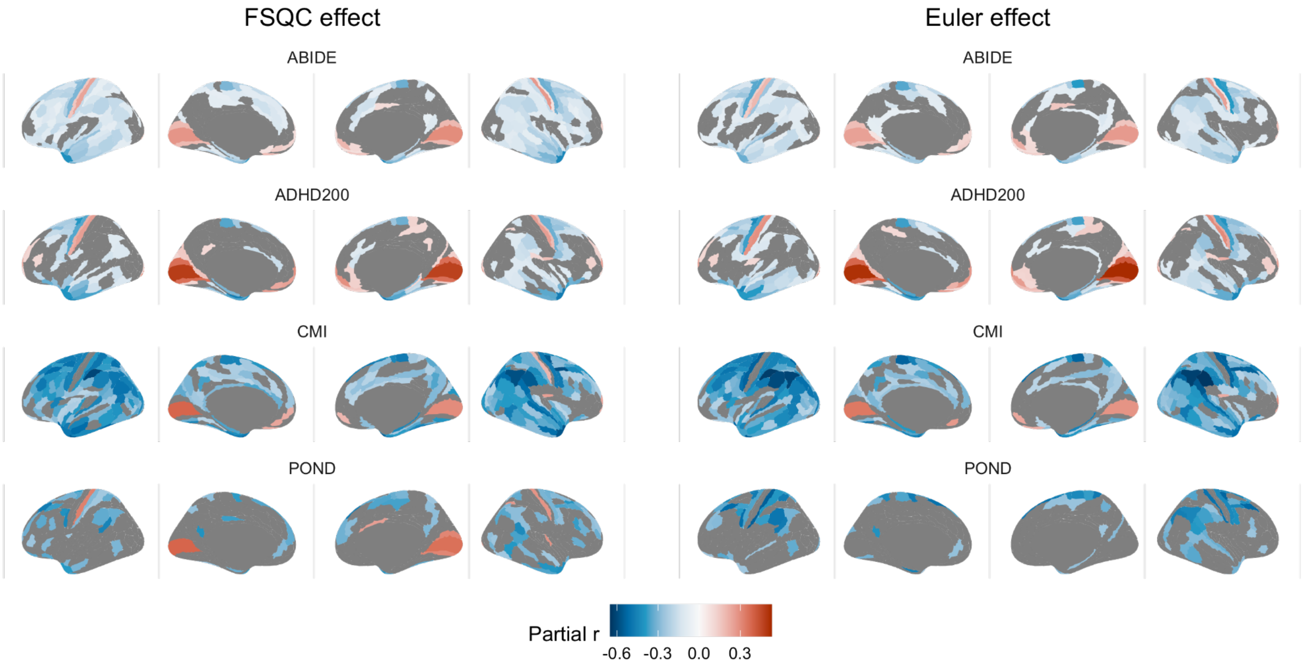

Fig S2.3.3. Relationship between each FSQC (left) and Euler (right) and cortical thickness in multiple datasets (ABIDE, ADHD200, CMI, POND), in regions passing FDR correction only. Results are largely consistent in direction of effect, but differ in strength and number of regions reaching significance.

2.4 Variance partitioning

The variance partitioning analysis on global brain measures indicated that, on average across all global brain measures, FSQC explained 3.7%, Euler number about 1%, and diagnosis 0.2% of the variance (Figure S2.4.1). For most measures, FSQC and Euler explained only a very small amount of the variance (<2%), but more substantial proportions of FSQC for cGMV (8.7%), tGMV (7.8%), and WMV (5.3%) (Figure S2.4.2). For all measures except ventricular volume, both quality metrics contributed to a substantially larger portion of the variance than diagnosis.

Across cortical thickness HPC parcellations, FSQC explained on average 1.4% of the variance, Euler 0.8% and diagnosis less than 0.1%. Site (16%) and age (11%) explained the most variance (Figure S2.4.3). A sample of 50 parcellations and their relative contributions is shown in Figure S2.4.4; FSQC and Euler appeared to explain significantly more variance than diagnosis for almost every parcellation. Brain maps of the percent of variance explained by FSQC and Euler are shown in Figure S2.4.5.

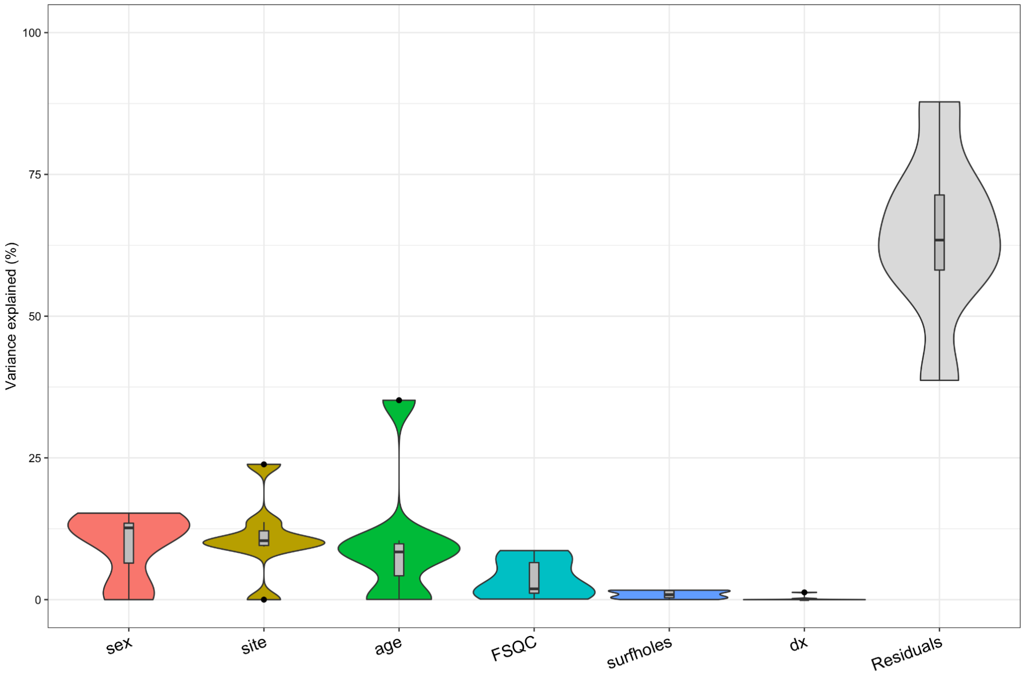

Fig S2.4.1. Average variance explained by each variable, across all global brain measures.

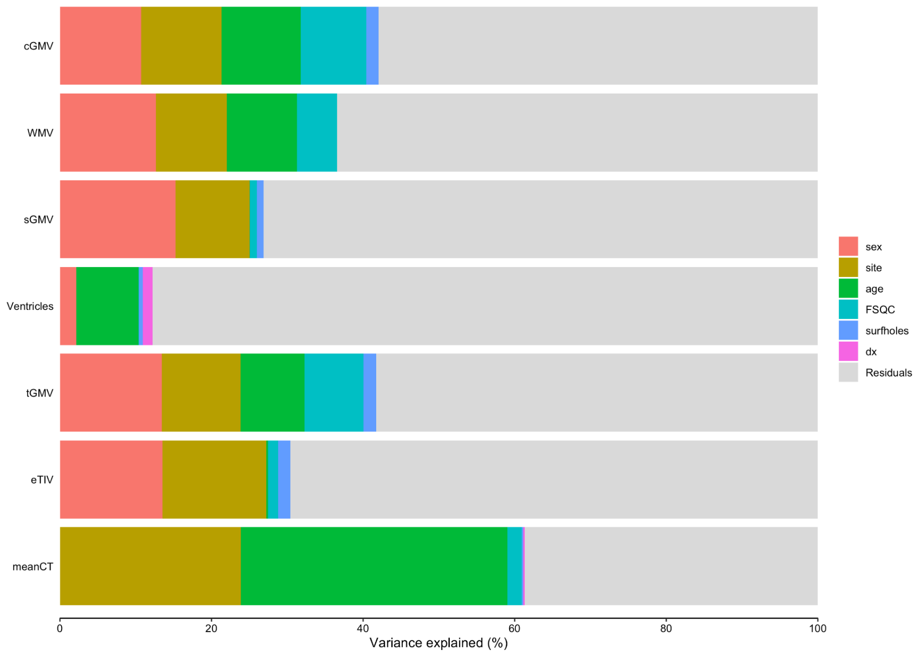

Fig S2.4.2. Percent of variance explained by each variable for each global brain measure.

