Large-Scale Evaluation of ANTs and FreeSurfer Cortical Thickness Measurements

Nicholas J. Tustison[a,1], Philip A. Cook[b], Arno Klein[c], Gang Song[b], Sandhitsu R. Das[b], Jeffrey T. Duda[b], Benjamin M. Kandel[b], Niels van Strien[c], James R. Stone[a], James C. Gee[b], Brian B. Avants[b]

[a]*Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA*
[b]*Penn Image Computing and Science Laboratory, University of Pennsylvania, Philadelphia, PA*
[c]*Sage Bionetworks, Seattle, WA*
[d]*Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*

Abstract

Many studies of the human brain have explored the relationship between cortical thickness and cognition, phenotype, or disease. Due to the tedium, time, inconsistency, and subjectivity in manual measurement of cortical thickness, scientists have relied on robust software tools for automating these measurements which facilitate the testing and refinement of neuroscientific hypotheses. The most widely used tool is the surface-based FreeSurfer package which owes its popularity largely to its public availability and good performance. Equally important to the adoption of such tools is a demonstration of their reproducibility and validity, in particular, the availability of robust parameter sets which have demonstrated good performance. To this end, we have developed the automated, volume-based, Advanced Normalization Tools (ANTs) cortical thickness pipeline comprising well-vetted components such as SyGN (multivariate template construction), SyN (image registration), N4 (bias correction), Atropos (n-tissue segmentation), and DiReCT (cortical thickness estimation). In this work, we have conducted the largest evaluation of automated cortical thickness measures in publicly available data, comparing FreeSurfer and ANTs measures computed on approximately 1200 images from four open data sets (IXI, MMRR, NKI, and OASIS), using standard processing protocols including the recently proposed ?Desikan-Killiany-Tourville? (DKT) cortical labeling protocol. We found good scan-rescan repeatability with both FreeSurfer and ANTs measures. Given that such assessments of precision do not necessarily reflect accuracy or ability to make statistical inferences, we tested validity of these thickness measures by their ability to predict demographic-based measures with a well-substantiated relationship to cortical thickness: age and gender. ANTs is shown to have a higher predictive performance than FreeSurfer for both of these measures. In addition to our use of open image data sets, to further promote open science, scientific reproducibility, and the use of the proposed ANTs pipeline, we make all of our scripts and results publicly available.

Keywords: advanced normalization tools, age prediction, MRI, gender prediction, open science, scientific reproducibility

1. Introduction

Magnetic resonance imaging-based structural analysis of the human brain plays a fundamental role in identifying the relationship between cortical morphology, disease, and cognition. Discriminative cortical thickness values have been demonstrated in normal aging [57, 13, 51, 95] and in gender [2, 60]. Thickness is also sensitive to conditional abnormalities such as Huntington's disease [77, 76, 81], schizophrenia [69], bipolar disorder [63], Alzheimer's disease and frontotemporal dementia [25, 23], Parkinson's disease [47], Williams syndrome [91], multiple sclerosis [74], autism [15, 41], migraines [20], chronic smoking [53], alcoholism [32], cocaine addiction [66], Tourette syndrome in children [89], scoliosis in female adolescents [96], early-onset blindness [44], chronic pancreatitis [35], obsessive-compulsive disorder [84], ADHD [1], obesity [73], and heritable [71] and elderly [8] depression. Evidence of cortical thickness variation has also been found in untreated male-to-female transsexuality [61], handedness [59, 2], intelligence [83], athletic ability [99], meditative practices [56], musical ability [9, 33], tendency toward criminality [72], childhood sexual abuse in adult females [43], and Tetris-playing ability in female adolescents [38]. Additionally, recent studies demonstrate correlated anatomical relationships using cortical thickness measures [102, 58, 42, 14]. Although these findings are subject to debate and interpretation [36], the availability of quantitative computational methods for extracting a measure of cortical thickness has proven invaluable for developing and refining fundamental neuroscience hypotheses.

Computational methods for analyzing the cortex may be broadly characterized as surface mesh-based or volumetric [79, 16]. Representative of the former is the FreeSurfer[2] cortical modeling software package [18, 30, 28, 29, 31] which owes its popularity to public availability, excellent documentation, good performance, and integration with other toolkits, such as the extensive FMRIB software library (FSL) [88]. Similar

[1]Corresponding author: PO Box 801339, Charlottesville, VA 22908; T: 434-924-7730; email address: ntustison@virginia.edu.

Partial support provided by US Army Medical Research and Materiel Command; Contract grant number: W81XWH-09-2-0055.

[2]http://surfer.nmr.mgh.harvard.edu/

to other surface-based cortical thickness estimation approaches (e.g., [21, 65, 64, 48]), the outer cortical and gray/white matter surfaces from individual subject MR data are modeled with polygonal meshes which are then used to determine local cortical thickness values based on a specified correspondence between the surface models.

Image volumetric (or meshless) techniques vary both in their algorithms as well as in the underlying definitions of cortical thickness. An early, foundational technique is the method of [45] in which the inner and outer surface geometry is used to determine the solution to Laplace's equation where thickness is measured by integrating along the tangents of the resulting field lines spanning the boundary surfaces. Subsequent contributions improved upon the original formulation. For example, in [104], a Eulerian partial differential equation approach was proposed to facilitate the computation of correspondence paths. Extending the surface-based work of [64], the hybrid approach of [48] uses the discrete Laplacian field to deform the white matter surface mesh towards the outer cortical surface. Although the Laplacian-based approach has several advantages including generally lower computation times and non-crossing correspondence paths, direct correlative assessments with histology are potentially problematic as the quantified distances are not necessarily Euclidean. Other volumetric algorithms employ coupled level sets [105], model-free intelligent search strategies either normal to the gray-white matter interface [79], or using a min-max rule [94]. Most relevant to this work is the DiReCT (Diffeomorphic Registration-based Cortical Thickness) algorithm proposed in [19] where generated diffeomorphic mappings between the gray/white matter and exterior cortical surfaces are used to propagate thickness values through the cortical gray matter. A unique benefit of DiReCT is that it naturally estimates the boundaries of buried sulci by employing a diffeomorphic constraint on the probabilistic estimate of the gray matter and cerebrospinal fluid interface.

