OXFORD UNIVERSITY PRESS | Cerebral Cortex

Influence of processing pipeline on cortical thickness measurement

Journal:	*Cerebral Cortex*
Manuscript ID	CerCor-2019-00888.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kharabian Masouleh, shahrzad; Research Centre Jülich, INM-7 Eickhoff, Simon B; Research Center Jülich, Institute of Neuroscience and Medicine, INM-1; Heinrich Heine University, Institute of Clinical Neuroscience and Medical Psychology Zeighami, Yashar; McGill University, Neuroscience Lewis, Lindsay; McGill University, Neuroscience Dahnke, Robert; Jena University Hospital, Department of Neurology Gaser, Christian; University of Jena, Department of Psychaitry Chouinard-Decorte, Francois; Montreal Neurological Institute, McConnell Brain Imaging Center Lepage, Claude; McGill University, Montreal Neurological Institute Scholtens, Lianne; Vrije Universiteit Amsterdam, Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research Hoffstaedter, Felix; Research Center Jülich, Institute of Neuroscience and Medicine Glahn, David; Yale, Psychiatry Blangero, John; University of Texas Rio Grande Valley School of Medicine, South Texas Diabetes and Obesity Institute Evans, Alan; Montreal Neurological Institute and Hospital, McGill University Genon, Sarah; Forschungszentrum Julich, Institute of Neuroscience and Medicine (Brain and Behaviour, INM-7) Valk, Sofie; Research Centre Jülich, INM-7
Keywords:	in-vivo cortical thickness, software comparison, reliability, replicability, inter-individual variability

Influence of processing pipeline on cortical thickness measurement

Shahrzad Kharabian Masouleh[1], Simon B. Eickhoff[1,2], Yashar Zeighami[3], Lindsay B. Lewis[3], Robert Dahnke[4,5], Christian Gaser[5], Francois Chouinard-Decorte[3], Claude Lepage[3], Lianne H. Scholtens[6], Felix Hoffstaedter[1], David C. Glahn[7], John Blangero[8], Alan C. Evans[3], Sarah Genon[1], Sofie Valk[1,2]

1. Institute of Neuroscience and Medicine (INM-7: Brain and Behavior), Research Centre Jülich, Jülich, Germany.
2. Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
3. Montreal Neurological Institute, McGill University, Montreal, Quebec.
4. Center of Functionally Integrative Neuroscience, Department of Clinical Medicine, Aarhus University.
5. Department of Psychiatry and Department of Neurology, Jena University Hospital, Jena, Germany.
6. Connectome Lab, CTG, CNCR, VU Amsterdam, Amsterdam, Netherlands.
7. Department of Psychiatry, Yale School of Medicine, New Haven, USA.
8. South Texas Diabetes and Obesity Institute and Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA.

Correspondence:

Shahrzad Kharabian Masouleh.
Brain and Behaviour (INM-7)
Institute for Neuroscience and Medicine
Research Centre Jülich
52425 Jülich
E-mail: s.kharabian@fz-juelich.de
Tel: +49 2461 61-1409

Abstract:

In recent years, replicability of neuroscientific findings, specifically those concerning correlates of morphological properties of gray matter, have been subject of major scrutiny. Use of different processing pipelines and, differences in their estimates of the macroscale gray matter may play important role in this context. To address this issue, here, we investigate the cortical thickness (CT) estimates of three widely-used pipelines. Based on analyses in two independent large-scale cohorts, we report high levels of within-pipeline reliability of the absolute CT-estimates, and comparable spatial patterns of CT-estimates across all pipelines. Within each individual, absolute regional thickness differed between pipelines, indicating that *in-vivo* thickness measurements are only a proxy of actual thickness of the cortex, which shall only be compared within the same software package and thickness estimation technique. However, at group level, CT-estimates correlated strongly between pipelines, in most brain regions. The smallest between-pipeline correlations were observed in para-limbic areas and insula. These regions also demonstrated the highest inter-individual variability and the lowest reliability of CT-estimates within each pipeline, suggesting that structural variations within these regions should be interpreted with caution.

Keywords: *in-vivo* cortical thickness, software comparison, reliability, replicability, inter-individual variability.

Gray matter (GM) contains the most neuronal cell bodies in the brain, and its structure changes considerably in the course of development, aging, and in disorders. These variations can be approximated *in-vivo* using structural MRI by measuring variations in macroscopic properties of the cortex. Over the last decade, many studies have utilized imaging-derived macroscopic GM properties to assess neurobiological changes in cortical structure as a function of development, aging, and pathology, as well as inter-individual behavioral variations (Kanai and Rees 2011; Fjell et al. 2014, 2015; Walhovd et al. 2017).

A frequently used macroscopic feature of the cortical structure is its thickness, which characterizes the distance between the gray-white interface (inner boundary) and the pial interface (outer boundary). Despite its seemingly straightforward definition and interpretation, calculation of cortical thickness from MR images of the highly folded human cortex is non-trivial, relying on precise identification of the boundaries between tissue types, as well as the metric used to quantify the thickness of the cortex that lies between these identified boundaries (MacDonald et al. 2000; Lerch 2001; Han et al. 2004; Lerch and Evans 2005; Das et al. 2009). In this respect, utility of manual assessments are limited, not only because they are labor intensive, but also due to the difficulty of accurate manual identification of tissue boundaries, for example due to blurring of the gray-white boundary, as well as within highly folded areas. Therefore, it is inevitable to rely on automatic methods for characterizing cortical thickness from *in-vivo* MRI, in particular among large cohorts of individuals.

Currently, several software packages provide algorithms for automatic estimation of cortical thickness. These routines can be broadly divided into surface-based and volume-based approaches and are commonly validated by focusing on assessment and comparison of cortical thickness of few selected cortical areas, with manual measurements from histological sections (Fischl and Dale 2000; Cardinale et al. 2014) from limited number of individuals. Though these validation studies report encouraging findings about correlation of cortical thickness estimated from automatic and manual segmentations of brain structure, they also demonstrate that the

thickness estimate from *in-vivo* MRI is only a proxy of the histological measurement of the thickness of the cortex, in particular within regions with blurred boundaries between gray and white matter.

Despite importance of such validation studies, the limited number of participants, as well as the focus on specific regions of interest (but see (Wagstyl et al. 2018) for a whole cortex solution using BigBrain) and pipelines, make interpretation and generalizability of these comparisons challenging in the era of 'Big Data Neuroscience'. Over the last decade the number of participants used in structural imaging studies has increased considerably, from tens to thousands of individuals (Van Essen et al. 2013; Miller et al. 2016; Bearden and Thompson 2017; Smith and Nichols 2018). Such investigations often relate small sub-millimeter variations of the automatically-determined cortical thickness to behavioral or clinical outcomes (Miller et al. 2016). These large-scale *in-vivo* assessments play central role in shaping our understanding of brain variability, and are thus increasing in frequency. Therefore, in the absence of an automatic pipeline, providing *in-vivo* ground truth of macro-scale neuroanatomical variation, it is crucial to ensure reliability of the conclusions drawn from these studies through assessment of within-tool robustness and between-tool reproducibility of the estimated thickness, which can further be useful in interpreting measured macro-anatomical variability as well as its modulating factors.

A handful of recent studies have attempted to perform such between-tool replicability assessment of cortical thickness estimates, within samples of healthy individuals or in patients (Clarkson et al. 2011; Gronenschild et al. 2012; Dahnke et al. 2013; Tustison et al. 2014; Li et al. 2015; Martínez et al. 2015; Dickie et al. 2017; Lewis et al. 2017; Righart et al. 2017; Seiger et al. 2018). While some studies have demonstrated high levels of within-pipeline test-retest reproducibility (Dahnke et al. 2013; Lewis et al. 2017), results of comparison of cortical thickness estimations across different versions of the same package have been less straightforward (Gronenschild et al. 2012; Dickie et al. 2017). Also, though global thickness

estimates have been shown to correlate highly across different software packages (Righart et al. 2017; Seiger et al. 2018), spatial distributions of cortical thickness estimates have demonstrated marked regional between-tool differences (Clarkson et al. 2011; Li et al. 2015; Martínez et al. 2015; Dickie et al. 2017). Moreover, an investigation of brain-behavior relationships, using thickness estimates from three different pipelines (Cortical pattern matching, CIVET and BrainSuite) demonstrated considerable variations in the spatial pattern of associations between cognitive scores and cortical thickness measurements across tools, as well as irreproducible associations within tools (Martínez et al. 2015). These ambiguous reports raise concerns about the reproducibility of cortical thickness estimates and the biological validity of the conclusions that are derived from its inter-individual variability.

However, calling the available cortical thickness estimation pipelines unreliable and questioning the replicability of all the literature studying variability of cortical thickness using these tools might be premature. In particular, comparing cortical thickness processing pipelines at different software development stages (i.e. beta version as opposed to stable releases (Righart et al. 2017; Seiger et al. 2018)) can partly explain the differences in the derived conclusions. Furthermore, proper quality control of the outcomes as well as characteristics of the samples under study may influence performance of the different pipelines.

In the current work, we aimed to comprehensively study and compare cortical thickness estimations from three software packages with large user-based communities (FreeSurfer, CIVET, CAT) using two large single-site datasets of healthy individuals, with different age ranges. In doing so, our approach considers potential biases due to scanning sites, population differences, and confounds due to age-related changes in macroscale gray matter anatomy. **Importantly, in each step of analysis, for example in the group-wise registration step, we try to adhere to the pipelines' originally implemented (or suggested) routine. As such our findings shall remain useful for comparison with group-level studies that use only one of these pipelines.**

After visual inspection of each pipeline's output, the quality-controlled thickness metrics were used to assess the following three main goals:

- *Within-pipeline* regional thickness variability across all participants, within each cohort.
- *Between-pipelines* thickness variability across all participants, globally and regionally, to identify regions in which cortical thickness estimations differed significantly between pipelines. Commonality of the pipelines in capturing similar *inter-individual* variabilities was also assessed by means of correlations, within each sample.
- Finally, for each pipeline *test-retest reliability* of the cortical thickness estimations was assessed within a subgroup of 143 individuals.

In doing so, we aimed to provide detailed insight on global and regional robustness of cortical thickness estimations of the commonly used pipelines, generating a basic understanding about factors that may influence replicability of studies that characterize structural variations and highlighting regions in which inter-individual variability in cortical thickness should be interpreted with caution.

Methods:

***Participants and acquisition parameters**:*

Two large single-site datasets with high quality T1-weighted MRIs of healthy adults were used to assess the between-individual variability and between-pipeline replicability of cortical thickness estimation. The first dataset is the publicly available data from the Human Connectome Project (HCP; http://www.humanconnectome.org/), consisting of young healthy adults. HCP comprises data from 1113 individuals (656 females), with mean age of 28.8 years (standard deviation (SD) = 3.7, range = 22–37). The full set of inclusion and exclusion criteria are described elsewhere (Glasser et al. 2013; Van Essen et al. 2013). **Two high-resolution (isotropic 700 µm) 3D MPRAGE T1-weighted images were acquired using the HCP's custom 3T Siemens Skyra scanner with the following parameters: TE/TR/TI = 2.14/2400/1000 ms, field of view (FOV) =224 mm, flip angle=8°, Bandwidth (BW)=210 Hz per pixel.** Two T2-weighted images were also acquired with identical geometry (TE/TR = 565/3200 ms, variable flip angle, BW = 744 Hz per pixel). Full description of MRI protocols of the HCP is previously described in (Glasser et al. 2013). Images underwent gradient nonlinearity correction (using acquired B1 bias field) and the two scans of each modality were co-registered and averaged. This average image is downloaded from (https://wiki.humanconnectome.org/display/PublicData/).

The second dataset consisted of 902 individuals (546 females), with mean age of 43.2 years (SD = 14.8; range = 18-85) from randomly selected families of Mexican-American descent who live in San Antonio, Texas, USA, enrolled in the Genetics of Brain Structure and Function Study (GOBS) before December 31st, 2012 (McKay et al. 2014). The GOBS study is a collaborative effort between Texas Biomedical Research Institute, University of Texas Health Science Center at San Antonio (UTHSCSA) and Yale University School of Medicine. Exclusion criteria were MRI contraindications, history of neurological illness, stroke or other major neurological events. All MRI images were acquired at the UTHSCSA Research Imaging

Center on a Siemens 3T Trio scanner (Siemens, Erlangen, Germany). **High-resolution (isotropic 800 µm) 3D Turbo-flash T1-weighted images were acquired with the following parameters: TE/TR/TI = 3.04/2100/785 ms, FOV = 200 mm, flip angle = 13◦, BW = 200 Hz per pixel.** Seven images were acquired consecutively using this protocol for each subject. Native T1-weighted MRI scans were co-registered and were further corrected for non-uniformity artifacts with the N3 algorithm (Sled et al. 1998) and averaged, for each participant over the seven scans, to increase signal-to- noise ratio and reduce motion artifacts **(Kochunov et al. 2006).**

Participants for assessment of within-tool reliability of cortical thickness:

143 participants of the GOBS cohort (100 females; mean age: 49.5 years; SD = 12.1; range = 26-81) with seven good quality scans (as defined subjectively by FCD) were selected for further analysis, to assess within-subject replicability of cortical thickness estimation for each pipeline. In order to increase signal-to-noise ratio (SNR) of the test-retest scans, the seven scans were divided into two groups, the *odd* and *even* acquisitions. Following estimation and correction of bias fields using the N3 algorithm (Sled et al. 1998), odd acquisitions (1st, 3rd, 5th and 7th acquired T1-weighted MRIs) were co-registered **(Kochunov et al. 2006)**. Separately, the same procedure was performed for even acquisitions (i.e., 2nd, 4th and 6th acquired scans). This procedure resulted in a pair of T1-weighted MRIs for each of the 143 participants. The pairs of scans were generated by averaging different numbers of images (four odd scans versus three even scans), to enhance the difference in SNR between them. Comparing the cortical thickness estimations within each pair hence examines robustness of the tools to slightly different levels of noise.