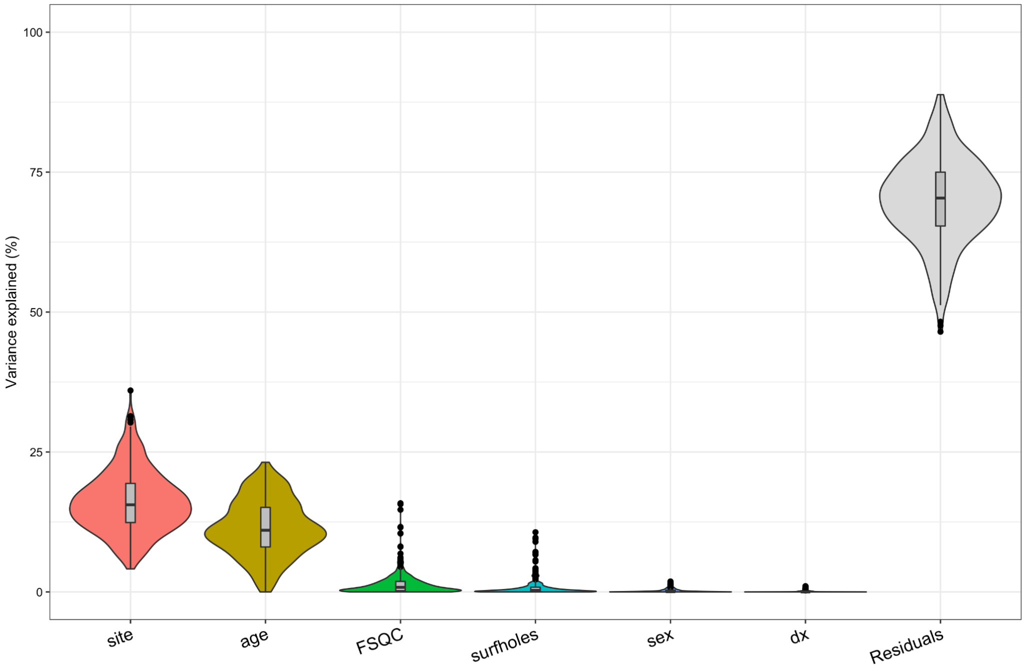

Fig S2.4.3 average variance explained by each variable, across all cortical thickness parcellations.

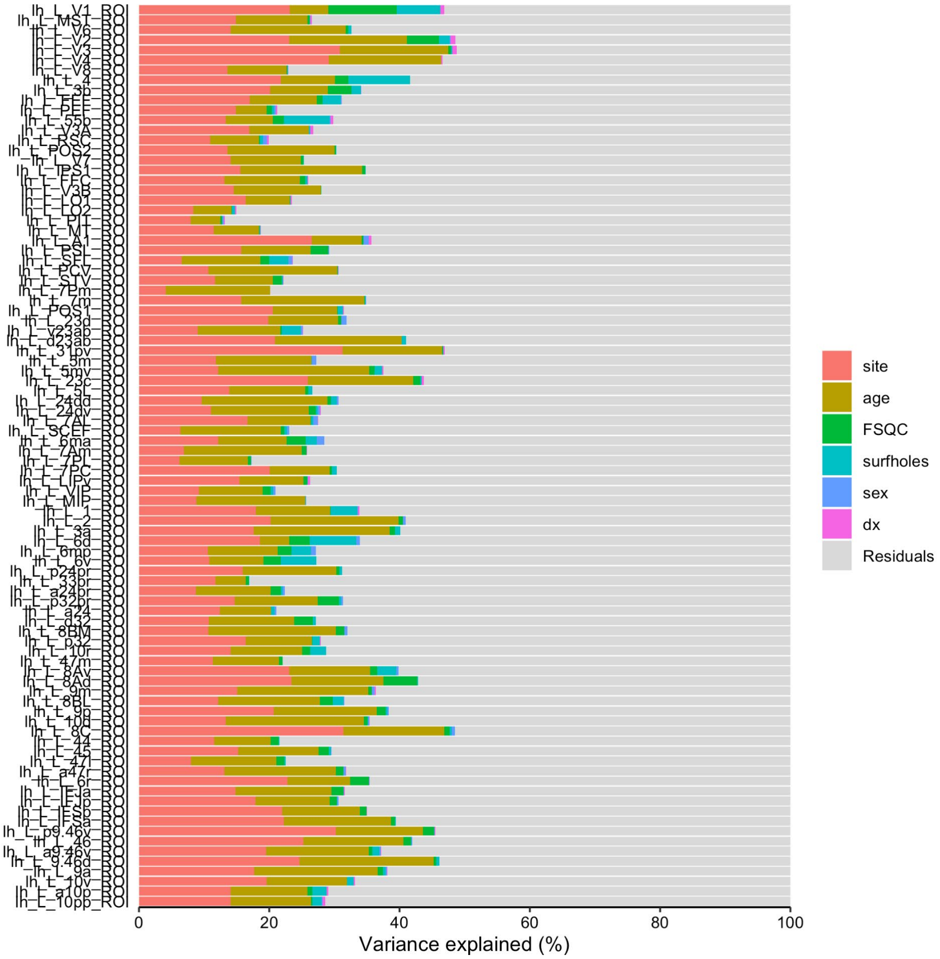

Fig S2.4.4 Percent of variance explained by each variable for the first 50 listed HPC parcellations.

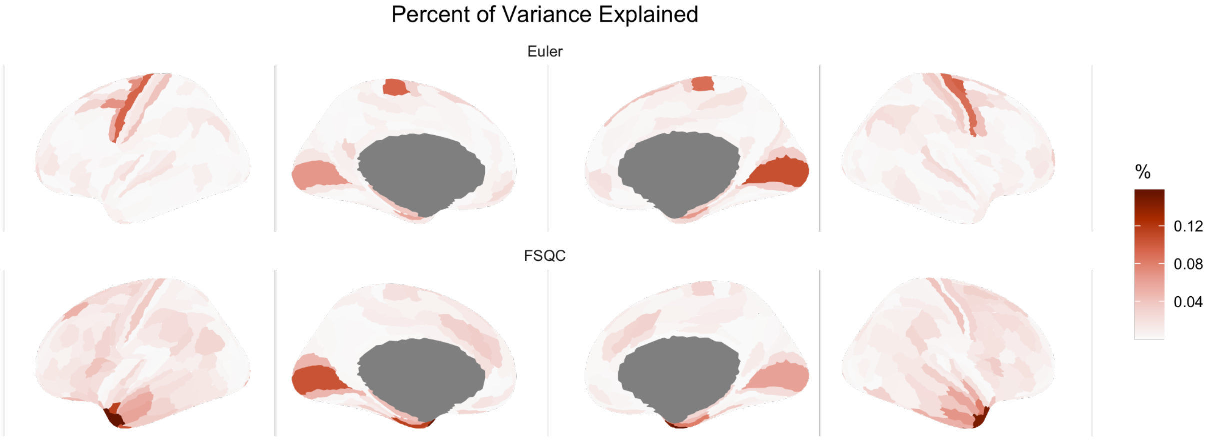

Fig S2.4.5. Percent of variance explained by Euler (top) and FSQC (bottom) for each HPC cortical thickness parcellation across the cortex.