Although a variety of techniques exist for estimating cortical thickness from imaging data (of which only a fraction are cited here), several common preprocessing components can be identified. The most fundamental of these include inhomogeneity correction, skull stripping, and *n*-tissue segmentation for differentiating gray and white matter. For statistical analysis across large populations, construction of population-specific unbiased templates is also potentially beneficial [27]. In addition, intermediate steps might include a crucial registration component (e.g., propagating template-based tissue priors for improved segmentation).

Cortical thickness studies are made more complex by the need for large neuroimaging data sets such as that provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) [100] and the need for packaging of state-of-the-art research methods so that other researchers can more easily use them. Currently, the National Institutes of Health (NIH) mandates that any NIH-funded data resources, including MRI, must be released to the public. In contrast to ADNI, which provides standardized data acquisition protocols used across all sites, these smaller-scale projects are collected in an unstructured way. Therefore, neuroimage processing tools must work reliably even when there is a relative lack of quality control over the input data. While robustness is a goal shared by all software development targeted at neuroscience research, very few methods have been thoroughly tested on large and unstructured neuroimaging data sets.

The general lack of availability of published algorithms [52] (not to mention critical preprocessing components) is a strong deterrent to the use or evaluation of these algorithms by external researchers. For example, one recent evaluation study [16] compared FreeSurfer (a surface-based method) with two volumetric methods [45, 19]. Whereas the entire FreeSurfer processing pipeline has been made publicly available, refined by the original authors and other contributors, and described in great detail (specifically in terms of suggested parameters), both volumetric methods were implemented and run by the authors of the evaluation (not by the algorithm developers) using unspecified parameters with relatively small, private data sets, making the comparisons less than ideal (see [93] for further discussion concerning the issue of instrumentation bias and scientific reproducibility in the use and evaluation of software). Further complicating such comparisons is the potential for bias, such as interpolation artifacts when converting surface to volume data or vice versa [49].

We provide below a brief description of our proposed pipeline, which produces a volumetric cortical thickness map from an individual subject's T1-weighted MRI. Additionally, we note that it is freely available as part of the Advanced Normalization Tools (ANTs) software package. This includes all the necessary preprocessing steps consisting of well-vetted, previously published algorithms for bias correction [92], brain extraction [4], *n*-tissue segmentation [6], template construction [7], and image normalization [5]. More importantly, we provide explicit coordination among these components within a set of well-documented shell scripts[3] which are also available in the ANTs repository where parameters have been tuned by ANTs developers (N.T. and B.A.).

Here we demonstrate the use of the described framework in processing 1205 publicly available, T1-weighted brain MR images drawn from four well-known data sets. For comparative evaluation we also process the same data using the standard FreeSurfer cortical thickness processing protocol. Similar to previous work [e.g., 16], we are able to report repeatability assessments for both frameworks using subsets of the data with repeated acquisitions.

However, repeatability (or, more generally, *precision*) is not conceptually equivalent to *accuracy* and, thus, does not provide a complete perspective for determination of measurement quality. Although FreeSurfer validation has included histological [77] and image-drawn [54] comparisons, such manual assessments were extremely limited in terms of number of subjects and the number of cortical regions. In addition, there was no mention in these studies of the number of human observers making these measurements nor discussion of quality assurance. Alternatively, without ground truth other forms of

[3] Ongoing improvements in documentation can be found in the form of online tutorials, self-contained github examples, and the ANTs discussion forums.

evidence can be adduced [e.g., 10] in making comparative inferences. In this work we use demographic-based predictive assessments to show that ANTs outperforms FreeSurfer-based thickness estimation for these data, based on well-studied relationships between cortical thickness and age/gender.

Finally, we make available all data from both ANTs and FreeSurfer processing outcomes. This includes derived image data, processing scripts, and tabulated results. The availability of both the code and data enables the set of results described in this work to be fully reproducible.

2. Methods and Materials

2.1. ANTs volume-based cortical thickness estimation pipeline

The ANTs cortical thickness estimation workflow is illustrated in Figure 1. The steps are as follows:

1. Initial N4 bias correction on input anatomical MRI
2. Brain extraction using a hybrid segmentation/template-based strategy
3. Alternation between prior-based segmentation and "pure tissue" posterior probability weighted bias correction using Atropos and N4
4. DiReCT-based cortical thickness estimation
5. Optional normalization to specified template and multi-atlas cortical parcellation

Each component, including both software and data, is briefly detailed below with the relevant references for additional information.