Processing pipelines for cortical thickness estimation:

FreeSurfer version 6.0:

FreeSurfer analysis of the T1-weighted images for both datasets was performed using the default recon-all options (version (v) 6.0; (www.surfer.nmr.mgh.harvard.edu)). Briefly, the raw T1-weighted images are affine-registered to a 1 mm template (MNI305) and after normalization, removal of intensity bias-field and skull-stripping, white-matter voxels are identified based on intensity and neighbor constraints. The two hemispheres are then separated. By tiling the boundary of white-matter mass an initial white surface is created, which is further refined following intensity gradients of the white and gray matter to generate the final gray-white surface. This surface is then expanded to follow the intensity gradient of gray matter and cerebrospinal fluid (CSF), fitting the pial surface.

Cortical thickness at each vertex is computed as the average of two distances, i.e., from each vertex in the gray-white surface to the nearest point in pial surface and from the corresponding vertex in the pial surface to the nearest point in gray-white surface (Fischl and Dale 2000; Fischl 2012).

For the HCP data, in addition to the bias-corrected T1-weighted images, cortical thickness estimates derived using Freesurfer v5.3-HCP pipeline (https://github.com/Washington-University/HCPpipelines/blob/master/FreeSurfer/FreeSurferPipeline-v5.3.0-HCP.sh), are provided through the Amazon Web Services for download. The outcome of this pipeline (Glasser et al. 2013), which is optimized for the HCP data and further incorporates T2-weighted images into the Freesurfer analysis pipeline, is frequently used to report cortical thickness variations within the HCP sample. **In the current study we use this opportunity to compare outcome of this customized pipeline with the thickness estimates of the FreeSurfer v6.0 default pipeline, as well as CIVET v2.1.1 and CAT v12.5, on the HCP cohort. However, due to fundamental differences between this pipeline (inclusion of additional imaging modality, as well as other modifications tailored for the HCP data), as compared to default pipeline, results of comparisons of its outcomes are only presented in the Supplementary**

material and detailed discussions about the factors influencing these outcomes are considered beyond the scope of this manuscript.

CIVET v2.1.1:

Surface-extraction and cortical thickness estimation using CIVET were performed using version 2.1.1 (http://www.bic.mni.mcgill.ca/ServicesSoftware/CIVET). Briefly, T1-weighted MRIs are first transformed to stereotaxic MNI152 space, at 0.5 mm processing voxel resolution (Collins et al. 1994) and non-uniformity artifacts are corrected with the N3 algorithm (Sled et al. 1998) using the recommended N3 spline distance of 125mm for 3T T1-weighted scans. The bias-corrected volumes are then masked in the stereotaxic space (Smith 2002) and segmented into tissue classes (CSF, cortical GM, sub-cortical GM, WM) (Zijdenbos et al. 1998), from which the partial volume estimations are derived (Tohka et al. 2004). For the purpose of the white surface extraction, the ventricles and sub-cortical gray tissue classes are masked as white matter, and hyperintense T1-weighted voxels (representing blood vessels) are masked out. The hemispheres are separated and, for each hemisphere, an initial white matter surface of genus zero is obtained using a marching-cubes algorithm, which enforces the spherical topology of the cortical mantle.

This white surface, resampled at 40,962 vertices, is then fitted to the position of the maximum local T1-weighted intensity gradient at the gray-white boundary. The adjusted white surface is expanded to the classified border of GM and CSF to create the pial surface (called “gray surface” in CIVET) (Kim et al. 2005).

Cortical thickness is then measured as the distance between the “white” and “gray” cortical surfaces, in the native space framework of the original MR images, using the same double average approach as previously described for FreeSurfer.

CAT v12.5:

Unlike CIVET and FreeSurfer pipelines, estimation of cortical thickness using CAT (version 12.5 (www.neuro.uni-jena.de/cat/)) is assessed without extensive reconstruction of the surfaces. After initial preprocessing of the T1-weighted images including denoising (spatial-adaptive Non-Local Means), spatial registration, bias-correction and skull-striping, the images are segmented by an adaptive maximum a posteriori approach (Rajapakse et al. 1997) with partial volume model (Tohka et al. 2004). The two hemispheres are then separated at this step and the cerebellum and hindbrain are removed, while the ventricles, subcortical regions and (detected) white matter hyperintensities are filled. Within each hemisphere, for every voxel in the cortical GM, the closest (without passing a different boundary) voxel on the white matter boundary is estimated. The accurate location of the final white matter boundary is then defined using the intensity gradients along the normalized vector between each gray matter voxel and each intermediate boundary voxel.

In order to avoid explicit reconstruction of the outer boundary, thickness of gray matter is defined directly using the distance information from the final white matter boundary, using a projection-based thickness estimation approach. For every gray matter voxel, the distance to the nearest voxel in the final white matter-boundary is calculated and successors of this voxel are defined. Successor of a voxel (v1), is defined as its neighboring voxel (v2) in gray matter, whose distance to the white matter boundary is around one voxel greater than v1. Accordingly, voxels with no successors are considered as local maximum, and are assumed to be located at the CSF boundary; their distance from the nearest voxel in the white matter boundary results in the thickness of the cortex at that voxel. This thickness value is further corrected for GM voxels with more than 50% CSF contribution. See (Dahnke et al. 2013) for more details.

Quality control of the outcomes:

Prior to the analyses, results of each pipeline were visually inspected to ensure their adequate quality. In particular, the extracted pial and gray-white surfaces of FreeSurfer were overlaid on

each individual's T1-weighted scan (in freeview; https://surfer.nmr.mgh.harvard.edu/fswiki/FreeviewGuide). For consistency with the other two pipelines, we did not perform any manual corrections on the extracted surfaces, and thus participants with gross artifacts in surface extraction (e.g. due to susceptibility artifacts in the original image, low signal, excessive movement, or existence of a neurological abnormality) were discarded from further analysis.

Quality assessment for the CIVET pipeline was done in a similar manner, by overlaying each participant's derived white and gray surfaces on its T1-weighted scan (in Display; http://www.bic.mni.mcgill.ca/software/Display/Display.html). Again, participants with gross artifacts in surface extraction were discarded from further analysis.

As the CAT pipeline by default does not generate the gray-pial and gray-white surfaces, quality of the outcomes was assured by visually inspecting the following three outputs per individual: (1) normalized gray matter segments; (2) the location of individual's central surface, which roughly determines the 50% distance between the final white matter boundary and the boundary between gray matter and CSF; and (3) the spatial distribution of the z-scored cortical thickness for each participant. Participants with gross error in the first two criteria or spatially implausible cortical thickness distribution were then discarded from further analysis.

For the 143 individuals from the test-retest reliability analysis, visual quality control (QC) was performed on outcomes of all three pipelines, for both test and re-test scans. Participants with adequate quality of outcomes for all pipelines and in both scans were then retained for further analysis.

Assessment of regional cortical thickness:

To perform further group level analysis, regional cortical thickness was regionally collapsed using the parcellation scheme of (Schaefer et al. 2018). This parcellation scheme is based on the combination of a local gradient approach and a global similarity approach using gradient-

weighted Markov Random models, and in the context of the current study, is used to improve SNR and interpretability of the subsequent between-software comparison.

For each individual, thickness estimations from FreeSurfer were registered to the fsaverage surface (fsaverage - 163,842 vertices per hemisphere), upon which the parcels are represented. In particular, FreeSurfer uses folding pattern quantification to drive a non-linear surface-based inter-subject registration procedure that aligns the cortical folding patterns of each subject to a standard surface (fsaverage) space (Fischl et al. 1999). The registration is performed in spherical space. First, the subject's white surface is "inflated" to the shape of a sphere, and the geometry quantification of the white surface is transferred to the sphere. The location of the vertices on the sphere are then adjusted to minimize the overall cost to establish the correspondence. For each target vertex (on the fsaverage surface), the closest source vertex in the individual's spherical surface is found. The value of the quantity to be mapped (e.g., thickness) is then assigned from the individual's vertex to the target vertex, thus assigning each target vertex a value (Greve and Fischl 2018).

Individuals' thickness estimates from the CAT pipeline were also registered to fsaverage space, via a surface registration approach using anatomic features of cortical depth and curvature information (Yotter et al. 2011).

The standard surface within the CIVET pipeline is the MNI-ICBM152 symmetric surface (with 40,962 vertices per hemisphere). Surface registration is performed to resample each individual's cortical measurements to this template surface (akin to fsaverage in FreeSurfer), enabling vertex-wise group comparisons and application of surface parcellations for regional analyses. **In order to assess cortical thickness estimation of the CIVET pipeline from the same regions as the other pipelines, registrations between the MNI-ICBM152 template surface and fsaverage were assessed, allowing projection of Schaefer's parcellation on the MNI-ICBM152 template (Lewis et al. 2019).**

Here, we demonstrate the regional analysis using 400 parcels. For each participant, regional cortical thickness is assessed as a trimmed mean (10%) of thickness estimates within each parcel. Global thickness, per participant, is assessed as the mean of regional cortical thickness estimations over all parcels.

Statistical analysis:

For each sample, the *within-pipeline* estimates of cortical thickness estimation are summarized in terms of the distribution of the global thickness, spatial distribution of regional means, and SD, across individuals. The standard-deviation maps depict regions in which each pipeline demonstrated the highest variability across participants of one cohort. Comparison of these spatial maps and global estimates provide insight about commonalities and differences between cohorts and across different pipelines.

Between-pipeline comparison of the cortical thickness estimations is further investigated using paired t-tests, within each region as well as on the global thickness estimates. Two-sided significance threshold was set on $p < [0.05/(\#\text{ parcels} \times \#\text{ pairs of tools})]$, to identify regions with significantly different cortical thickness estimation between each pair of pipelines.

Furthermore, to investigate the degree to which each pipeline-pair depicts similar inter-individual variations, Spearman's rank correlations were assessed between tools, both globally, showing overall consistency of the cortical thickness estimations, and regionally, depicting regions with highest and lowest divergence between processing pipelines.

Assessment of test-retest reliability of cortical thickness measurements:

Evaluation of reliability of cortical thickness estimations within each pipeline is also performed at both the global as well as regional level. For each pipeline, mean absolute percent error (MAPE %), over all test-retest participants, is measured using the following formula:

$$(MAPE;\%) = \left(\frac{1}{n}\sum \frac{|c_1 - c_2|}{\frac{(c_1 + c_2)}{2}}\right) \times 100$$

Where c_1 and c_2 are cortical thickness estimations (global or parcel-wise cortical thickness) of the two average scans for each participant, and n denotes the number of participants. Given that the two average scans belong to the same participant, their cortical thickness estimates should, in theory, be identical. Therefore, the observed variations will be considered as measurement errors.

Results:

General qualitative remarks from visual quality control

4602 surface reconstructions, 2301 volumetric segmentations, and spatial CT-distribution maps were visually inspected to ensure the quality of the processing outcomes. Both datasets used in the current study had high quality T1-weighted MRIs, with very few images being affected by motion artifacts. **Within the HCP sample, 75 participants were discarded due to anatomical abnormalities, severe movement artefacts or other technical problems (see https://wiki.humanconnectome.org/pages/viewpage.action?pageId=88901591, issue codes A and B, for detailed explanation and examples of such issues).** Ten additional participants were excluded following visual QC of the FreeSurfer outcomes, while 21 participants were excluded following visual QC of CIVET results. Additional 24 participants were excluded following inspection of spatial distribution of CT estimates from the CAT-pipeline, resulting in a sample of 981 participants for further comparisons.

From the total of 902 GOBS participants, 30 individuals with substantial surface extraction errors were discarded in the FreeSurfer pipeline. Additionally, 10 were discarded in the CIVET pipeline and 30 participants were discarded after quality control of CAT results, due to abnormal regional distribution in cortical thickness, leaving 832 participants from the GOBS cohort for further analysis.

These exclusions were caused by gross errors in the processing pipelines. **Modest imperfections, such as small issues with brain extraction, specifically in the post-central gyrus within the FreeSurfer pipeline, and infrequent defects in white surface extraction resulting in generation of bridges across a sulcus in the CIVET pipeline, were ignored (see Supplementary Figure I-A(ii-iv)). Also, both the FreeSurfer and CIVET pipelines occasionally showed difficulties in placing the inner surface (gray/white boundary) accurately near the insular cortex (e.g. see Supplementary Figure I-A(i) and B(i)). Table 1 summarizes demographic characteristics of the remaining participants.**

--------Table 1--------

Distribution of regional and global mean cortical thickness and its variability across participants:

In both datasets, CAT and CIVET resulted in higher global cortical thickness estimations, compared to FreeSurfer (Table 2 and Figure 1). However, the difference between global cortical thickness estimations using CAT and FreeSurfer were less pronounced in GOBS cohort compared to the participants of the HCP sample.