3. Thresholding analyses

3.1 Euler threshold per site (CT)

When calculating and applying the MAD-based Euler threshold to each site individually rather than to the whole sample, results were comparable but with a few notable differences. Generally speaking, less regions remained significant at each cut-off point when applying the MAD-based threshold to each site individually compared to the whole sample; however, differences were minimal.

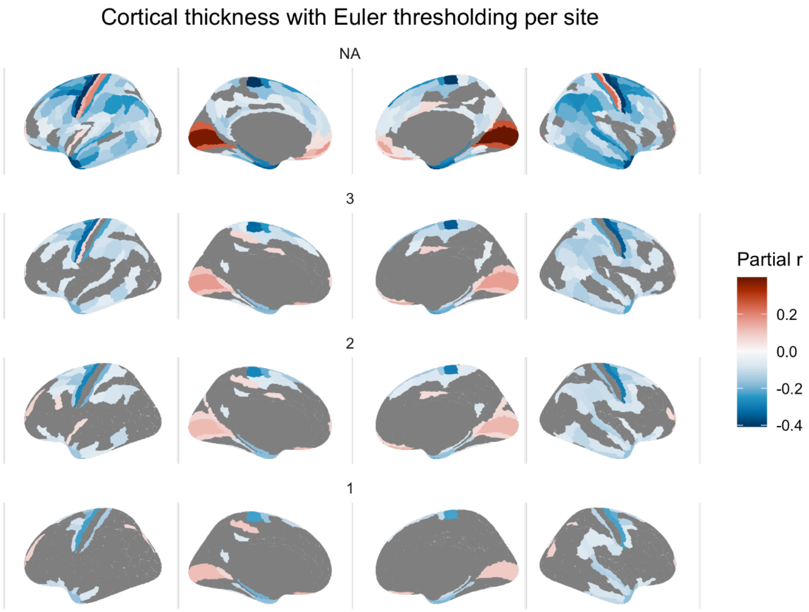

Fig S3.1. Relationship between CT and Euler at various Euler thresholds based on MADs applied per individual site (significant regions only).

3.2 FSQC median split analysis (CT)

Our median split analysis essentially compared all those with perfect or near perfect scans and surface reconstructions, as rated by our FSQC, to those with minor or major errors (split below and above 1.1, with 1 being a rating of “good” on all 10 images). Here, we observed significant and widespread cortical thickness differences between groups, in similar areas to those in which we observed significant correlations between scan quality and CT. The group differences we observed were largely of greater CT in the “high” quality group relative to the “lower” quality group. This is consistent with our previous results suggesting apparent cortical thinning related to poor scan quality. Again, as in these previous analyses, the exceptions were in the medial occipital cortex, inferior medial prefrontal cortex, and (right) postcentral gyrus, where we observed significantly lower CT values in the “high quality” group.

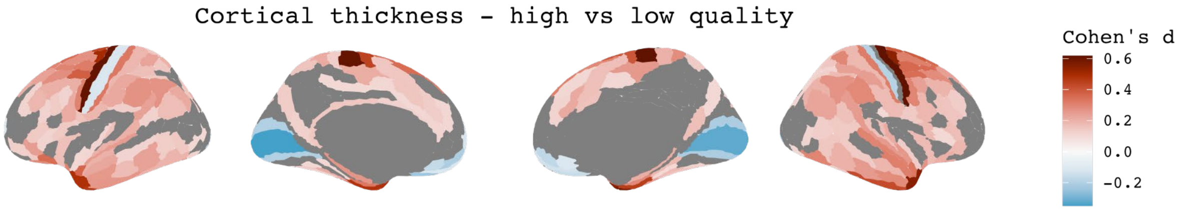

Fig S3.2 Median split analysis comparing scans with high (<1.1) vs low (>1.1) quality scans and outputs based on FSQC ratings.

3.3 Percent exclusion analysis

In both analyses (exclusion based on whole sample and per site), results were further attenuated at every 5% threshold interval, such that after excluding 50% of the data, only a few significant associations remained. These included negative correlations in the precentral gyrus, inferior temporal cortex, and left postcentral gyrus, and positive correlations in the prefrontal cortex and left post central gyrus. However, in the whole sample analysis, more positive associations were observed in the frontal and parietal cortices at cut offs of 30-45%, whereas in the per site analysis, these effects were observed after 40-50% exclusion.

In these analyses, we also modelled the relationship between the percent of participants excluded, and the number of significant regions, as well as between the percent excluded and the effect size of the regions with the strongest effects. For both analyses, there was an initially strong negative relationship between percent excluded and number of significant regions. For the per-site analysis, the inflection point where the curve began to level off was 20; for the whole sample analysis, it was 15. The relationship between percent excluded and partial r values varied by region, but a similar pattern was seen to that of the number of significant regions .