The coordination of all the algorithmic components is encapsulated in the shell script `antsCorticalThickness.sh` with subcomponents delegated to `antsBrainExtraction.sh` and `antsAtroposN4.sh`. A representative script command is reproduced in Listing 1 for a single IXI subject to demonstrate the simplicity and mature status of what we propose in this work and a comparison with the analogous FreeSurfer command. Option descriptions are provided by invoking the help option, i.e., "`antsCorticalThickness.sh -h`".

```
# Processing calls for subject IXI002-Guys-0828-T1

# ANTs
antsCorticalThickness.sh \
  -a IXI/T1/IXI002-Guys-0828-T1.nii.gz \
  -e IXI/template/T_template0.nii.gz \
  -m IXI/template/T_template0ProbabilityMask.nii.gz \
  -f IXI/template/T_template0ExtractionMask.nii.gz \
  -p IXI/template/Priors/priors%d.nii.gz \
  -o IXI/ANTsResults/IXI002-Guys-0828-

# FreeSurfer
recon-all \
  -i  IXI/T1/IXI002-Guys-0828-T1.nii.gz \
  -s  IXI002-Guys-0828 \
  -sd IXI/FreeSurferResults/ \
  -all
```

Listing 1: Analogous ANTs and FreeSurfer command line calls for a single IXI subject in the evaluation study.

2.1.1. N4 bias field correction

Critical to quantitative processing of MRI is the minimization of field inhomogeneity effects which produce artificial low frequency intensity variation across the image. Large-scale studies, such as ADNI, employ perhaps the most widely used bias correction algorithm, N3 [86], as part of their standard protocol [11].

In [92] we introduced an improvement of N3, denoted as "N4", which demonstrates a significant increase in performance and convergence behavior on a variety of data. This improvement is a result of an enhanced fitting routine (which includes multi-resolution capabilities) and a modified optimization formulation. For our workflow, the additional possibility of specifying a weighted mask in N4 permits the use of a "pure tissue" probability map (described below) calculated during the segmentation pipeline for further improvement of bias field estimation.

N4 is used in two places during the individual subject processing (cf Figure 1). It is used to generate an initial bias-corrected image for use in brain extraction. The input mask is created by adaptively thresholding the background from the foreground using Otsu's algorithm [70]. Following brain extraction, six-tissue (cerebrospinal fluid, cortical gray matter, white matter, deep gray matter, brain stem, and cerebellum) segmentation involves iterating between bias field correction using the current pure tissue probability map as a weight mask and then using that bias-corrected image as input for the Atropos segmentation step (described below).

2.1.2. Brain extraction

Brain extraction using ANTs combines template building, high-performance brain image registration, and Atropos segmentation with topological refinements. An optimal template [7], i.e., a mean shape and intensity image representation of a particular cohort, is first constructed using structural MRI data. Template construction iterates between estimating the optimal template and registering each subject to the optimal template. In this work, we perform the additional step of building separate templates for each cohort and propagating the probabilistic mask to each cohort template using registration of the T1-weighted templates (cf Figure 3). A probabilistic brain extraction mask for the new template can then be generated by warping an existing mask to the template or by averaging the warped, whole brain labels of subjects registered to the new template, if such labels are available. Further refinements include thresholding the warped brain probability map at 0.5 and dilating the resulting mask with a radius of two. Atropos is used to generate an initial three-tissue segmentation estimate within the mask region. Each of the three tissue labels undergoes separate morphological operations including hole-filling and erosion. These results are then combined to create the brain extraction mask which is further refined by additional dilation, erosion, and hole-filling operations. The complete workflow is found in the script `antsBrainExtraction.sh`.[4]

[4] A self-contained 2-D example is available at https://github.com/ntustison/antsBrainExtractionExample.

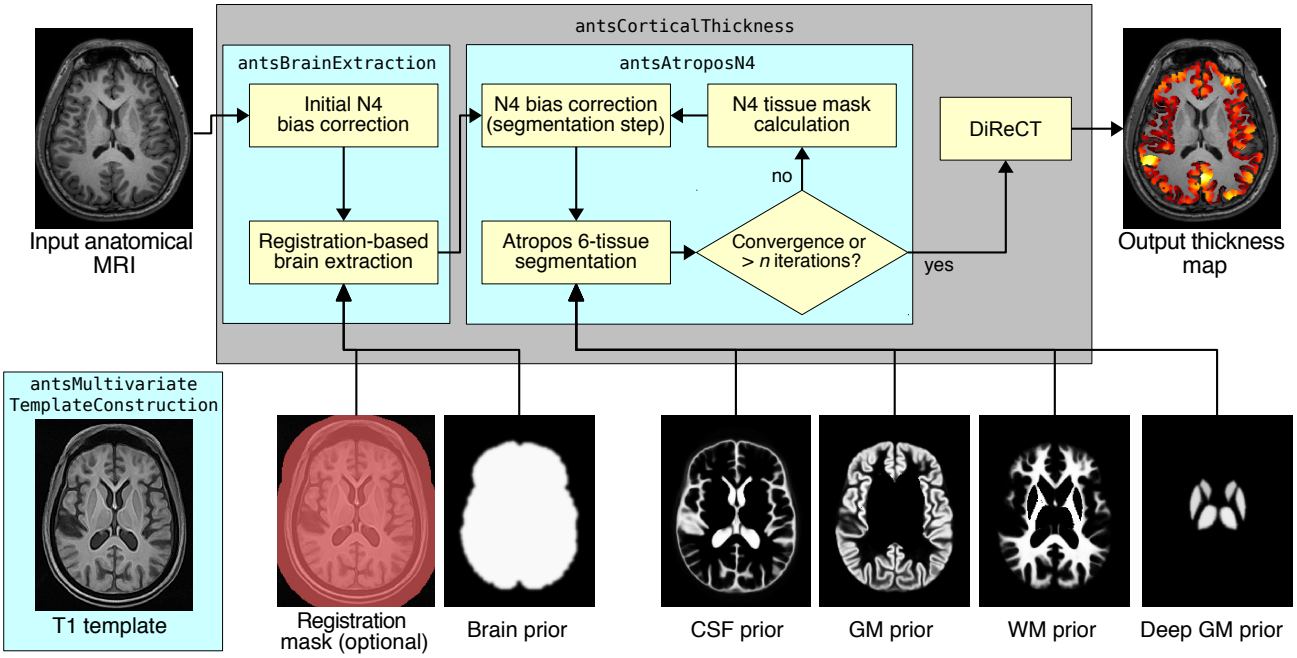

Figure 1: Illustration of the main components of the ANTs processing workflow containing all elements for determining cortical thickness. We also included the domain of operations for the selected scripts. Not shown are the probability maps for the brain stem and cerebellum priors.