----------Table 2----------

----------Figure 1----------

Spatial distributions of mean and standard deviation of regional cortical thickness across individuals, for each pipeline, are shown for HCP and GOBS samples in Figure 2. Overall, for all pipelines, cortical thickness showed the expected distribution of thinner estimated cortex in the visual and somatosensory cortices, while insular cortex, temporal poles as well as dorsal-medial prefrontal cortex showed highest cortical thickness measurements.

For comparison, Supplementary Figure II-A demonstrates distribution of the width of the cortical mantle (only on the left hemisphere) as measured by von Economo and Koskinas, by means of histological examinations of post-mortem brains (age: 30-40 years) (von Economo and Koskinas 1925). For detailed information about the manual segmentation of the von Economo–Koskinas atlas onto individual T1 scans and the subsequent construction of the digital atlas, see (Scholtens et al. 2016). Accordingly, the thickest cortex in their experiments was measured in the motor and premotor regions as well as the inferior medial temporal lobe, while the thinnest cortex was measured in the post-central gyrus and occipital lobe. Despite differences, specifically in the insular cortex and cingulate area, these results partly corroborate the distribution we observed in the MR-derived thickness estimates using all pipelines. The similarity of the spatial distributions between MR-derived cortical thickness and the histological assessments are quantified in Supplementary Figure II-B, for the HCP sample.

Here, labels of the von Economo–Koskinas atlas are identified on the fsaverage surface. Regional cortical thickness of each pipeline is then identified using similar procedure, as described for the Schaefer's parcels. In line with a previous publication (Scholtens et al. 2015), we observed moderate correlations (ranging between 0.57 and 0.66) between *in-vivo* thickness estimates of all pipelines and the histological measurements.

----------Figure 2----------

Considering the variability of the regional cortical thickness across participants, CIVET and CAT showed the lowest and highest regional SD, respectively.

In all tested pipelines, we observed highest SD in cortical thickness estimates in insular cortex, medial temporal pole, parahippocampal gyri, cingulate gyri and temporo-parietal junction across participants and cohorts. Furthermore, in both samples, CAT showed high SD in regions around the central sulcus (i.e. pre- and postcentral gyrus, specifically in the left hemisphere); see Figure 2. For each cohort, tables of mean and SD of cortical thickness within each parcel are shared (https://github.com/shahrzadkh/CT_replicability_reliability), for all three tested pipelines.

Between-pipelines comparison of cortical thickness estimations:

Comparison of the actual thickness estimates:

Regional paired t-test comparison of the cortical thickness estimates showed cortex-wide significant differences between all pipelines (Figure 3 and 4 show HCP and GOBS samples, respectively). Across both samples, CAT demonstrated consistently lower cortical thickness estimates within the pre- and post-central gyrus as well as within the visual cortex, compared to CIVET and FreeSurfer. In contrast, insular cortex and temporal lobes were estimated systematically thicker in CAT. Relative to both FreeSurfer versions (see also Supplementary Figure V), CIVET estimated the cortex as thicker, specifically within the medial occipital lobe.

While thickness estimations of CAT and CIVET were minimally different in the dorsal lateral frontal lobes, FreeSurfer comparably resulted in much lower thickness estimates within this region, in both samples.

As depicted in the Supplementary Figure V, Within the HCP sample, the two FreeSurfer processing pipelines also showed slight differences; thicker medial structures and thinner parahippocampal gyri and medial temporal poles were measured by FreeSurfer v5.3-HCP compared to the standard pipeline of FreeSurfer v6.0.

----------Figure 3----------

----------Figure 4----------

Between-pipelines correlation of cortical thickness estimates:

Despite the differences in the absolute value of thickness estimations between pipelines, correlation analysis indicated that inter-individual variabilities of the estimated global cortical thickness correlated highly across all pipelines (all correlations > 0.75; Figure 5 A, C), in both HCP and GOBS datasets. Accordingly, global cortical thickness in all pipelines correlated similarly with age of participants, within both samples (HCP: age-correlations $\approx$ -0.2; GOBS: age-correlations $\approx$ -0.6; Figure 5 B,D), emphasizing again that the inter-individual variabilities are depicted similarly in the global thickness estimates.

Regionally, in the HCP sample, thickness estimates of the two FreeSurfer versions correlated highly within most cortical parcels, with the exception of medial frontal lobe, insular cortex, parahippocampal gyri and medial temporal poles (Supplementary Figure V). Considering in addition the CAT and CIVET pipelines, cortical thickness estimates of the medial frontal cortex, insular cortex, cingulate, precentral and parahippocampal gyri as well as temporal poles showed lower correlations between all the pairs of tested pipelines in this cohort (Figure 3). In contrast, cortical thickness estimates within postcentral gyrus and superior temporal lobes showed highest between-pipeline correlation in this sample. Comparable observations were

made in the GOBS cohort. In particular, cortical thickness estimation of CAT showed lowest correlation with CIVET and FreeSurfer estimates within the orbitofrontal and pre-central gyri, superior parietal and occipital lobes, parahippocampal and cingulate gyri (Figure 4). Lowest correlations between CIVET and FreeSurfer thickness estimates were found in the dorsal lateral frontal lobe, parahippocampal and cingulate gyri as well as in the parietal lobule. In line with the HCP sample, highest correlations were found in the postcentral gyrus and superior temporal lobes.

Importantly, most regions with low between-pipeline correlations, including the parahippocampal and cingulate gyri, insular cortices, pre-central gyrus and occipital lobes, also showed high within-sample standard deviation of thickness estimates in each pipeline (comparing Figures 3 and 4 with Figure 2).

Regional associations between cortical thickness and age of participants are also demonstrated in the Supplementary Figure X, for both cohorts. Accordingly, these maps show a similar pattern of association across all pipelines, while differences are most visible within regions in which we have observed the lowest between-pipeline correlation of the thickness estimates, namely insular cortices, regions around the central sulcus, occipital lobes and temporo-parietal junction.

Test-retest reliability of cortical thickness measurements:

Out of the 143 participants for the test-retest analysis, visual inspection defined 115 individuals to have adequately good quality following processing by all three tested pipelines, and for both test-retest scans (2/5 participants excluded from FreeSurfer v6.0, 5/5 participants excluded from CIVET v2.1.1 and 6/8 individuals excluded from CAT v12.5, referring to the odd (first) and even (second) scan-groups, respectively).

All three tested pipelines showed high test-retest reliability. In particular, pairwise Spearman's correlation coefficients between the global cortical thickness estimations of both test and the

retest scans (i.e., odd and even scans) were above 0.96 (ρ_{CAT}: 0.99; ρ_{CIVET}: 0.97; $\rho_{FreeSurfer}$: 0.96). Regional distribution of MAPE, demonstrating the relative measurement error, showed that for all tested pipelines, more than 50% of the parcels had a MAPE bellow 3% (Table 3).

----------Table 3----------

----------Figure 6----------

Insular cortex, medial temporal pole, parahippocampal gyri and cingulate gyri showed reduced reliability of thickness estimations using all three tested pipelines (MAPE in these areas reach above 10%). In addition, CAT showed reduced reliability within the precentral gyrus and FreeSurfer demonstrated higher MAPE, i.e., reduced reliability, also in medial occipital lobe (Figure **6**).

Discussion:

In the current work, we investigated reliability and replicability of cortical thickness estimations from three software packages with large user-based communities (FreeSurfer, CIVET, CAT) using two large single-site datasets of healthy individuals. While all three pipelines, on average, demonstrated similar pattern of thickness distributions throughout the cortex, the estimated values differed significantly between pipelines. Using pair-wise comparisons in both samples, we highlighted regions in which each pipeline tends to under- or over-estimate cortical thickness relative to the two other pipelines. At the same time, inter-individual variability of cortical thickness covaried strongly between tools, suggesting relatively consistent under- or over-estimations across pipelines. Last, test-retest analyses demonstrated high reliability of the thickness estimations in most parts of the brain within each processing pipeline. However, regions with low between-pipeline correlation, mainly located in the parahippocampal and cingulate gyri, insular cortex, pre-central gyrus and occipital lobe, consistently demonstrated higher inter-individual variability, as well as low test-retest reliability, within each pipeline. In sum, using a big-data approach to reliability and replicability of cortical thickness, we observed that while inter-individual variabilities in cortical thickness estimates are comparably depicted across samples and pipelines, absolute values differ. Moreover, estimates of thickness in para-limbic and mid-occipital regions vary most within, as well as across individuals and also across pipelines. These results suggest that inter-individual variability of the cortical thickness within these regions is less reliably estimated and should be interpreted carefully.

Regional distribution of cortical thickness estimations:

In two large samples of healthy individuals, and across processing pipelines, regional cortical thickness estimates demonstrated a similar distribution pattern, with thinner cortex estimated in the visual and somatosensory regions and thicker cortex within insular cortex, temporal poles as well as medial prefrontal cortex and premotor areas. These observations are in line with the

previous findings in the literature and histological studies, demonstrating lower thickness of the cortical mantle in the somatosensory regions as well as in the visual cortex (Rowley et al. 2015). Notably, we observed a positive correlation between MRI-derived measures of cortical thickness, using all three pipelines, and width of the cortical mantle as reported in the histological work of von Economo and Koskinas (von Economo and Koskinas 1925). We acknowledge that the accuracy of this analysis is restricted by lack of correspondence between participants used for *in-vivo* and *ex-vivo* examinations and further inherent limitations of most *ex-vivo* measurements, including use of tissue samples only from the limited number individuals with unknown amounts of under-estimation due to post-mortem volume shrinkage and over-estimation due to selected sectioning procedure (Amunts et al. 2007; Scholtens et al. 2015). Despite these limitations, our results suggest a moderate level of agreement ($\rho > 0.57$) in the spatial distribution pattern of the cortical thickness, albeit with different absolute values, between the histological and MR-derived estimates, using all three pipelines.

At the same time, we observed that, in all samples and across pipelines, and unlike the histological assessments (see supplementary Figure II), insular cortex and the temporal pole were the regions with the thickest cortex in MR-based measurements. Visual quality control showed that these regions are also particularly prone to errors. For example, Supplementary Figure I-B (i) depicts one such failure, where the thin white matter between insula and claustrum is missed and hence the white surface of the posterior insula is misplaced. Similarly, the thin white matter within the temporal lobe and relatively higher impact of partial volume effects between gray and white matter within these regions, may lead to misplacement of white surface inside white matter. Such occasional imperfect surface definitions can in turn result in inflated cortical thickness estimates within these regions and its higher variability across participants.

Differences between *in-vivo* and *ex-vivo* thickness estimations can also arise from inherent limitation of the T1-weighted MRI signal (Natu et al. 2019). For example, higher myelin

content within motor and premotor areas increases the intensity of voxels in T1-weighted anatomical MR images, shifting the apparent gray–white boundary in these regions and thus resulting in under-estimation of cortical thickness from *in-vivo* measurements, in all pipelines.

***Distribution of inter-individual variability of cortical thickness estimates*:**

Our results demonstrated an important aspect of inter-individual variability that is depicted in standard deviation maps by all pipelines. In both young and older adults all pipelines showed high inter-individual variation of thickness within the posterior cingulate gyri, parahippocampal gyri, temporal pole, insular cortex, and temporo-parietal junction. The pattern of spatial distribution of average cortical thickness over all participants and its variation, within each tool, was confirmed in both HCP and GOBS sample, highlighting the replicability of these findings despite the characteristic differences between the two samples.

This observation can reflect true biological inter-individual variability of these regions. **Alternatively, the consistently higher variability in cortical thickness estimations within these regions might demonstrate higher frequency of processing errors that can occur in these areas. Our test-retest analyses, demonstrating low reliability of cortical thickness estimations within these regions, support this later explanation. This suggests that the heightened inter-individual variability in these areas is, at least partly, driven by common difficulty of all pipelines in correctly characterizing tissue boundaries in these regions. Specific structural properties, such as very thin local structures (white matter and/or CSF) or lower contrast in the MR-images are possible factors that can negatively influence accuracy of automatic pipelines in correctly identifying tissue boundaries of these regions. Accordingly, our results highlight that cortical thickness estimates within these regions may be prone to errors and are thus less reliable.** Higher variability of estimated cortical thickness within these less reliable regions can in turn obscure true biological variations and result in spurious findings, for example in studies linking brain structure to variation in non-

brain phenotype (Loken and Gelman 2017), but also group-wise differences. As such, findings within these regions should be taken with caution.

Between-pipeline comparison of cortical thickness estimation:

In line with previous work (Lerch and Evans 2005; Dahnke et al. 2013; Dickie et al. 2017; Lewis et al. 2018), we showed that different pipelines produce different absolute values of cortical thickness estimates, highlighting the dependency of *in-vivo* cortical thickness estimates on the processing software and the metric that is used to estimate the thickness of the cortex. Thus, *in-vivo* values of cortical thickness are only a proxy of the actual thickness of the cortex and are limited to the specific analysis pipeline (e.g. (Alvarez et al. 2019)). These differences complicate straightforward use of cortical thickness estimates, acquired from different pipelines, in a single study.