Participants excluded per site

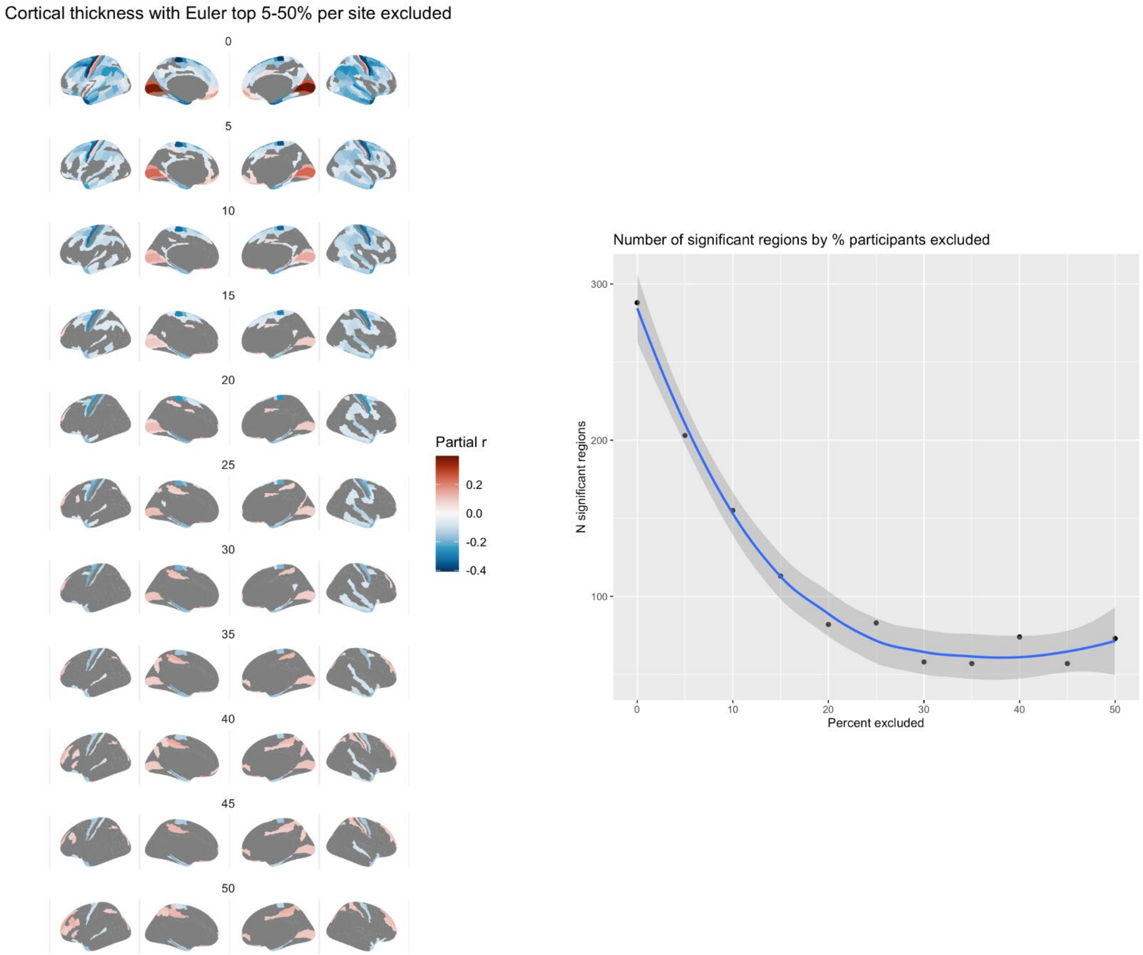

Fig S3.3.1. Effect of Euler number on CT after excluding top percent of participants per site, in intervals of 5% (left), and relationship between number of regions showing a significant relationship and percent of participants excluded (right).

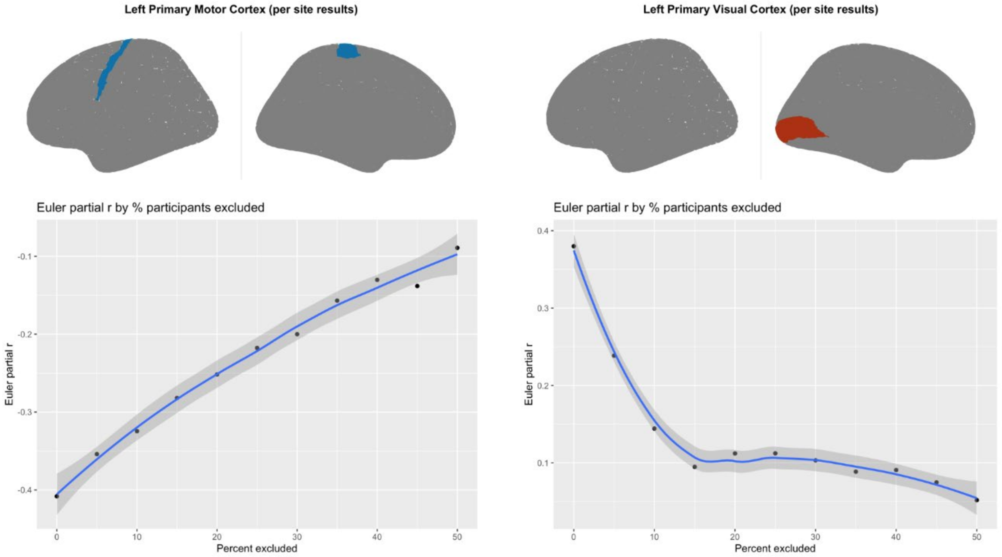

Fig S3.3.2. Relationship between correlation strength (partial r correlation for Euler number) and percent of participants excluded for two of the regions showing the strongest relationship: primary motor cortex (left) and primary visual cortex (right).

Participants excluded from whole sample

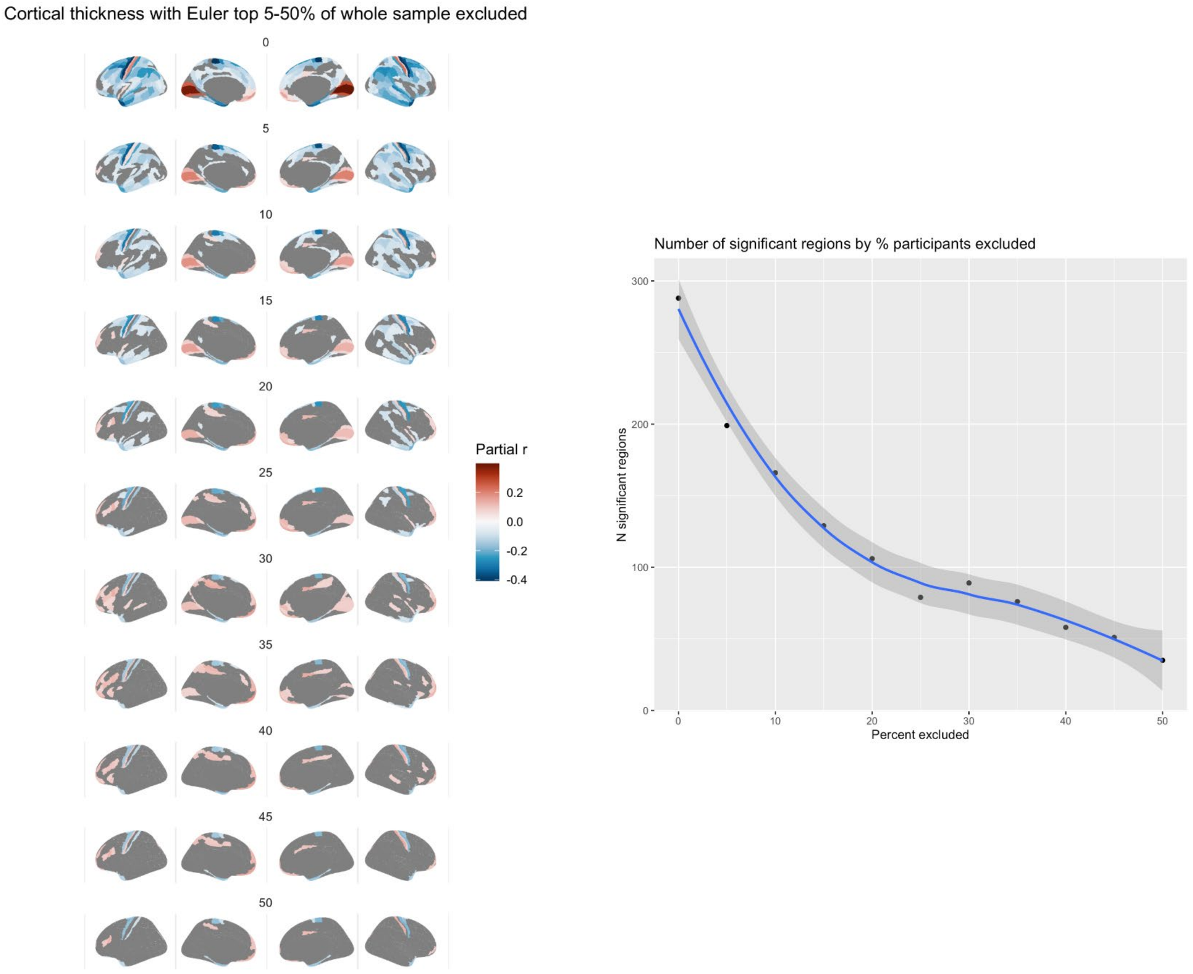

Fig S3.3.3 Fig Effect of Euler number on CT after excluding top percent of participants based on the whole sample, in intervals of 5% (left), and relationship between number of regions showing a significant relationship and percent of participants excluded (right).