In an evaluation study, we compared an earlier version of our extraction method with publicly available brain extraction algorithms, including AFNI's `3dIntracranial` [98], FSL's `BET2` [87], FreeSurfer's `mri_watershed` [80], and BrainSuite [24]. We demonstrated that our combined registration/segmentation approach [4] performed with an accuracy comparable to FreeSurfer and a parameter-tuned version of BrainSuite. Figure 2 presents a visual comparison of results derived with the current ANTs brain extraction method and results obtained using FreeSurfer.

2.1.3. *Atropos six-tissue segmentation*

In [6] we presented an open source n-tissue segmentation software tool (which we denote as "Atropos") that attempts to distill over 20 years of active research in this area, in particular some of the most seminal work (e.g., [106, 3]). Specification of prior probabilities includes spatially varying Markov Random Field modeling, prior label maps, and prior probability maps typically derived from our template building process. Additional capabilities include handling of multivariate data, partial volume modeling [82], a memory-minimization mode, label propagation, a plug-and-play architecture for incorporation of novel likelihood models which includes both parametric and non-parametric models for both scalar and tensorial images, and alternative posterior formulations for different segmentation tasks.

Due to the important interplay between segmentation and bias correction, we perform multiple N4 $\rightleftharpoons$ Atropos iterations. Atropos and N4 are integrated in the script `antsAtroposN4.sh`.[5] A pure tissue probability weight mask generated from the posterior probabilities is derived from the segmentation step. Given N labels and the corresponding N posterior probability maps $\{P_1, \ldots, P_N\}$ produced during segmentation, the N4 weight mask is created at each N4 $\rightleftharpoons$ Atropos iteration from

$$P_{pure\ tissue}(\mathbf{x}) = \sum_{i=1}^{N} P_i(\mathbf{x}) \prod_{j=1, j\neq i}^{N} \Big(1 - P_j(\mathbf{x})\Big). \qquad (1)$$

One of the key insights of the original N3 development is the observation that inhomogeneities cause the intensity values of pure tissue peaks to spread in the intensity histogram as though convolved with a Gaussian. A core contribution of N3 is the proposed corrective step of deconvolving the intensity histogram to accentuate the tissue peaks, coupled with a spatial smoothing constraint. The pure tissue probability mask weights more heavily the voxels corresponding to pure tissue types (as determined by the segmentation) during the deconvolution process while minimizing the contribution of regions such as the gray/white matter interface where peak membership is ambiguous.

Atropos enables prior knowledge to guide the segmentation process where template-based priors are integrated into the optimization with a user-controlled weight. Modulating the likelihood and prior contributions to the posterior probability is essential for producing adequate segmentations. Atropos weights the likelihood and priors according to $P(x|y) \propto P(y|x)^{1-\alpha} P(x)^{\alpha}$