For example, **unlike CIVET and FreeSurfer pipelines, cortical thickness in CAT v12.5 is defined using a projection-based thickness estimation approach (PBT). Calculating the thickness using the same double average approach that is implemented in FreeSurfer (Tfs), instead of the PBT, would result in lower thickness values throughout the cortex. Influence of this modification is demonstrated in Supplementary Figures III-VIII.** As these figures show CAT cortical thickness using Tfs approach results in smaller values globally and in all cortical areas, across participants, bringing CAT's thickness estimates in the range of FreeSurfer's measurements. Yet, despite this change in the absolute value of thickness, correlations between the modified CAT-thickness estimates and CIVET and FreeSurfer outcomes remain unchanged, showing that the modification to the thickness measurement approach has not influenced the pattern of inter-individual variabilities.

Importantly, we observed that difference between pipelines is not uniformly distributed across the cortical mantle. Irrespective of the thickness estimation approach, CAT under-estimated thickness around the central sulcus, compared to the other pipelines. **The under-estimation of**

the cortical thickness around the central sulcus and specifically in motor cortex, might be attributed to its proportional higher myelin content (Rowley et al. 2015), which as mentioned earlier results in higher intensity of voxels in T1-weighted images and influence the apparent the gray/white contrast at this regions. While this effect can result in under-estimation of the cortical thickness in all MR-based automatic pipelines, it particularly is problematic for CAT, which relies strongly on actual intensities of the MRI images for identification of the white matter boundary (Dahnke et al. 2013). CIVET, on the other hand, estimated the cortex systematically slightly thicker than both the FreeSurfer pipelines. This is in line with recent observation, comparing surface estimates of CIVET v2.1.1 and FreeSurfer v6.0 with ground truth measurements from BigBrain (Lewis et al. 2018). Their findings demonstrate that both pipelines extract a similar and relatively accurate white surface (especially at higher resolutions), but that CIVET commonly over-estimates the gray surface placement (yielding higher thickness values), while FreeSurfer commonly under-estimates the gray surface placement (yielding lower thickness values). Accordingly, findings of the current study suggest that such over- and under-estimations in both pipelines are also occurring in in-vivo measurements, resulting in systematically thicker thickness estimates in CIVET, compared to FreeSurfer estimates.

Comparing the two FreeSurfer pipelines (Supplementary Figure V), version 5.3-HCP resulted in thicker medial structures and thinner parahippocampal gyri and medial temporal poles, compared to the default version 6.0.

Though our results indicate tool-based differences in the cortical thickness estimates within each participant, pair-wise correlations suggest that inter-individual variation that is depicted by these pipelines are comparable. In particular, cortical thickness estimates within the post-central gyrus, the superior-temporal lobes, as well as the lateral frontal lobes showed strongest between-pipelines correlation.

Comparability of the outcomes of neuroimaging software, in particular in the absence of *in-vivo* ground-truth measurements, plays a central role in the discussions about replicability of neuroscientific findings (Muhlert and Ridgway 2016). In this regard, previous studies showing strong discrepancies of the outcomes from widely used cortical thickness estimation pipelines are particularly alarming (Martínez et al. 2015; Dickie et al. 2017). However, our comparative examinations on two samples from two sites and with different characteristics, accompanied by thorough quality inspection, suggest that estimated thickness from different pipelines show high level of correlation globally and in most brain regions. It is thus likely that lack of reproducibility in studies looking at between-subject variations in cortical thickness, defined using either of these pipelines, in association with behavior or disease status cannot be solely attributed to the difference in applied software.

Quality assurance of the outcomes of neuroimaging software is an important, yet frequently overlooked task, particularly in studies with high number of participants. Accordingly, careful exclusion of participants with suboptimal processing outcomes for each tool is best achieved by interactively visualizing outcome surfaces, as opposed to quality assurance using selected screenshots of the surface overlays. This may partially explain the superior similarity between cortical thickness estimations of the three tested pipelines in this study as compared to some previous studies (Martínez et al. 2015; Dickie et al. 2017).

Yet, the spatial maps of pair-wise correlations also highlighted regions in which cortical thickness estimates of different pipelines correlated poorly. Medial orbitofrontal cortex, cingulate and parahippocampal gyri as well as occipital lobe showed the lowest correlations in all comparisons and within both samples. Furthermore, most regions that showed high standard deviation within one pipeline (for example regions around the central sulcus, showing high SD in CAT pipeline) also showed lower between-pipeline correlations, suggesting that the variability in those brain regions are potentially driven by noise and errors and therefore depicted differently by different pipelines.

Test-retest reliability of cortical thickness measurements and influence of noise:

Test-retest reliability of neuroimaging outcomes is the stepping stone towards reliable science, and its importance is on par with between-pipeline replicability of outcomes. In this regard, our results showing near perfect (i.e. MAPE < 3%) test-retest reliability of cortical thickness estimation across most brain regions, for all tested pipelines, are very encouraging. It is important to note that in order to assess robustness of the cortical thickness estimations to slight differences in noise level, we generated the two test-retest scans by averaging different numbers of MR images. Similarity of the cortical thickness estimates between these two scans, despite the possible different noise levels, confirms that all the tested tools can robustly deal with small variations of noise in the input images.

Another informative aspect of the test-retest analysis is the regional distribution of MAPE. In particular while all pipelines had very high reliability (i.e. less than 2% variation) in the frontal and superior temporal lobes, the stability of the cortical thickness estimates is lower within the insular cortex, temporal poles, temporo-parietal junction, posterior-central and cingulate gyri as well as in occipital lobes. Also, irrespective of thickness estimation metric, regions around the central sulci and parahippocampal gyri showed lowest reliability in CAT. As stated earlier, these regions also show the highest inter-individual variability within each pipeline (see Figure 2 and Supplementary Figure IX) and have the lowest between-pipeline correlations. This demonstrates that the cortical thickness estimates within these regions are less reliable and thus calls for particular caution in interpreting findings within these regions.

Summary and concluding remarks:

Using two large samples of younger and older healthy adults and three widely used packages for cortical thickness estimation, the current study addressed two fundamental factors that can potentially compromise the reproducibility of the neuroscientific findings: similarity of the

cortical thickness estimations between-pipelines, and within-pipeline test-retest reproducibility of cortical thickness estimations.

Our investigations demonstrated that: (1) estimates of the three tested packages, on average, result in similar spatial pattern of cortical thickness distribution, suggesting that the relative thickness estimates of the cortical areas are comparably depicted by these three pipelines; (2) actual measured thickness differs significantly between-pipelines, in most cortical areas, indicating that *in-vivo* thickness measurements are a proxy of actual thickness of the cortex, and actual values shall only be compared within the same software package and thickness estimation technique; (3) despite the different actual values estimated by different pipelines, in most cortical areas the inter-individual variability is depicted comparably by different tools, supporting general replicability of the findings of studies, assessing inter-individual variability in cortical thickness, using either of these pipelines; (4) furthermore, we observe a high level of within-tool test-retest reliability of cortical thickness in all three pipelines, demonstrating a general high level of reliability of the estimates within each pipeline; (5) yet, at the same time, we observed that in insula, cingulate gyrus, medial occipital lobe, parahippocampal gyrus and medial temporal pole, cortical thickness is estimated less reliably. Accordingly, these results imply regional variation in reliability and replicability of cortical thickness indices, and encourage attempts seeking to reproduce findings about variation in structural properties, particularly, within these less reliable regions.

Acknowledgements:

The authors thank Compute Canada (https://www.computecanada.ca) and Jülich Supercomputing Centre (https://www.fz-juelich.de/ias/jsc/) for the usage of the computing facilities in the development of this work. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/1-1, EI 816/4-1), the Helmholtz Portfolio Theme 'Supercomputing and Modelling for the Human Brain' and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) and Grant Agreement No. 785907 (HBP SGA2).

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Competing interests: The authors declare no competing interests.

References*:*

Alvarez I, Parker AJ, Bridge H. 2019. Normative cerebral cortical thickness for human visual areas. Neuroimage. 201:116057.

Amunts K, Armstrong E, Malikovic A, Homke L, Mohlberg H, Schleicher A, Zilles K. 2007. Gender-Specific Left-Right Asymmetries in Human Visual Cortex. J Neurosci. 27:1356–1364.

Bearden CE, Thompson PM. 2017. Emerging Global Initiatives in Neurogenetics: The Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA) Consortium. Neuron. 94:232–236.

Cardinale F, Chinnici G, Bramerio M, Mai R, Sartori I, Cossu M, Lo Russo G, Castana L, Colombo N, Caborni C, De Momi E, Ferrigno G. 2014. Validation of FreeSurfer-Estimated Brain Cortical Thickness: Comparison with Histologic Measurements. Neuroinformatics. 12:535–542.

Clarkson MJ, Cardoso MJ, Ridgway GR, Modat M, Leung KK, Rohrer JD, Fox NC, Ourselin S. 2011. A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage. 57:856–865.

Collins DL, Neelin P, Peters TM, Evans AC. 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr. 18:192–205.

Dahnke R, Yotter RA, Gaser C. 2013. Cortical thickness and central surface estimation. Neuroimage. 65:336–348.

Das SR, Avants BB, Grossman M, Gee JC. 2009. Registration based cortical thickness measurement. Neuroimage. 45:867–879.

Dickie E, Hodge S, Craddock R, Poline J-B, Kennedy D. 2017. Tools Matter: Comparison of Two Surface Analysis Tools Applied to the ABIDE Dataset. Res Ideas Outcomes. 3:e13726.

Fischl B. 2012. FreeSurfer. Neuroimage. 62:774–781.

Fischl B, Dale AM. 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci. 97:11050–11055.

Fischl B, Sereno MI, Tootell RBH, Dale AM. 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum Brain Mapp. 8:272–284.

Fjell AM, Amlien IK, Sneve MH, Grydeland H, Tamnes CK, Chaplin TA, Rosa MGP, Walhovd KB. 2015. The roots of Alzheimer's disease: Are high-expanding cortical areas preferentially targeted? Cereb Cortex. 25:2556–2565.

Fjell AM, Westlye LT, Grydeland H, Amlien I, Espeseth T, Reinvang I, Raz N, Dale AM, Walhovd KB. 2014. Accelerating cortical thinning: unique to dementia or universal in aging? Cereb Cortex. 24:919–934.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, Van Essen DC, Jenkinson M, WU-Minn HCP Consortium for the W-MH. 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 80:105–124.

Greve DN, Fischl B. 2018. False positive rates in surface-based anatomical analysis. Neuroimage. 171:6–14.

Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, van Os J, Marcelis M. 2012. The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. PLoS One. 7:e38234.

Han X, Pham DL, Tosun D, Rettmann ME, Xu C, Prince JL. 2004. CRUISE: Cortical reconstruction using implicit surface evolution. Neuroimage. 23:997–1012.

Kanai R, Rees G. 2011. The structural basis of inter-individual differences in human behaviour and cognition. Nat Rev Neurosci. 12:231–242.

Kim JS, Singh V, Lee JK, Lerch J, Ad-Dab'bagh Y, MacDonald D, Lee JM, Kim SI, Evans

AC. 2005. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. Neuroimage. 27:210–221.

Kochunov P, Lancaster JL, Glahn DC, Purdy D, Laird AR, Gao F, Fox P. 2006. Retrospective motion correction protocol for high-resolution anatomical MRI. Hum Brain Mapp. 27:957–962.

Lerch J. 2001. Measuring Cortical Thickness.

Lerch JP, Evans AC. 2005. Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage. 24:163–173.

Lewis L, Lepage C, Khalili-Mahani N, Omidyeganeh M, Jeon S, Bermudez P, Zijdenbos A, Vincent R, Adalat R, Evans A. 2017. Robustness and reliability of cortical surface reconstruction in CIVET and FreeSurfer. In: Annual Meeting of the Organization for Human Brain Mapping.

Lewis LB, Lepage CY, Evans AC. 2018. Utilizing the BigBrain as ground truth for evaluating of CIVET and FreeSurfer structural MRI pipelines. In: Annual Meeting of the Organization for Human Brain Mapping.

Lewis LB, Lepage CY, Glasser MF, Coalson TS, Essen DC Van, Alan C. Evans. 2019. An extended MSM surface registration pipeline to bridge atlases across the MNI and the FS/HCP worlds. In: Annual Meeting of the Organization for Human Brain Mapping.

Li Q, Pardoe H, Lichter R, Werden E, Raffelt A, Cumming T, Brodtmann A. 2015. Cortical thickness estimation in longitudinal stroke studies: A comparison of 3 measurement methods. NeuroImage Clin. 8:526–535.

Loken E, Gelman A. 2017. Measurement error and the replication crisis. Science (80-).

MacDonald D, Kabani N, Evans AC, Avis D. 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage. 12:340–356.

Martínez K, Madsen SK, Joshi AA, Joshi SH, Román FJ, Villalon-Reina J, Burgaleta M,

Karama S, Janssen J, Marinetto E, Desco M, Thompson PM, Colom R. 2015. Reproducibility of brain-cognition relationships using three cortical surface-based protocols: An exhaustive analysis based on cortical thickness. Hum Brain Mapp. 36:3227–3245.

McKay DR, Knowles EE, Winkler AA, Sprooten E, Kochunov P, Olvera RL, Curran JE, Kent Jr. JW, Carless MA, Goring HH, Dyer TD, Duggirala R, Almasy L, Fox PT, Blangero J, Glahn DC. 2014. Influence of age, sex and genetic factors on the human brain. Brain Imaging Behav. 8:143–152.

Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, Bartsch AJ, Jbabdi S, Sotiropoulos SN, Andersson JLR, Griffanti L, Douaud G, Okell TW, Weale P, Dragonu I, Garratt S, Hudson S, Collins R, Jenkinson M, Matthews PM, Smith SM. 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat Neurosci. 19:1523–1536.

Muhlert N, Ridgway GR. 2016. Failed replications, contributing factors and careful interpretations: Commentary on Boekel et al., 2015.

Natu VS, Gomez J, Barnett M, Jeska B, Kirilina E, Jaeger C, Zhen Z, Cox S, Weiner KS, Weiskopf N, Grill-Spector K. 2019. Apparent thinning of visual cortex during childhood is associated with myelination, not pruning. PANS.

Rajapakse JC, Giedd JN, Rapoport JL. 1997. Statistical approach to segmentation of single-channel cerebral mr images. IEEE Trans Med Imaging. 16:176–186.

Righart R, Schmidt P, Dahnke R, Biberacher V, Beer A, Buck D, Hemmer B, Kirschke JS, Zimmer C, Gaser C, Mühlau M. 2017. Volume versus surface-based cortical thickness measurements: A comparative study with healthy controls and multiple sclerosis patients. PLoS One. 12:e0179590.

Rowley CD, Bazin PL, Tardif CL, Sehmbi M, Hashim E, Zaharieva N, Minuzzi L, Frey BN, Bock NA. 2015. Assessing intracortical myelin in the living human brain using

myelinated cortical thickness. Front Neurosci. 9:396.

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT. 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex. 28:3095–3114.

Scholtens LH, de Reus MA, de Lange SC, Schmidt R, van den Heuvel MP. 2016. An MRI Von Economo – Koskinas atlas. Neuroimage. 170:249–256.

Scholtens LH, de Reus MA, van den Heuvel MP. 2015. Linking contemporary high resolution magnetic resonance imaging to the von economo legacy: A study on the comparison of MRI cortical thickness and histological measurements of cortical structure. Hum Brain Mapp. 36:3038–3046.

Seiger R, Ganger S, Kranz GS, Hahn A, Lanzenberger R. 2018. Cortical Thickness Estimations of FreeSurfer and the CAT12 Toolbox in Patients with Alzheimer's Disease and Healthy Controls. J Neuroimaging. 28:515–523.

Sled JG, Zijdenbos AP, Evans AC. 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 17:87–97.

Smith SM. 2002. Fast robust automated brain extraction. Hum Brain Mapp. 17:143–155.

Smith SM, Nichols TE. 2018. Statistical Challenges in "Big Data" Human Neuroimaging. Neuron. 97:263–268.

Tohka J, Zijdenbos A, Evans A. 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage. 23:84–97.

Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, Kandel BM, van Strien N, Stone JR, Gee JC, Avants BB. 2014. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. Neuroimage. 99:166–179.

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. 2013. The WU-Minn Human Connectome Project: An overview. Neuroimage. 80:62–79.

von Economo C, Koskinas G. 1925. Die Cytoarchitektonik der Hirnrinde des erwachsenen

Menschen. Berlin: Julius Springer.

Wagstyl K, Lepage C, Bludau S, Zilles K, Fletcher PC, Amunts K, Evans AC. 2018. Mapping cortical laminar structure in the 3D bigbrain. Cereb Cortex. 28:2551–2562.

Walhovd KB, Fjell AM, Giedd J, Dale AM, Brown TT. 2017. Through Thick and Thin: a Need to Reconcile Contradictory Results on Trajectories in Human Cortical Development. Cereb Cortex. 27:1472–1481.

Yotter RA, Thompson PM, Gaser C. 2011. Algorithms to Improve the Reparameterization of Spherical Mappings of Brain Surface Meshes. J Neuroimaging. 21:e134–e147.

Zijdenbos A, Forghani R, Evans A. 1998. Automatic quantification of MS lesions in 3D mri brain data sets: Validation of INSECT. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg. p. 439–448.

Tables:

Table 1: demographic characteristics of samples under study:

	n (% female)	Age (mean ± SD)
HCP Sample	981 (55%)	28.8 ± 3.6 [22-37]
GOBS Sample	832 (60%)	42.7 ± 14.6 [18-85]
Test-retest Sample	115 (70%)	49.3 ± 11.44 [27-77]

Table2. Distribution of global cortical thickness, across all participants of each sample (mean ± SD).

	HCP	GOBS
Fs 6.0	2.66 ± 0.077 [2.4 – 2.90]	2.56 ± 0.1 [2.2 – 2.87]
CAT	2.82 ± 0.09 [2.53 – 3.17]	2.64 ± 0.12 [2.18 – 3.09]
CIVET	2.81 ± 0.08 [2.54 – 3.07]	2.72 ± 0.09 [2.45 – 3.02]

Table3. Test retest results MAPE

	Mean ± SD	min	25%	50%	75%	max
Fs 6.0	2.488 ± 0.899	1.4	2.02	2.28	2.68	11.32
CAT	2.18 ± 1.1	0.945	1.51	1.85	2.51	9.11
CIVET	2.026 ± 1.09	1.07	1.54	1.77	2.14	14.9

Figure legends:

Figure 1 Distribution of global cortical thickness across samples and pipelines. FS: FreeSurfer.

Figure 2 regional mean (left columns) and SD (right columns) of cortical thickness estimates of each pipeline. For each pipeline, parcels with the highest SD (top 10%) are depicted with a black surrounding. FS: FreeSurfer.

Figure 3 Regional between-pipeline comparison of cortical thickness estimates within the HCP sample. Top rows: paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Only regions with significant paired t-test comparisons are colored ($p < 4 \times 10^{-5}$); Lower rows: Between-pipeline regional Spearman's correlation). FS: FreeSurfer.

Figure 4 Regional between-pipeline comparison of cortical thickness estimates within the GOBS sample. Top rows: paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Only regions with significant paired t-test comparisons are colored ($p < 4 \times 10^{-5}$); Lower rows: Between-pipeline regional Spearman's correlation). FS: FreeSurfer.

Figure 5 Between-pipeline comparison of global cortical thickness and its inter-individual variability. Scatter plots of pair-wise comparison of global cortical estimates and their Spearman's correlations, within two cohorts (A, C; dashed line depicts the identity line(y=x)). B, D: Scatter plots of association between participants' age and global cortical thickness from each pipeline, for each cohort. Legends depict Spearman's correlation between age and global thickness of each pipeline. FS: FreeSurfer.

Figure 6 Test-retest reliability. Left: pairs of violin plots demonstrate distribution of global cortical thickness estimates from odd and even scans, over all the subjects. Right: regional

distribution of mean absolute percent error (MAPE) for each pipeline. Ten percent of the parcels, showing the lowest test-retest reliability, are surrounded with black line. The lighter colors in the spatial maps depict higher MAPE. FS: FreeSurfer.

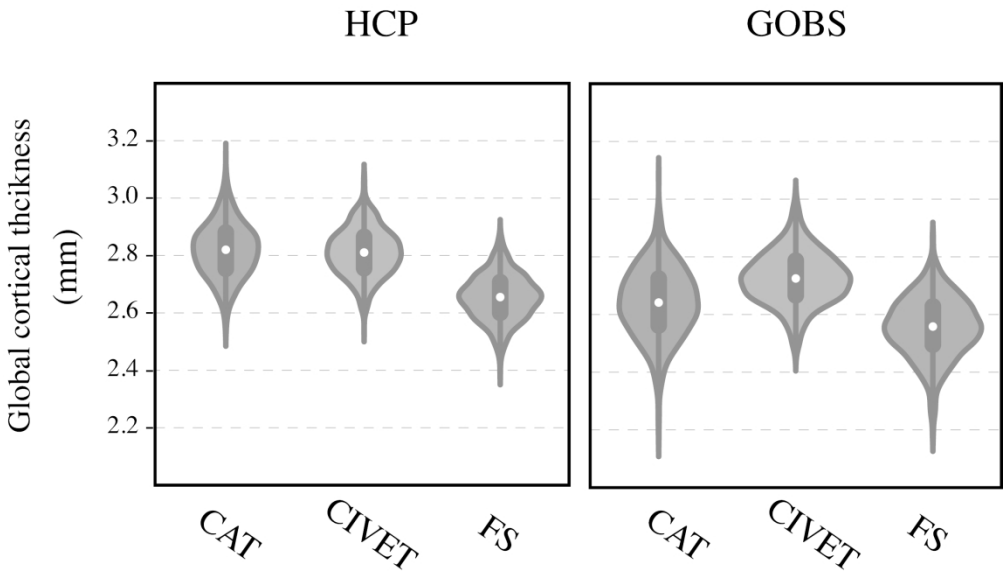


Figure 1 Distribution of global cortical thickness across samples and pipelines. FS: FreeSurfer.

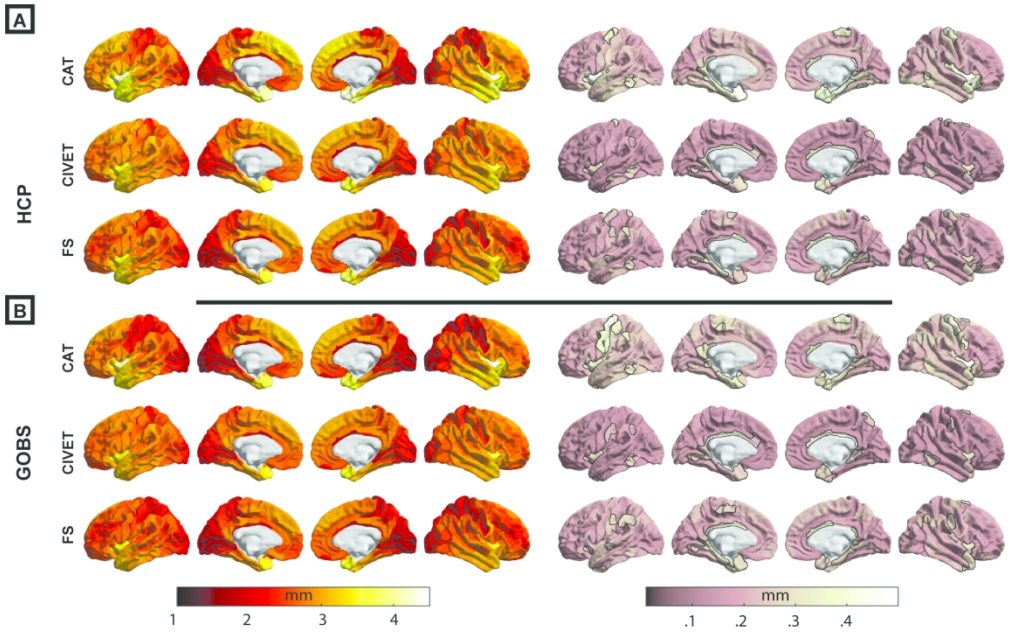


Figure 2 regional mean (left columns) and SD (right columns) of cortical thickness estimates of each pipeline. For each pipeline, parcels with the highest SD (top 10%) are depicted with a black surrounding. FS: FreeSurfer.

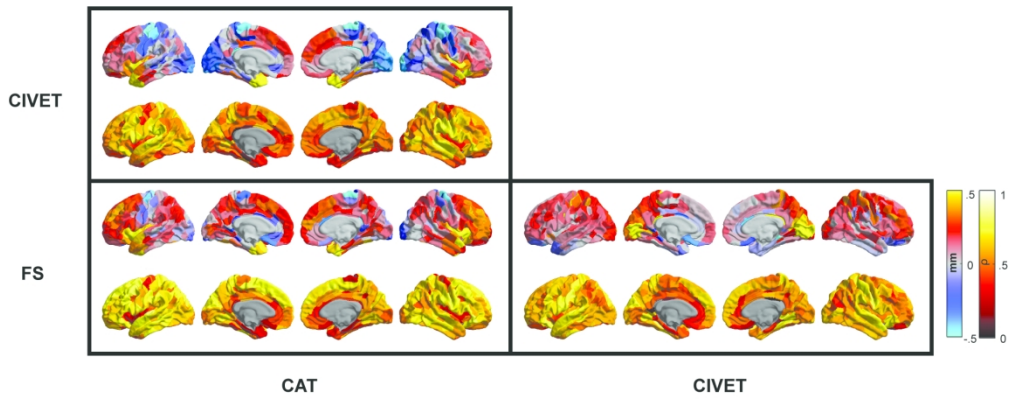


Figure 3 Regional between-pipeline comparison of cortical thickness estimates within the HCP sample. Top rows: paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Only regions with significant paired t-test comparisons are colored ($p < 4 \times 10\text{-}5$); Lower rows: Between-pipeline regional Spearman's correlation). FS: FreeSurfer.