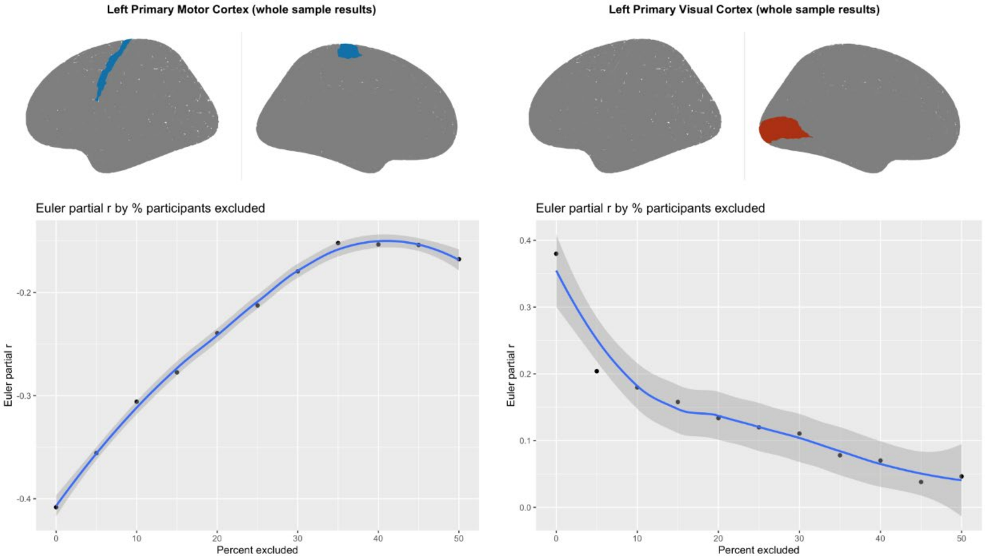

Fig S3.3.4. Relationship between correlation strength (partial r correlation for Euler number) and percent of participants excluded for two of the regions showing the strongest relationship: primary motor cortex (left) and primary visual cortex (right).

3.4 SA and CV: FSQC thresholding

When thresholding by FSQC for SA and CV, there was a stark drop off in significant effects even when excluding only those above a score of 2.5, with very few significant regions remaining. Subthreshold maps (i.e. including non significant regions) showed a mix of positive and negative partial correlations after initial thresholding

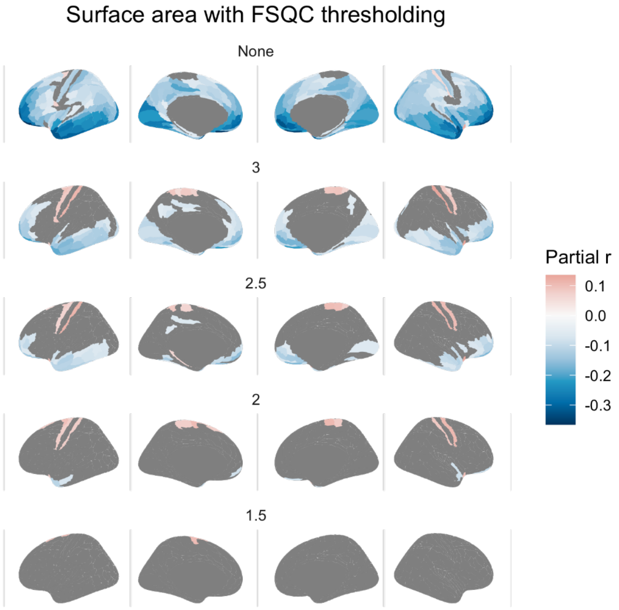

Fig S3.4.1 Relationship between SA and FSQC at various FSQC thresholds

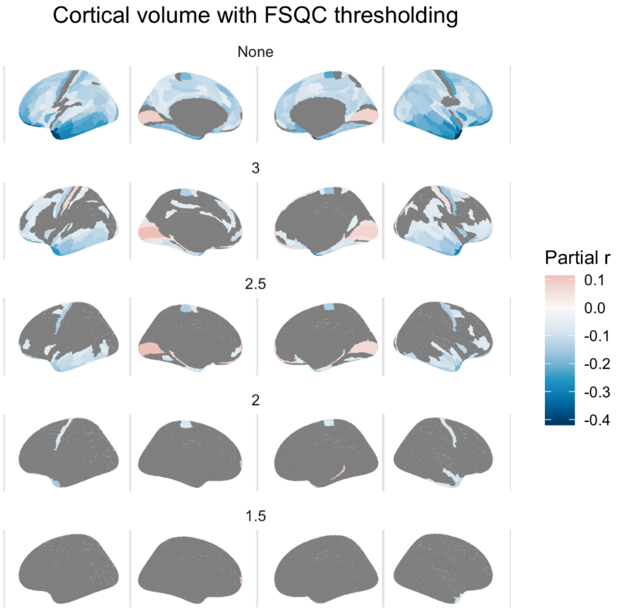

Fig S3.4.2 Relationship between CV and FSQC at various FSQC thresholds

3.5 SA and CV: Euler thresholding

For SA and CV, when thresholding by Euler number, results largely demonstrated positive rather than negative correlations after even the least stringent threshold.

This appears to be largely due to different regions becoming significant, rather than a reversing of effect direction within regions. Positive results were not attenuated with lowering thresholds; in fact, they were observed in more widespread regions by the most stringent threshold. For SA, after thresholding at any of the MAD cut off points, largely positive significant correlations were observed, primarily in the superior frontal and temporal cortices, and parietal regions. For CV, the same is true, but primarily in prefrontal and parietal cortices. It should also be noted that when thresholding using FSQC, while very few regions remain significant after exclusions, spatial patterning is similar to those observed here.