[5] https://github.com/ntustison/antsAtroposN4Example

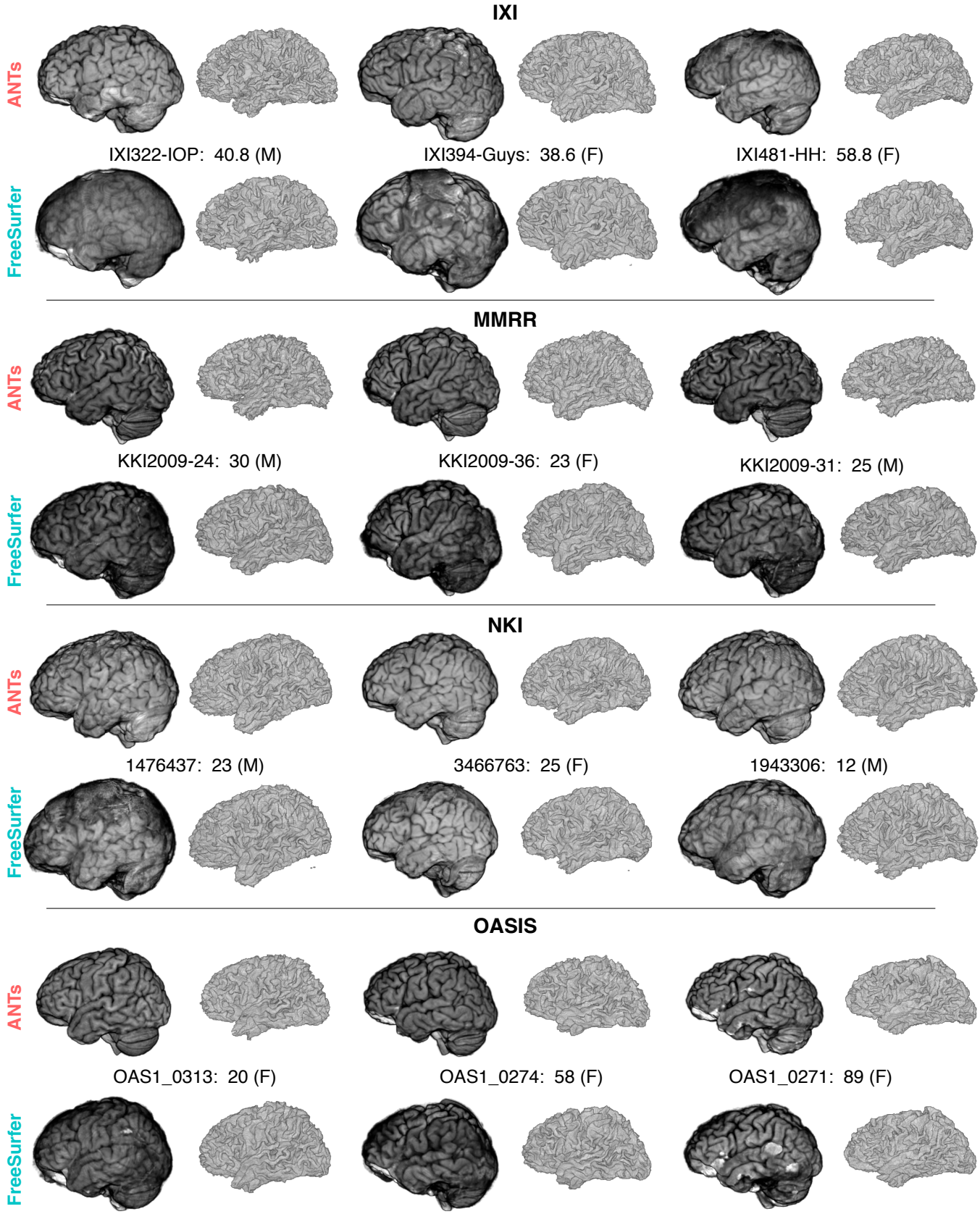

Figure 2: Representative sample of volume brain renderings from the four different cohorts (IXI = rows 1 and 2, MMRR = rows 3 and 4, NKI = rows 5 and 6, OASIS = rows 7 and 8), illustrating the qualitative difference between ANTs and FreeSurfer results, which are arranged top-and-bottom for each subject. Each brain was rigidly registered to the OASIS template for rendering purposes. With each subject we provide subject ID, age, and gender.

where α is a user-selected parameter which weights the trade-off between the likelihood and priors terms. A weighting of $\alpha = 0.25$ is the default value based on our extensive experimentation with different parameter weights.

Since cortical thickness estimation only requires the cortical gray and white matter, the deep gray and white matter (both labels and posterior maps) are combined to form a single "white matter" set. This white matter set and the cortical gray matter are the only results from the segmentation step that are used by the DiReCT algorithm (described below).

2.1.4. DiReCT cortical thickness estimation

DiReCT was introduced in [19] and was made available in ANTs as the program `KellySlater`. Since then several improvements have been made and incorporated into the program `KellyKapowski`.[6] The more recent implementation has made numerous advances, including: it is multi-threaded, written in rigorous ITK coding style,[7] and has been made publicly available through ANTs, complete with a unique command line interface design developed specifically for ANTs tools. Additionally, a fully functional, self-contained example using 2-dimensional image data is provided. [8]

2.1.5. Anatomical template construction

Normalizing images to a standard coordinate system reduces intersubject variability in population studies. Various approaches exist for determining the standardized coordinate space, such as the selection of a preexisting template based on a single individual (e.g., the Talairach atlas [90]) or a publicly available average of multiple individuals (e.g., the MNI [17] or ICBM [67] templates), or an average template constructed from the individuals under study. The work of [7] explicitly models the geometric component of the normalized space during optimization to produce such mean templates. Coupling the intrinsic symmetry of SyN pairwise registration [5] and an optimized shape-based sharpening/averaging of the template appearance, Symmetric Group Normalization (SyGN) is a powerful framework for producing optimal population-specific templates.

The ANTs implementation of this technique is currently available as a shell script, `buildtemplateparallel.sh`. A generalized, multivariate version is also available as `antsMultivariateTemplateConstruction.sh`. Both scripts are distributed as part of the ANTs software. The multivariate script permits the construction of multimodal templates created from a cohort of subjects, each with a set of pre-aligned multimodal images per subject. All modalities are used during the template-building process. For example, for N subjects where each subject has T1, T2, and fractional anisotropy (FA) images which are all aligned in the same subject space, a multimodal template consisting of T1, T2, and FA components can be created using antsMultivariateTemplateConstruction.sh. Both scripts accommodate a variety of computational resources for facilitating template construction.[9] These computational resource possibilities include:

- serial processing on a single workstation
- parallelized processing on a single workstation with multiple cores using `pexec`[10]
- parallelized processing using Apple's XGrid technology[11]
- parallelized processing using Sun Grid Engine for cluster-based systems[12]
- parallelized processing using the Portable Batch System for cluster-based systems[13]

For this work, cohort-specific templates were used during cortical thickness pipeline processing for both brain extraction and segmentation steps. Further details regarding the actual templates used (along with corresponding prior probability maps) are given in the section describing the public data used. The quality of prior probability images, particularly for the segmentation step, is crucial for performance.