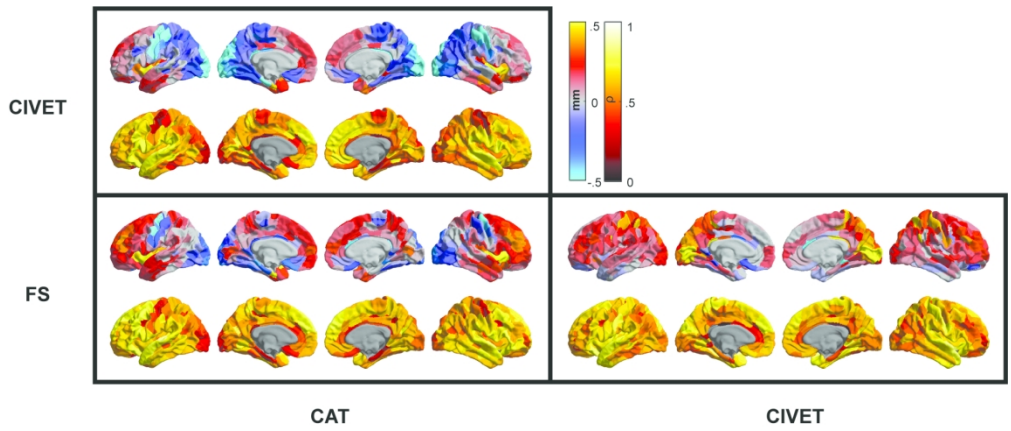


Figure 4 Regional between-pipeline comparison of cortical thickness estimates within the GOBS sample. Top rows: paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Only regions with significant paired t-test comparisons are colored ($p < 4 \times 10\text{-}5$); Lower rows: Between-pipeline regional Spearman's correlation). FS: FreeSurfer.

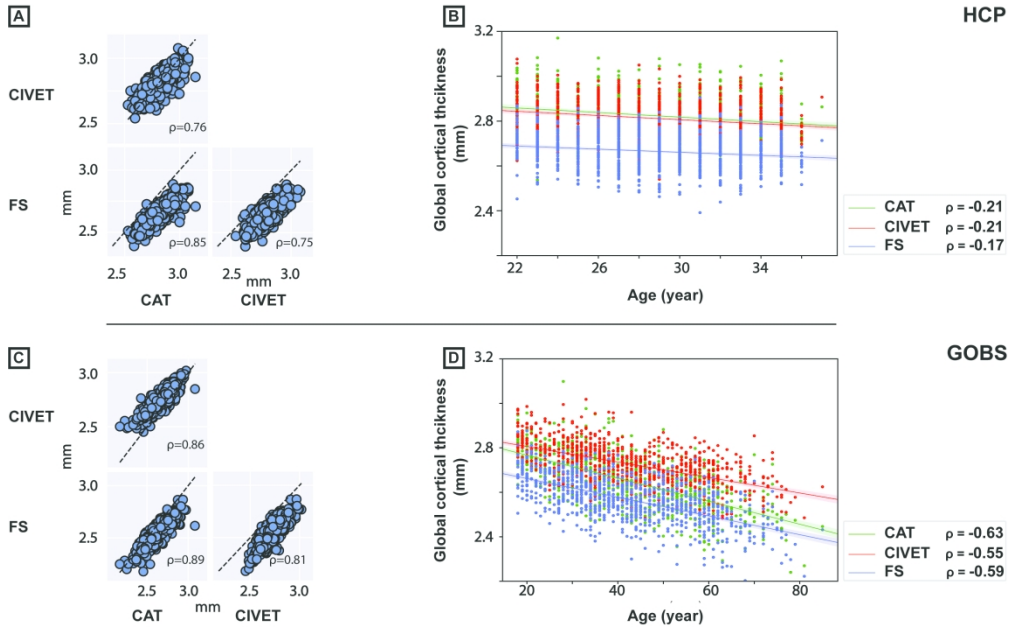
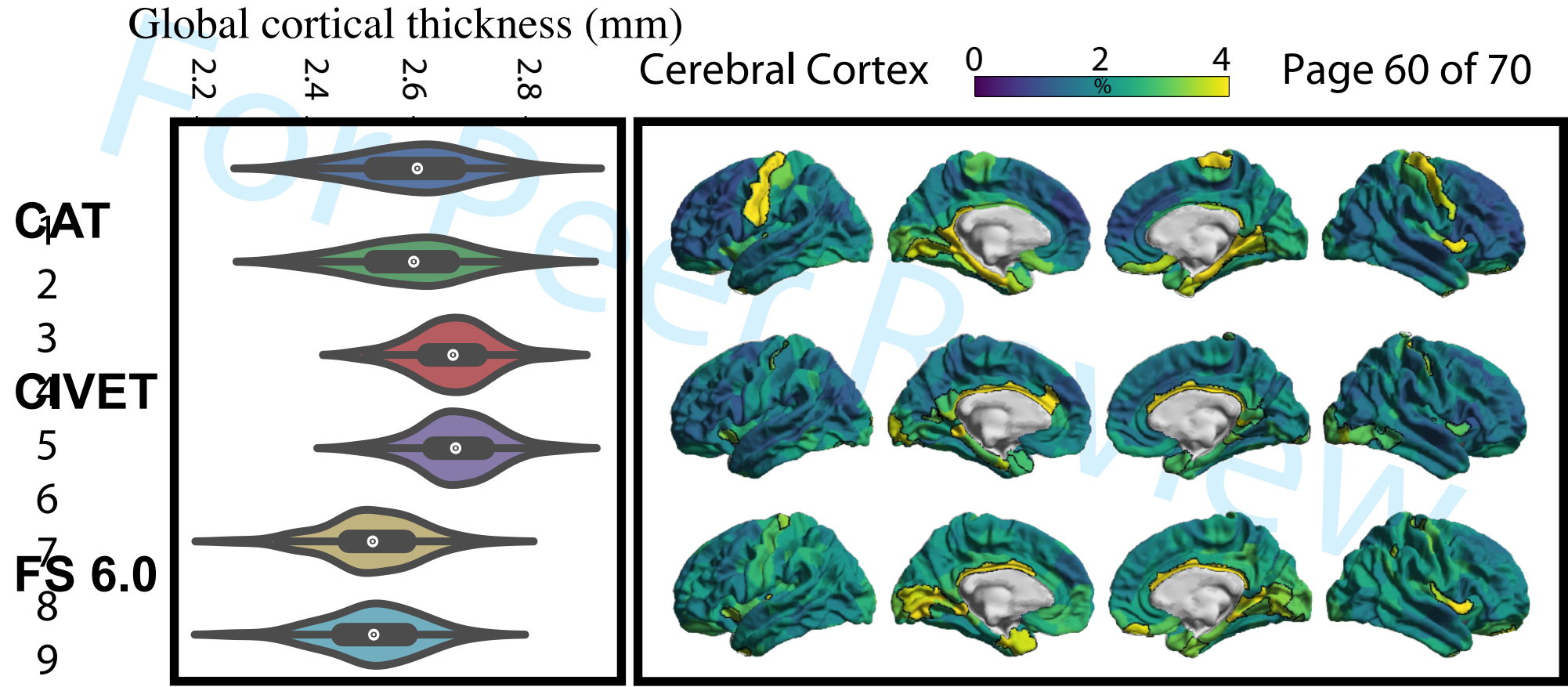
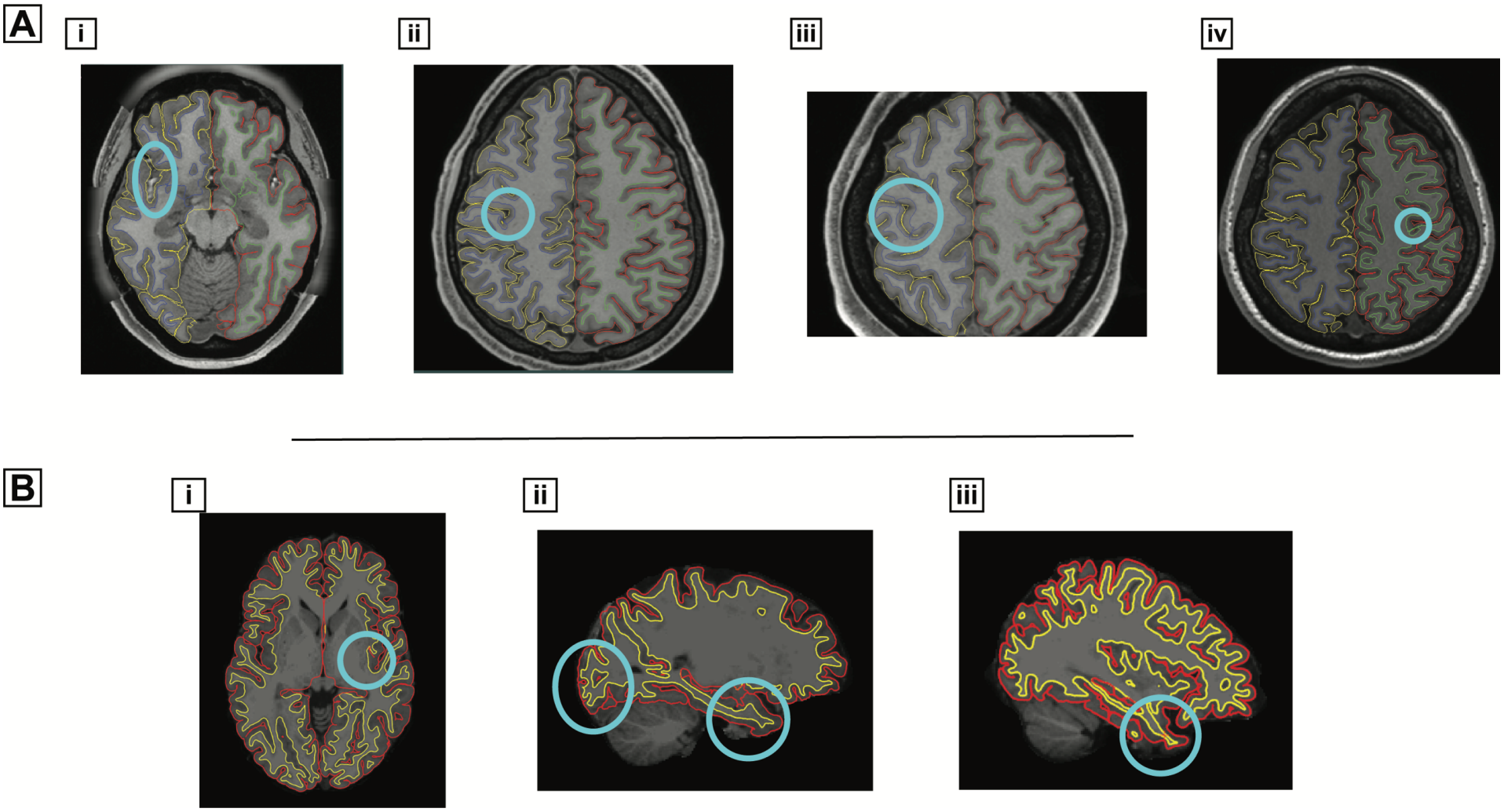
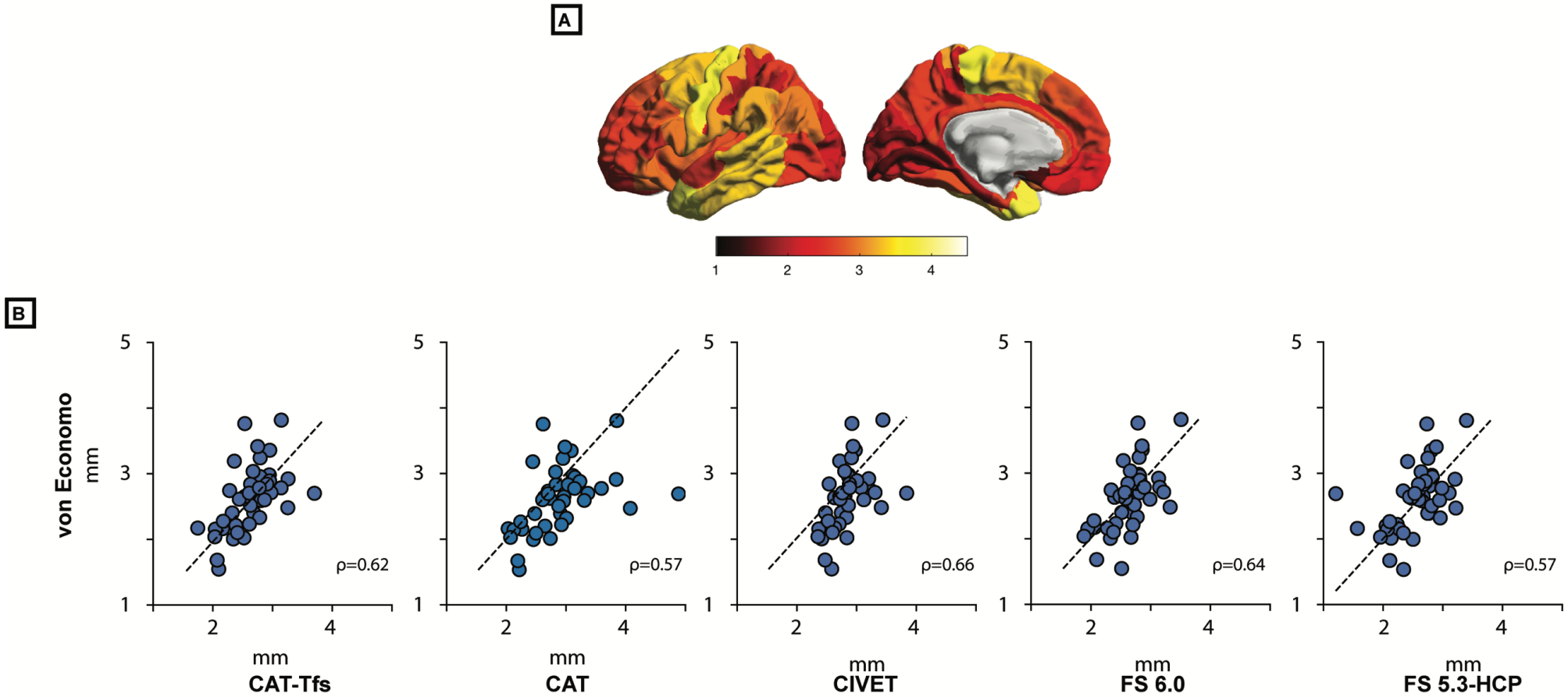


Figure 5 Between-pipeline comparison of global cortical thickness and its inter-individual variability. Scatter plots of pair-wise comparison of global cortical estimates and their Spearman's correlations, within two cohorts (A, C; dashed line depicts the identity line(y=x)). B, D: Scatter plots of association between participants' age and global cortical thickness from each pipeline, for each cohort. Legends depict Spearman's correlation between age and global thickness of each pipeline. FS: FreeSurfer.