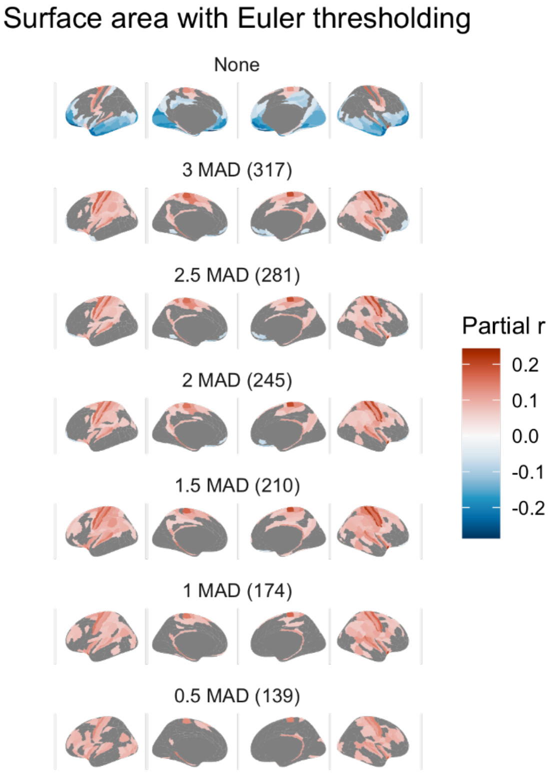

Fig S3.5.1 Relationship between SA and Euler at various Euler thresholds

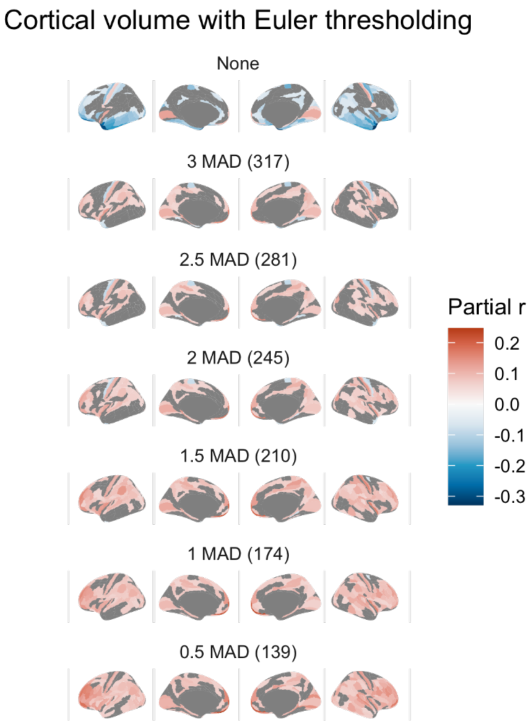

Fig S3.5.2 Relationship between CV and Euler at various Euler thresholds

4. Diagnosis

4.1 Replication in CMI and POND datasets

Most results in both CMI and POND indicate thicker cortex in autism relative to controls. In CMI, more associations are observed after accounting for quality, most notably the superior temporal gyrus (though not passing FDR correction). Negative associations disappear after accounting for quality. In POND, results do not change drastically with quality control. More regions pass FDR correction, most notably significantly thicker cortex in the superior temporal gyrus, which has previously been reported (and observed in ABIDE and CMI), and many negative associations between diagnosis and CT (autism < controls) disappear or are attenuated, similar to effects observed in ABIDE.

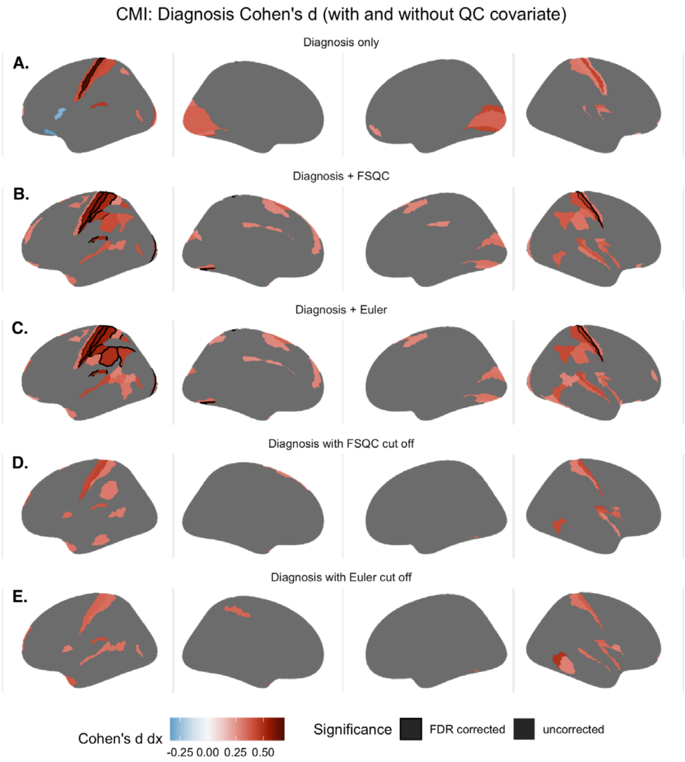

Figure S4.1.1. Impact of autism diagnosis on cortical thickness (Cohen's *d*) in the CMI Healthy Brain Network dataset without accounting for image quality (A), when controlling for FSQC (B) or Euler (C), and thresholding by FSQC (D) and Euler (E). Significant regions passing 5% FDR are shown with a black border; other regions are subthreshold (i.e., not surviving FDR) differences. Most results indicate thicker cortex in autism relative to controls and more associations are observed after accounting for quality.

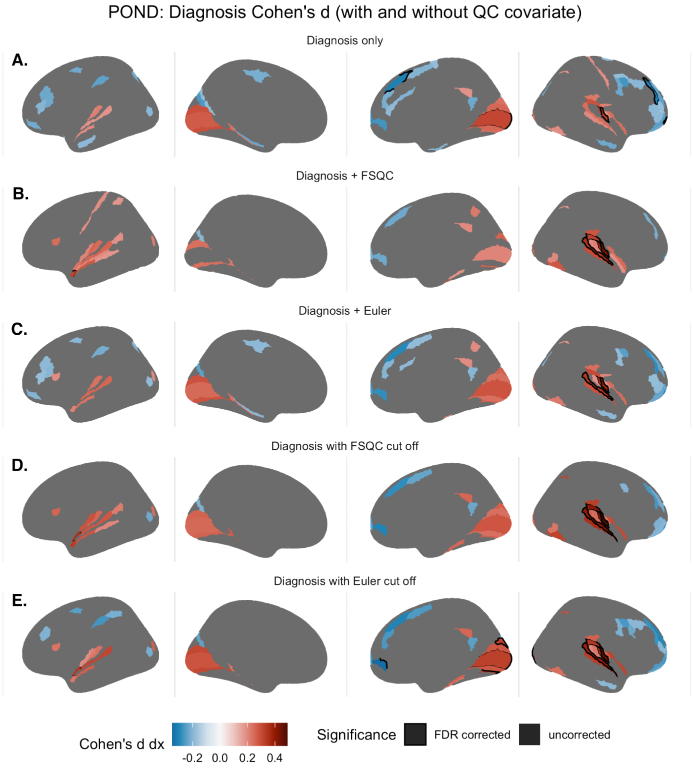

Figure S4.1.2. Impact of autism diagnosis on cortical thickness (Cohen's *d*) in the POND dataset without accounting for image quality (A), when controlling for FSQC (B) or Euler (C), and thresholding by FSQC (D) and Euler (E). Significant regions passing 5% FDR are shown with a black border; other regions are subthreshold (i.e., not surviving FDR) differences. More significant associations are observed after accounting for quality, and almost all indicate thicker cortex in autism relative to controls.

4.2 Diagnosis with cut off points

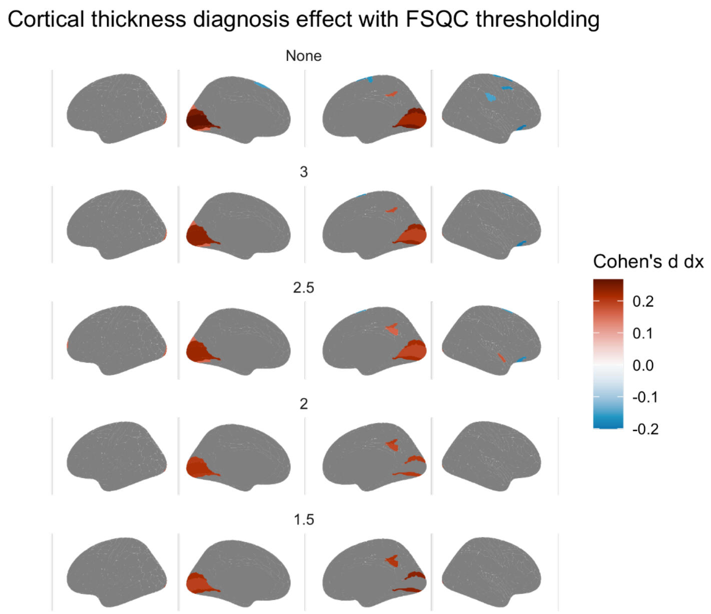

Fig S4.2.1. Effect of diagnosis (Cohen’s *d*) on cortical thickness at various thresholds of FSQC