To generate the priors for each T1 template, we used the multi-atlas label fusion (MALF) algorithm of [97] which is also distributed with ANTs and simplified with the `antsMalfLabeling.sh` script. We also used the training data associated with the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling[14] consisting of 20 labeled brain data taken from the OASIS data set.[15] Once the labels were propagated to the template, we condensed the 100+ labels associated with the training data to the six needed for our analysis, viz., cerebrospinal fluid (CSF), gray matter (GM), white matter (WM), deep gray matter, brain stem, and cerebellum. These binary masks were then smoothed using Gaussian convolution with a one voxel-width kernel (cf Figure 4). Since the labelings did not describe the extracerebral CSF, we augmented the CSF prior image with the CSF posterior output from running each template through the segmentation component of the above-described pipeline with a template that we had built previously. This new CSF prior was then subtracted from each of the other five prior probability images.

[6] Traditional academic discourse encountered in the published literature rarely contextualizes peculiarities such as algorithmic nomenclature. We briefly mention that this was the source of a rare disagreement between the first and last authors based, as many disagreements are, on a simple misunderstanding and not an affronting existential statement concerning a certain favorite sitcom of the first author's youth.

[7] http://www.itk.org/ITK/help/documentation.html

[8] https://github.com/ntustison/antsCorticalThicknessExample

[9] A self-contained 2-D example is available at https://github.com/ntustison/TemplateBuildingExample.

[10] http://www.gnu.org/software/pexec/pexec.1.html

[11] https://developer.apple.com/hardwaredrivers/hpc/xgrid_intro.html

[12] http://www.oracle.com/technetwork/oem/grid-engine-166852.html

[13] http://www.pbsworks.com/

[14] https://masi.vuse.vanderbilt.edu/workshop2012/index.php/Main_Page

[15] The data was released under the Creative Commons Attribution-NonCommercial license. Labelings were provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription.

1) caudal anterior cingulate	17) pars orbitalis
2) caudal middle frontal	18) pars triangularis
3) cuneus	19) pericalcarine
4) entorhinal	20) postcentral
5) fusiform	21) posterior cingulate
6) inferior parietal	22) precentral
7) inferior temporal	23) precuneus
8) isthmus cingulate	24) rosterior anterior cingulate
9) lateral occipital	25) rostral middle frontal
10) lateral orbitofrontal	26) superior frontal
11) lingual	27) superior parietal
12) medial orbitofrontal	28) superior temporal
13) middle temporal	29) supramarginal
14) parahippocampal	30) transverse temporal
15) paracentral	31) insula
16) pars opercularis	

Table 1: The 31 cortical labels (per hemisphere) of the DKT atlas.

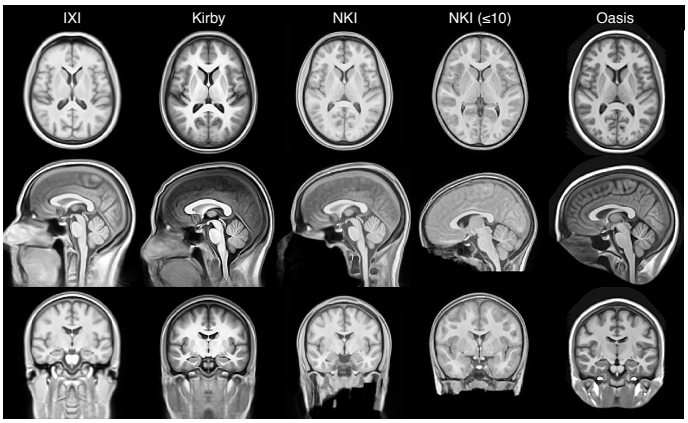

Figure 3: Population-specific templates for each of the four public data sets used for cortical thickness estimation. These templates were generated using the script `antsMultivariateTemplateConstruction.sh`. The benefit of using such population-specific templates is obvious when one sees the variability in acquisition and data preparation (e.g., defacing protocols).

2.2. Public data resources

The above pipeline was run on four publicly available data sets: IXI, MMRR, NKI, and OASIS. In addition, we used a subset of the MindBoggle-101 data[16] labeled using the Desikan-Killiany-Tourville (DKT) protocol [50] to define the regions of interest (ROI). This latter data set was not included in the thickness analysis. All five data sets are described below.

2.2.1. MindBoggle-101 data for ROI definitions

In [50] the authors proposed the DKT cortical labeling protocol—a modification of the popular Desikan-Killiany protocol [22] to improve cortical labeling consistency and to improve FreeSurfer's cortical classification of 31 cortical regions per hemisphere, listed in Table 1. For the latter, forty manually labeled brains were used to construct the DKT40 Gaussian classifier atlas, which is now bundled with current versions of FreeSurfer and used to automate anatomical labeling of MRI data. Since the regional thickness values generated by FreeSurfer follow this protocol, these anatomical labels provide a common standard for comparison between ANTs and FreeSurfer.

The work of [50] also resulted in a publicly available set of manually edited labels following the DKT protocol in 101 T1-weighted brain images from different sources, including a subset of 20 images from the OASIS data set (specifically, the test-retest data). These 20 images are used in the MALF step that defines the volumetric cortical regions for each subject.

2.2.2. Public data for thickness estimation evaluation

We applied the same pipeline to diverse publicly available data sets collected from multiple sites and with a mixture of 3T and 1.5T T1-weighted brain images. Subjects in this data set span the age range from 4 to 96 years old. This strategy tested robustness to variation in head position, brain shape, defacing, image contrast, inhomogeneity, imaging artifacts, and the broad variation in extracerebral tissue. Failure can occur in initial brain extraction, segmentation, registration, or bias correction, any of which can lead to an inaccurate cortical thickness measurement.