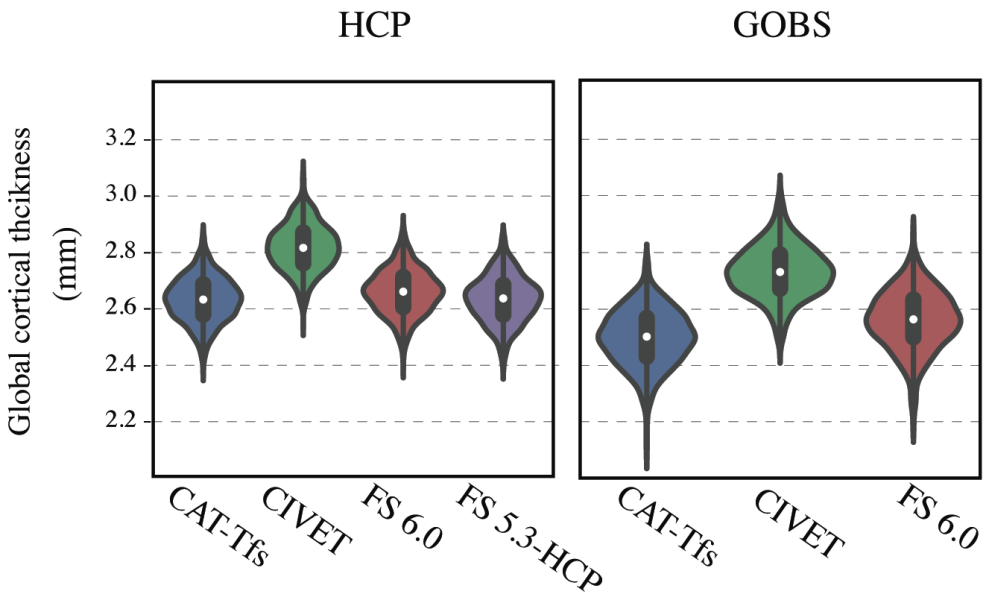




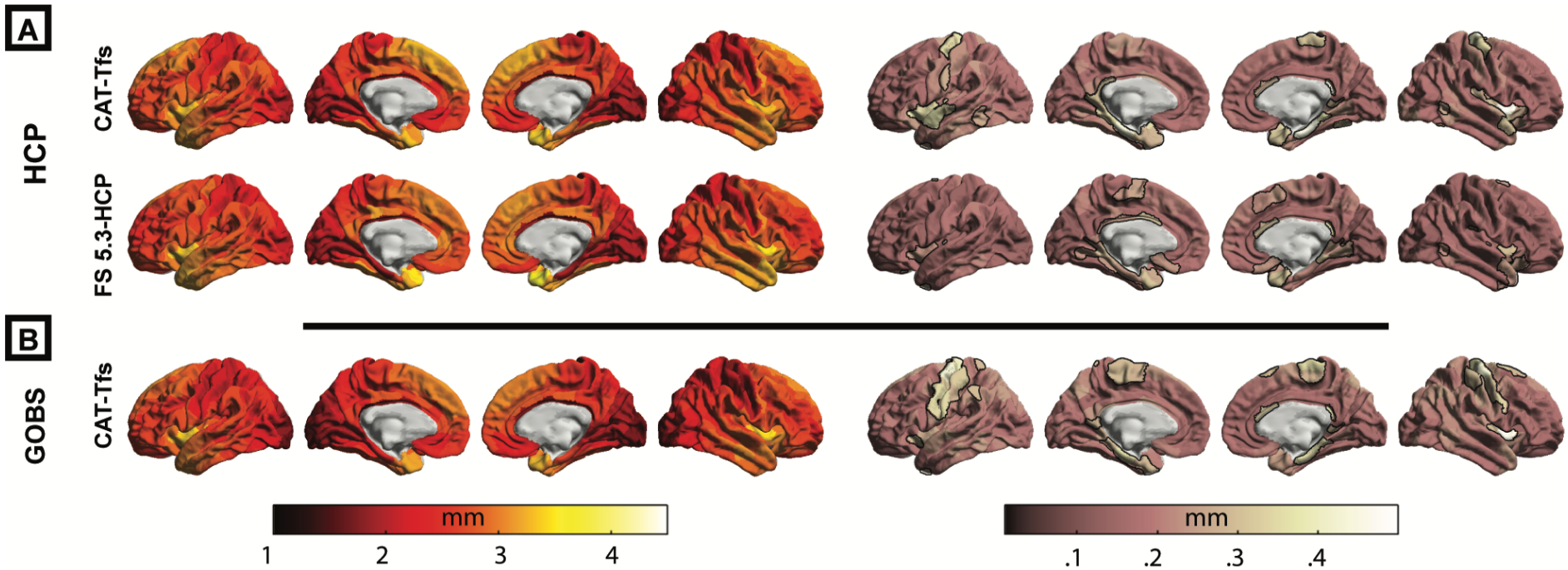
Supplementary Figure I Several examples of errors in the CIVET (A) and FreeSurfer (B) pipelines.



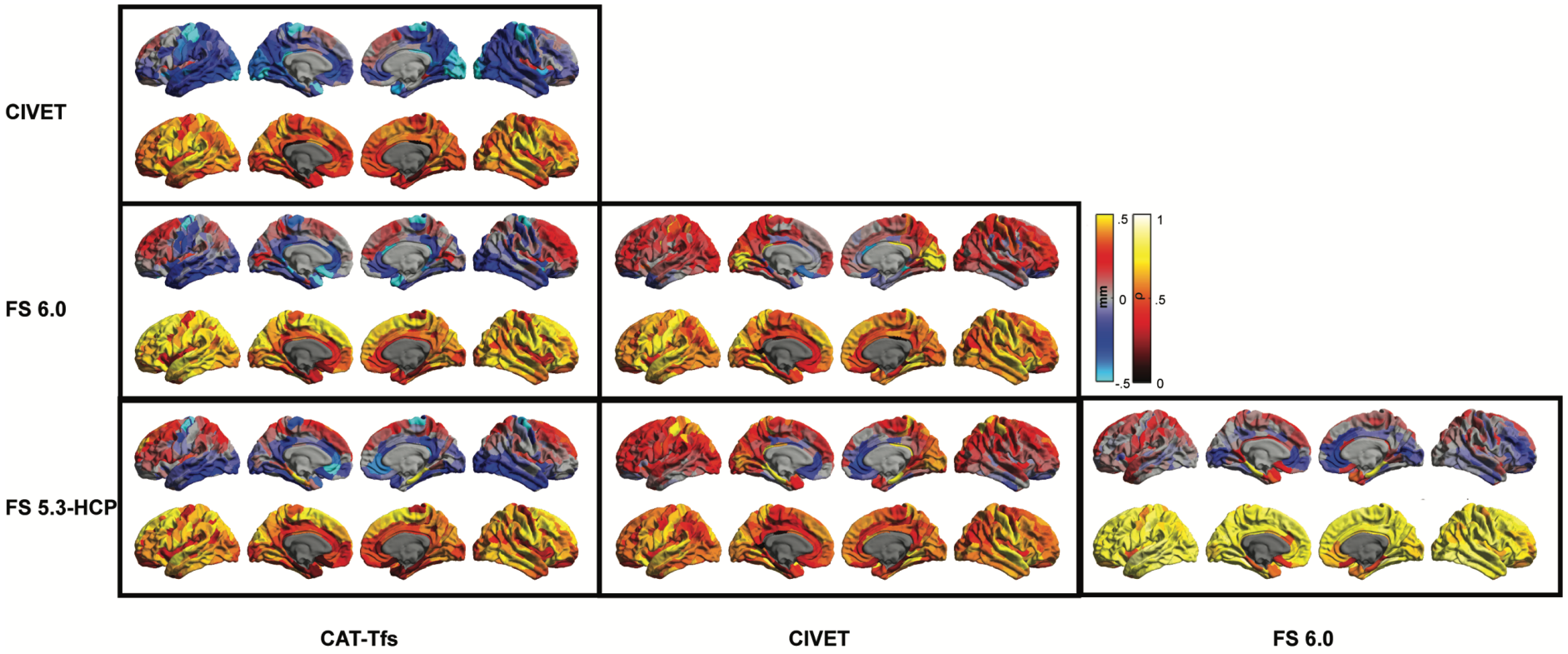
Supplementary Figure II Comparison with the manual measurements. Von Economo and Koskinas' reported cortical width values, within the boundaries of the manually segmented von Economo (1925) regions on fsaverage space (A); Scatter plots and Spearman's correlation between von Economo and Koskinas' reported cortical width values in all regions and thickness estimates from each automatic pipeline, within HCP cohort (B; dashed lines depicts the identity line(y=x)). CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer. For detailed information about the regions of the von Economo-Koskinas atlas, see http://www.dutchconnectomelab.nl/economo/.



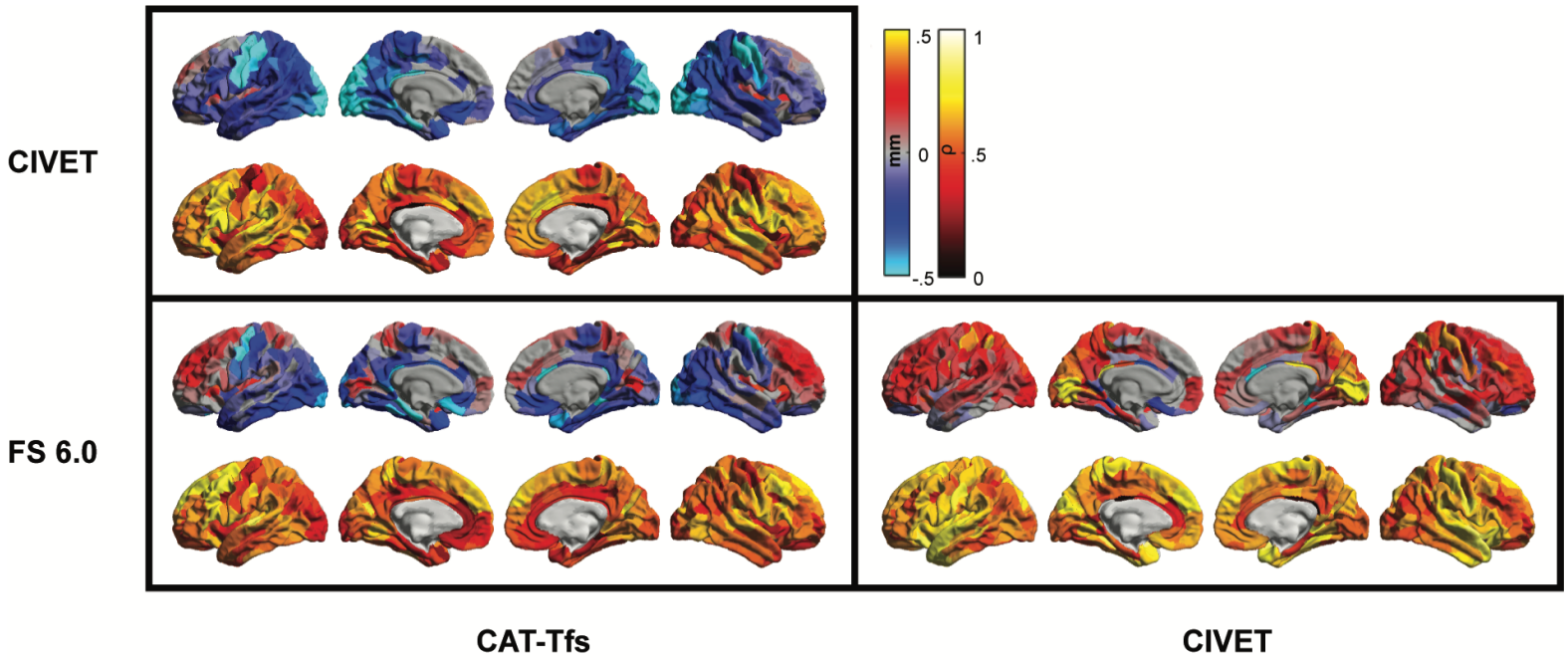
Supplementary Figure III Distribution of global cortical thickness across samples and pipeline. CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