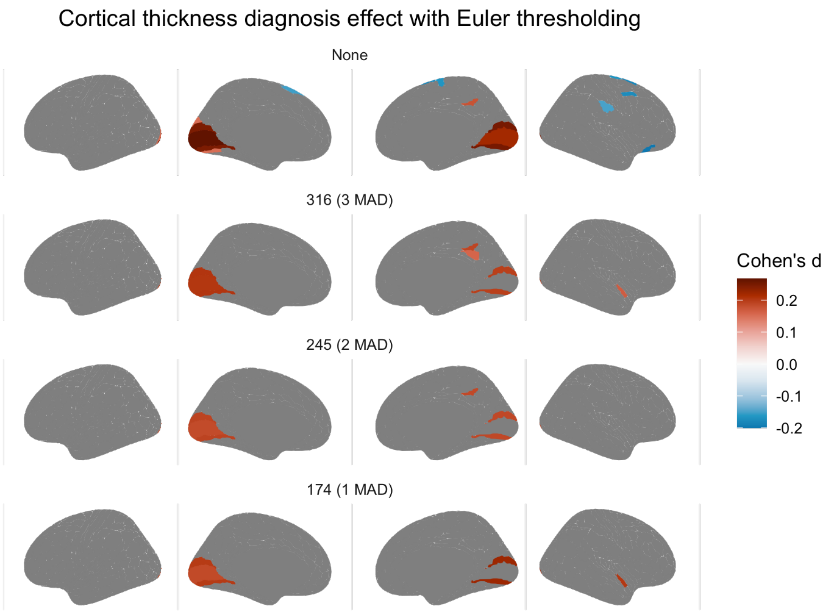

Fig S4.2.2. Effect of diagnosis (Cohen’s *d*) on cortical thickness at various thresholds of Euler

4.3 Diagnosis FSQC threshold + Euler

We also combined these two approaches, by applying a threshold based on FSQC, and also controlling for Euler. In this analysis, results were very similar to the previous FSQC thresholding analysis, with the exception of additional significant regions (autism > controls) in the superior frontal and temporal regions.

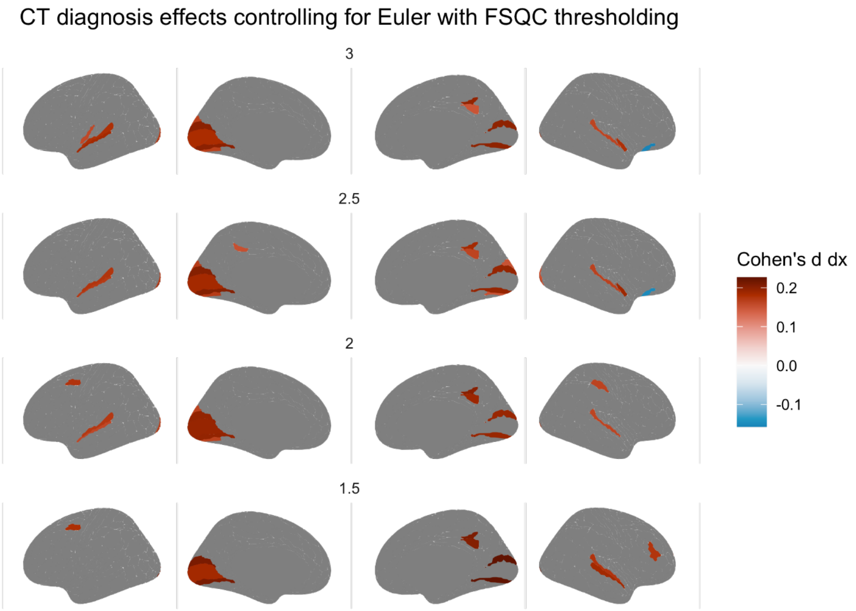
CT diagnosis effects controlling for Euler with FSQC thresholding
3
2.5
2
1.5
Cohen's d dx
0.2
0.1
0.0
-0.1

Fig S4.3. Effect of diagnosis on cortical thickness at various FSQC thresholds and controlling for Euler number.

4.4 Diagnosis interaction

We also examined the interaction between diagnosis and scan quality for both Euler and FSQC. For FSQC, a significant interaction was observed in multiple frontal, temporal and parietal regions, in particular in the right hemisphere. For Euler index, very few significant regions were observed, including small regions in the medial frontal cortices and left posterior parietal cortex. To explore this further, we examined the association between quality and CT in autistic individuals and controls separately. In regions where we observed a significant interaction, correlations appeared to be stronger in autistic individuals than controls, though in the same direction.

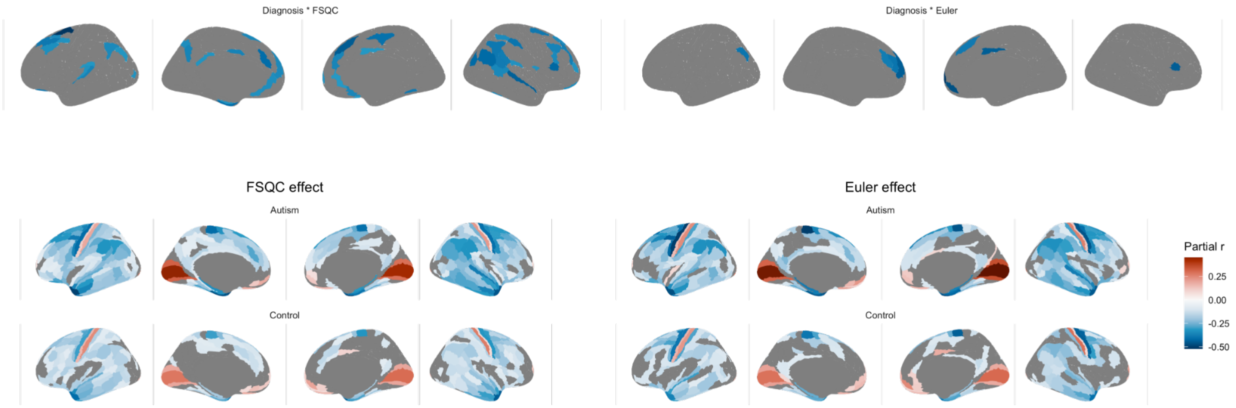

Fig 4.4 Interaction between diagnosis and FSQC (top right) and diagnosis and Euler number (top left), and effect of FSQC (bottom right) and Euler number (bottom right) in the autism and control groups, respectively.

4.5 Diagnosis SA and CV

Only very minimal group differences in SA and CV were observed, both with and without accounting for scan quality. For SA, without accounting for quality, lower SA in autistic individuals compared to controls was observed in right V1 only. After controlling for either FSQC or Euler, no significant results remained. For CV, no significant associations were observed before controlling for quality; however, after controlling for either FSQC or Euler, a significant effect was observed in the superior temporal gyrus, in which autistic individuals had greater cortical volume relative to controls.