In total, we processed 1,205 T1-weighted images from four different public data sets to obtain cortical thickness maps. Below we describe the four data sets and their corresponding templates:

IXI. Initially, we processed 581 T1-weighted images from the IXI[17] data set, but only 563 subjects (313 females, 250 males) were included in the post-processing analysis due to missing demographic information, which would have prevented an accurate estimate of the age at the time of image acquisition. These data were imaged at three sites with several modalities acquired (T1-weighted, T2-weighted, proton density, magnetic resonance angiography, and diffusion tensor imaging). The database also includes demographic information such as date of birth, date of scan, weight, height, ethnicity, occupation category, educational level, and marital status.

The IXI template had been built previously during an investigation of multimodal (T1, T2, FA, and proton density) template generation using different aged cohorts. Templates consisting of subjects within decade ranges, i.e., 10 to 20, 20 to 30, etc. were built. These age-based multimodal templates were then used to build an average "meta-template" of which the T1 component is used in this study as seen, respectively, in the first column and row of Figures 3 and 4.

MMRR. The Multi-Modal MRI Reproducibility Resource (MMRR)[18] data set, was originally described in [55] consisting of 21 subjects (10 females, 11 males) and features a rich set of modalities, as well as repeated scans.

[16] http://mindboggle.info/data.html

[17] http://biomedic.doc.ic.ac.uk/brain-development/

[18] http://www.nitrc.org/projects/multimodal/

In previously testing our template construction protocol, we used all MMRR data sets to construct a multimodal template consisting of FA, mean diffusivity (MD), FLAIR, T1, T2, magnetic transfer (MT), and vascular space occupancy (VASO) components. More details concerning data acquired are provided at [55]. The T1 component used in this study can be seen, respectively, in the second column and row of Figures 3 and 4.

NKI. In support of open science, the 1,000 Functional Connectomes Project[19] was initiated on December 11, 2009 by various members of the MRI community seeking to form collaborative partnerships among imaging institutions for sharing well-documented multimodal image sets accompanied by phenotypic data. One such contribution is the Nathan Klein Institute (NKI)/Rockland sample, consisting of 186 T1-weighted images (87 females, 99 males).[20]

The T1-weighted template was constructed originally for the study performed in [93]. It was created from 30 randomly sampled subjects which we repurpose here for our cortical thickness evaluation. Since the NKI cohort contains younger individuals, for this study we also created a template for subjects of 10 years and younger generated from 13 individuals. Both templates and corresponding priors are also seen in Figures 3 and 4.

OASIS. The initial Open Access Series of Imaging Studies (OASIS)[21] data set consisted of 433 T1-weighted images. We processed all of these, but 100 were excluded from our analysis due to probable Alzheimer's disease ($CDR > 0$) and an additional 20 repeat scans were excluded, resulting in 313 individual subject scans included in the normal group statistical analysis (118 males, 195 females). Ages were between 18 and 96 and all subjects are right-handed.

As part of the MICCAI 2012 Grand Challenge mentioned earlier, the first and last authors were asked to provide the "canonical" registrations for each pairwise mapping. Instead of performing n^2 registrations where n is the number of subjects, we decided to produce all mappings using a "pseudo-geodesic" approach where all subjects are warped to an optimal template thus providing any pairwise transformation and thereby reducing the number of image registrations required from n^2 to n. The optimal template for this pseudo-geodesic processing was constructed from the combined 35 testing and training OASIS data of the MICCAI 2012 Grand Challenge which we repurpose for this study.

2.3. Processing miscellany

Given the documented variability in FreeSurfer results with version and operating system [37] (as we would expect with our own ANTs pipeline), all data were processed using the same ANTs and FreeSurfer versions on the same hardware platform. Processing was performed using the Linux (CentOS release 6.4) cluster at the University of Virginia[22] using single-threading with a maximal requested memory footprint of 8 gb for ANTs and 4 gb for FreeSurfer. The development version of ANTs was used for processing (git commit tag: 69d3a5a6c7125ccf07a9e9cf6ef29f0b91e9514f, date Dec. 11, 2013). FreeSurfer version 5.3 x86_64 for CentOs was downloaded on 5 December, 2013 ("freesurfer-Linux-centos6_x86_64-stable-pub-v5.3.0", release date: 15 May, 2013).

We visually inspected all brain extraction and segmentation results from the ANTs and FreeSurfer pipelines. This cursory quality control step consisted of a script which displayed three cross-sectional views, one along each viewing axis in a time-series of ITK-SNAP window displays. No manual changes were made for any component of either pipeline and no change was made to the settings of either processing pipeline.

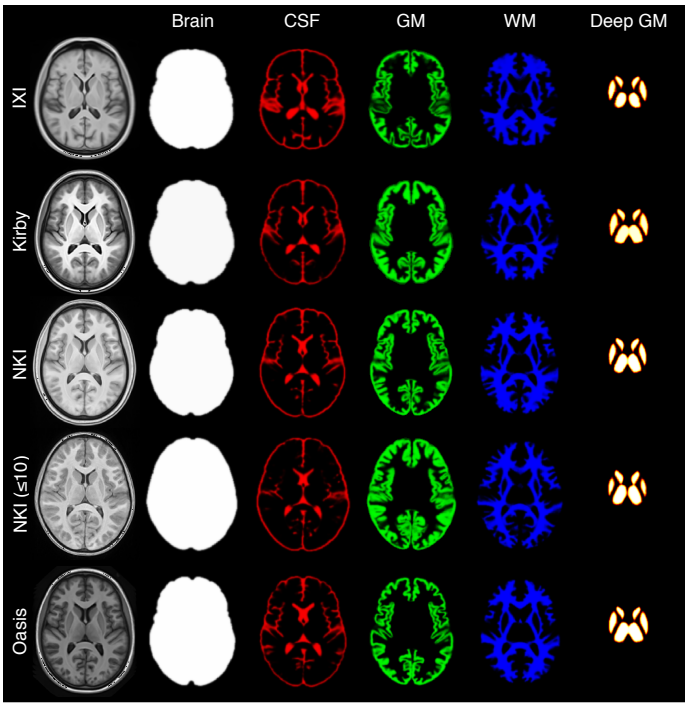

Figure 4: Axial slices from each of the five T1 templates including the corresponding probability masks used for brain extraction and brain tissue segmentation. Not shown are the prior probability maps for brain stem and cerebellum regions.