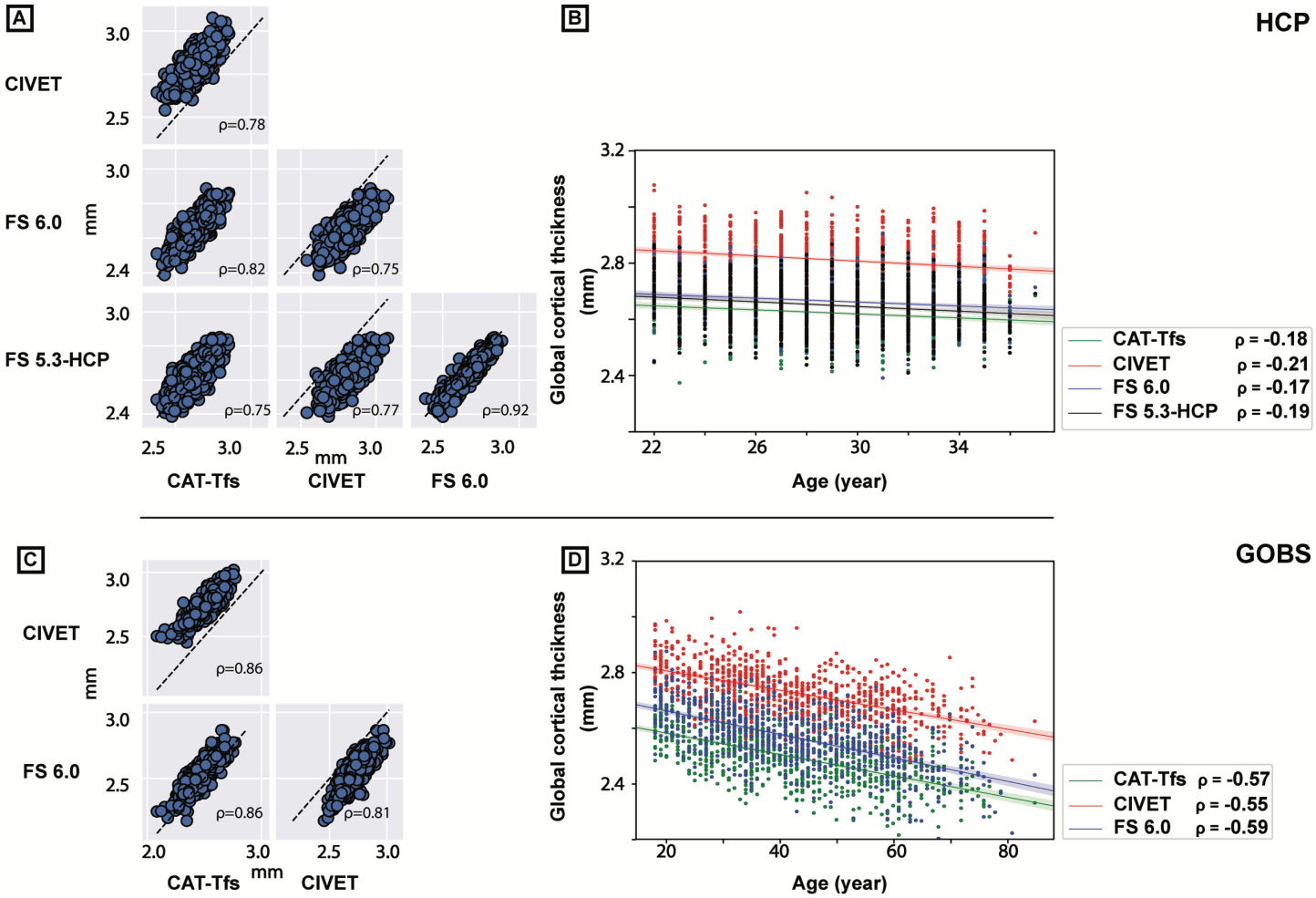
Supplementary Figure IV Regional mean (left columns) and SD (right columns) of cortical thickness estimates of CAT-Tfs and FreeSurfer v5.3-HCP (only for HCP sample) pipelines. For each pipeline, parcels with the highest SD (top 10%) are depicted with a black surrounding. CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



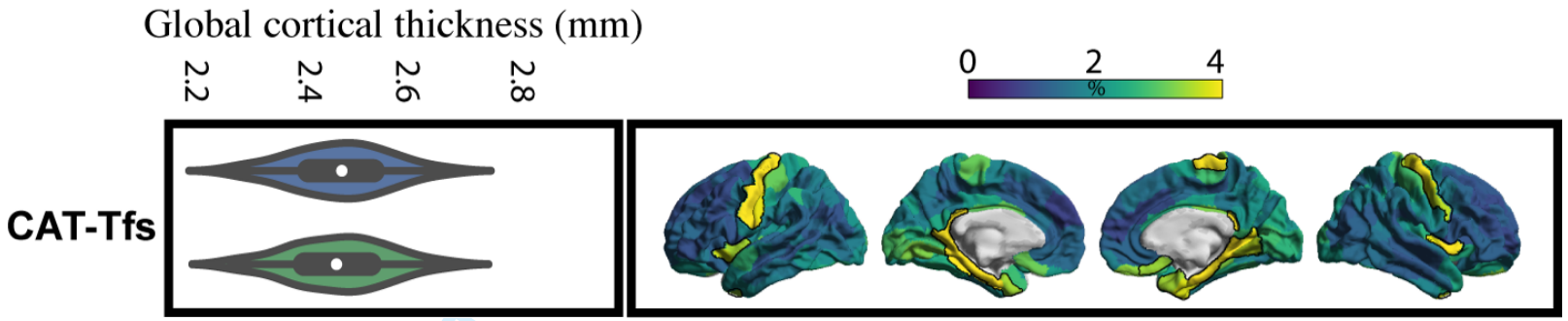
Supplementary Figure V Regional between-pipeline comparison of cortical thickness estimates within the HCP sample (modified CAT thickness estimates-Tfs). Top rows: Top rows: average paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Here only regions with significant paired t-test are shown ($p < 2 \times 10^{-5}$); Lower rows: Between-pipeline regional Spearman's correlation). CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



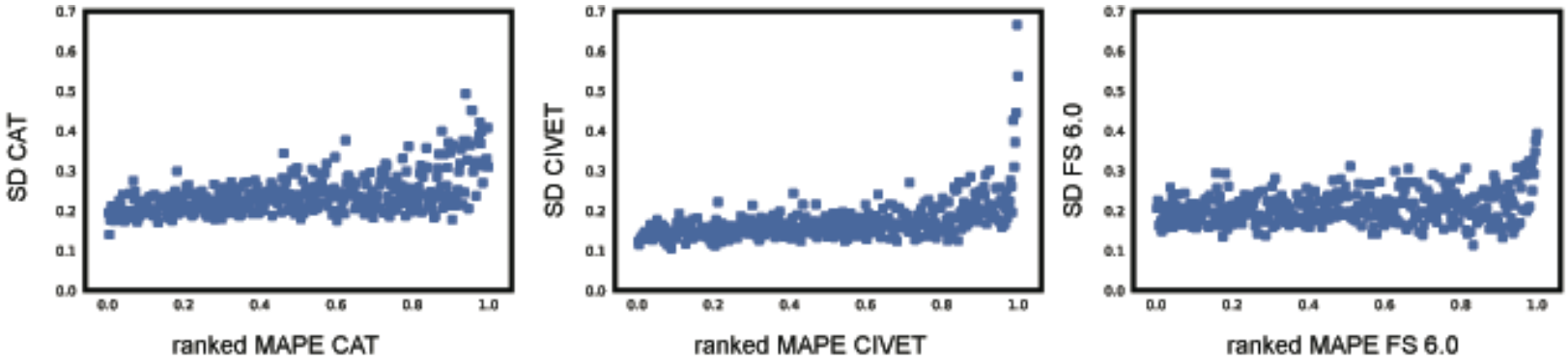
Supplementary Figure VI Regional between-pipeline comparison of cortical thickness estimates within the GOBS sample (modified CAT thickness estimates-Tfs). Top rows: average paired-difference map of cortical thickness estimates; Red-yellow colors depict higher thickness estimate of the pipeline mentioned in column vs row and the dark-light blue depicts the opposite direction. Here only regions with significant paired t-test are shown ($p < 4 \times 10^{-5}$); Lower rows: Between-pipeline regional Spearman's correlation). CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



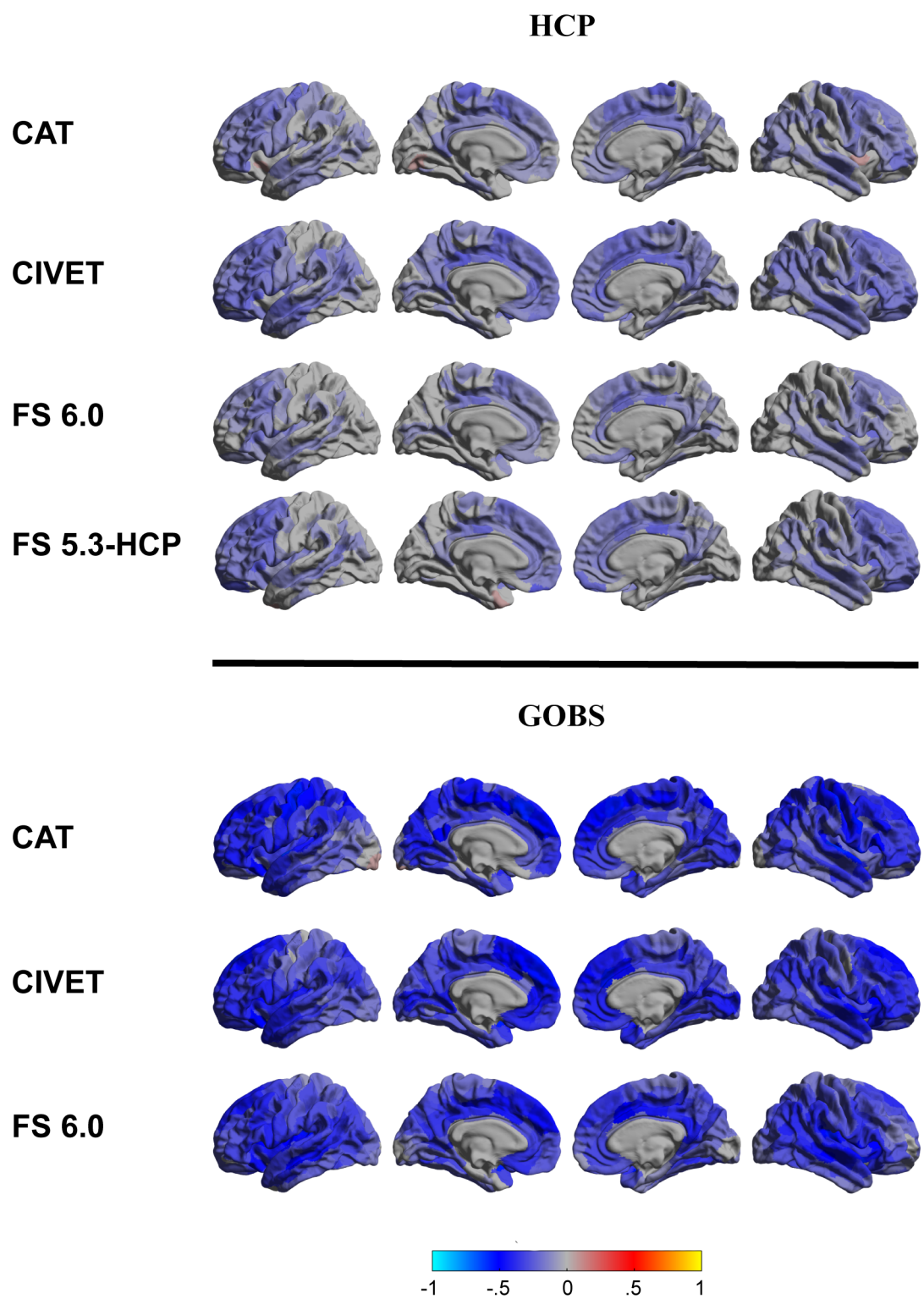
Supplementary Figure VII Between-pipeline comparison of global cortical thickness and its inter-individual variability. Scatter plots of pair-wise comparison of global cortical estimates and their Spearman's correlations, within two cohorts (A, C; dashed line depicts the identity line(y=x)). B, D: Scatter plots of association between participants' age and global cortical thickness from each pipeline, for each cohort. Legends depict Spearman's correlation between age and global thickness of each pipeline. CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



Supplementary Figure VIII Test-retest reliability. Left: pairs of violin plots demonstrate distribution of global cortical thickness estimates from odd and even scans, over all the subjects. Right: regional distribution of mean absolute percent error (MAPE) for each pipeline. Ten percent of the parcels, showing the lowest test-retest reliability, are surrounded with black line. The lighter colors in the spatial maps depict higher MAPE. CAT-Tfs: modified CAT thickness estimates-Tfs; FS: FreeSurfer.



Supplementary Figure IX Relationship between standard deviation in each parcel and its rank in the MAPE distribution. In general, parcels with highest standard deviation (SD) across individuals, within the GOBS cohort, also demonstrated higher MAPE, i.e. more error in the test-retest experiment.



Supplementary Figure X Regional cortical thickness-age associations (color bar shows spearman's correlation).