SA:

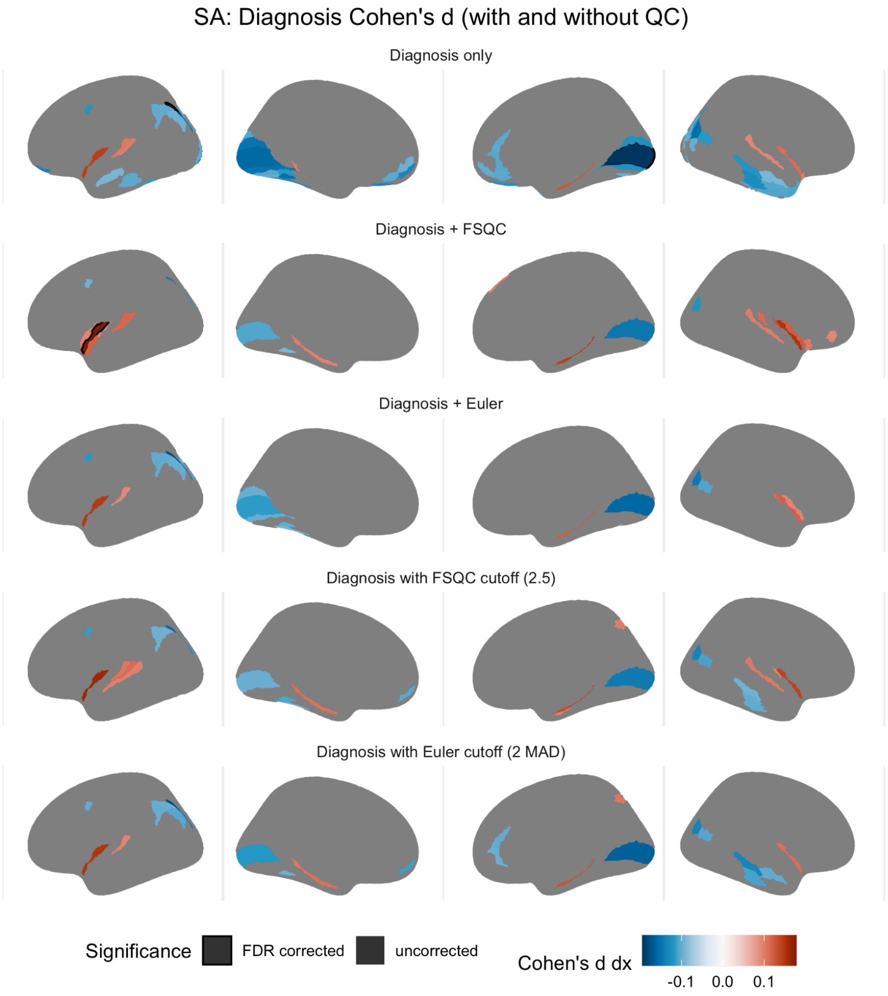
SA: Diagnosis Cohen's d (with and without QC)
Diagnosis only
Diagnosis + FSQC
Diagnosis + Euler
Diagnosis with FSQC cutoff (2.5)
Diagnosis with Euler cutoff (2 MAD)
Significance
FDR corrected
uncorrected
Cohen's d dx
-0.1
0.0
0.1

Fig S4.5.1. Impact of autism diagnosis on cortical surface area (Cohen's *d*) without accounting for scan quality (A), when controlling for FSQC (B) or Euler (C), and thresholding by FSQC (D) and Euler (E). Significant regions passing 5% FDR are shown with a black border; other regions are subthreshold (not surviving FDR) differences.

CV:

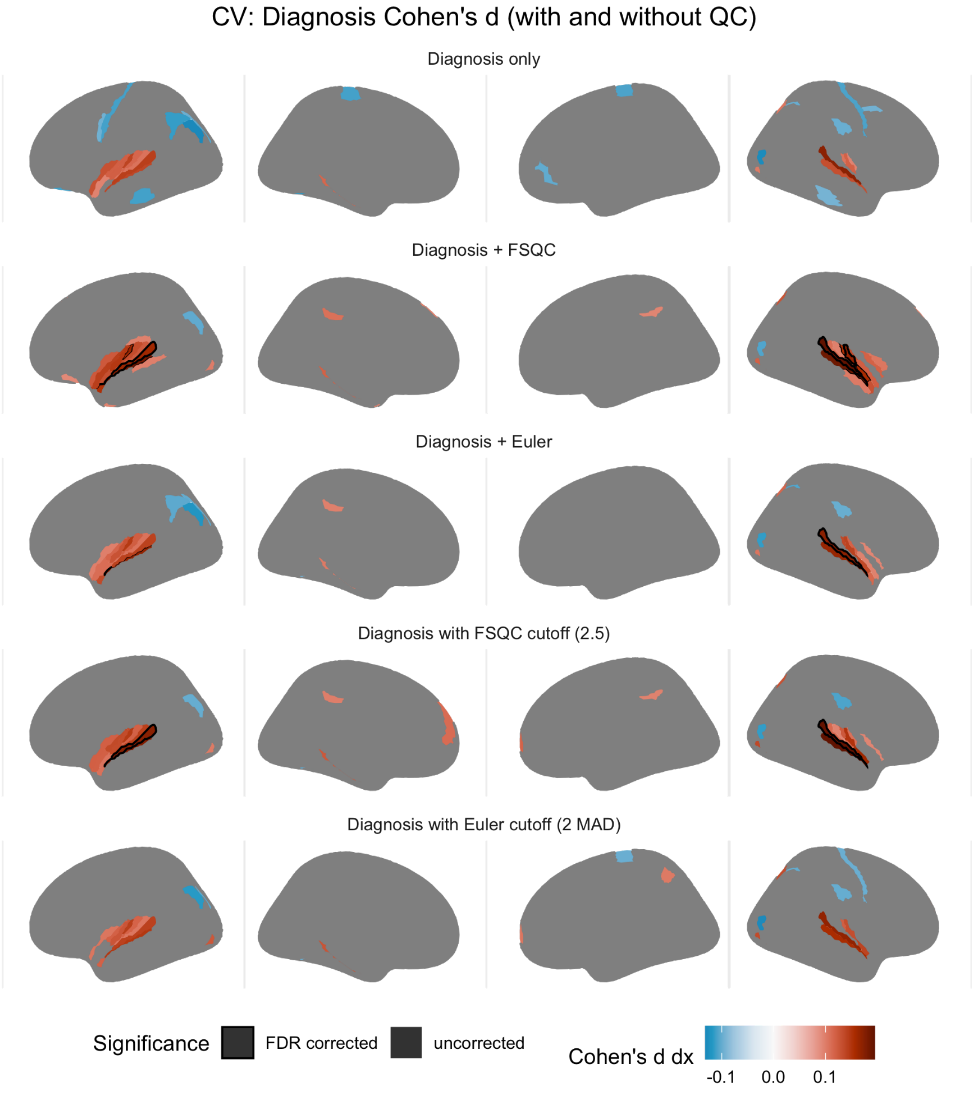
CV: Diagnosis Cohen's d (with and without QC)
Diagnosis only
Diagnosis + FSQC
Diagnosis + Euler
Diagnosis with FSQC cutoff (2.5)
Diagnosis with Euler cutoff (2 MAD)
Significance
FDR corrected
uncorrected
Cohen's d dx
-0.1
0.0
0.1

Fig S4.5.2. Impact of autism diagnosis on cortical volume (Cohen's *d*) without accounting for scan quality (A), when controlling for FSQC (B) or Euler (C), and thresholding by FSQC (D) and Euler (E). Significant regions passing 5% FDR are shown with a black border; other regions are subthreshold (not surviving FDR) differences.

References

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., … Milham, M. P. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*, 170181.

Bethlehem, R. A. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*(7906), 525–533.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS One*, *12*(9). https://doi.org/10.1371/JOURNAL.PONE.0184661

Hoffman, G. E., & Schadt, E. E. (2016). variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, *17*(1), 483.

Zarrar Shehzad, Steven Giavasis, Qingyang Li, Yassine Benhajali, Chaogan Yan, Zhen Yang, Michael Milham, Pierre Bellec and Cameron Craddock. (2015). *The Preprocessed Connectomes Project Quality Assessment Protocol - a resource for measuring the quality of MRI data*. *9*. https://doi.org/10.3389/conf.fnins.2015.91.00047