3. Evaluation

Traditional assessment approaches, such as manual labeling, are inadequate for evaluating large-scale performance. We therefore sought to minimize failure rate, quantify the repeatability of cortical thickness measures, and determine whether the ANTs pipeline reveals biologically plausible relationships between the cortex, gender,[23] and age and how its performance

[19] http://fcon_1000.projects.nitrc.org
[20] Downloaded on September 22, 2012.
[21] http://www.OASIS-brains.org/
[22] http://www.uvacse.virginia.edu/
[23] We recognize the distinction often made between "sex" and "gender" (cf `http://www.who.int/gender/whatisgender/en/`). As the demographic

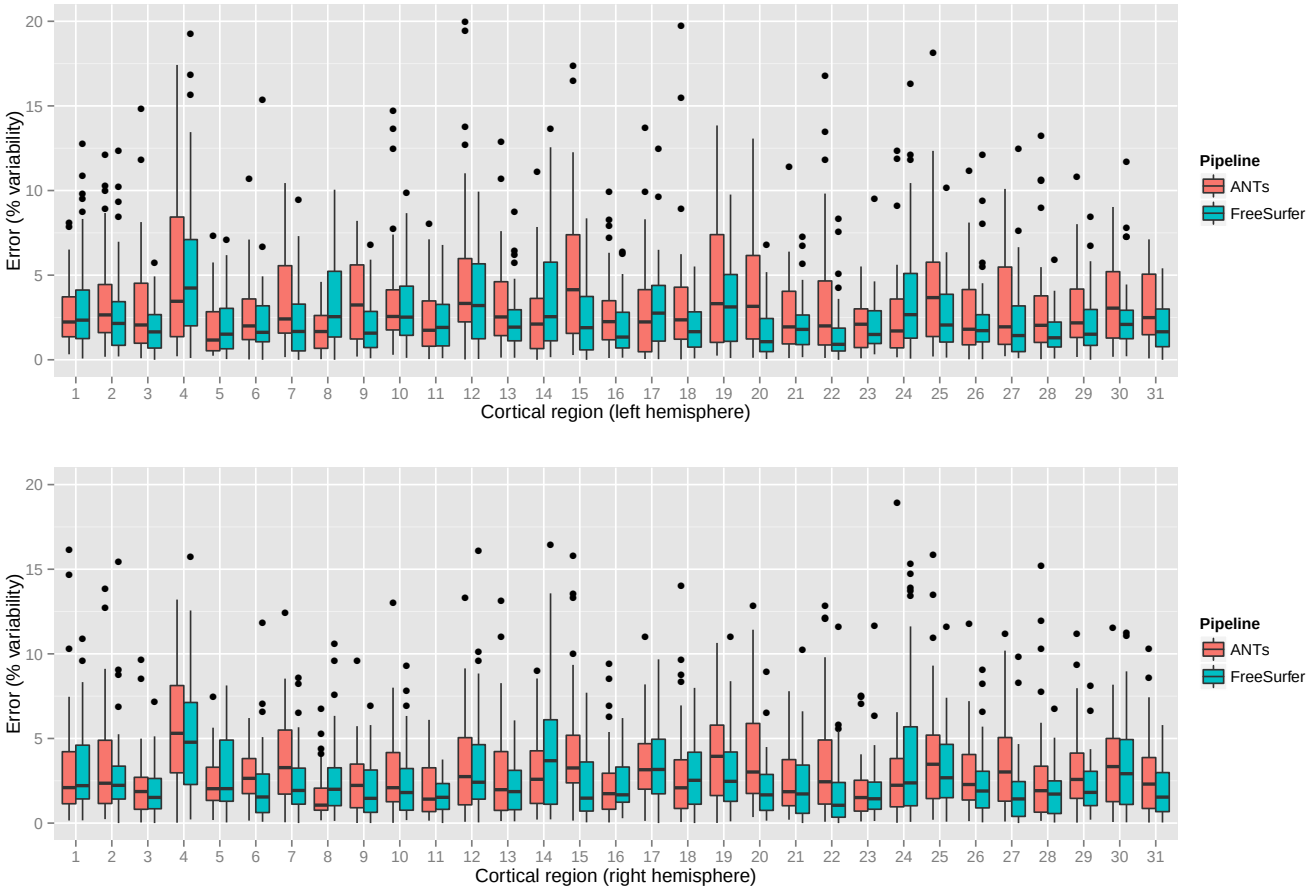

Figure 5: Percent error variability for both ANTs and FreeSurfer pipelines over the left and right hemispheres of both the MMRR and OASIS data subsets within the 62 regions defined by the Desikan-Killiany-Tourville atlas. Both methods demonstrate good repeatability qualities.

compares to the current de facto standard of FreeSurfer-derived thickness estimation. Collectively, these surrogate measurements allow us to establish data-derived relative performance standards. Additionally, for completeness, we include timing results as that factors into usability.

3.1. Repeatability

Repeat scans of 40 subjects (20 MMRR subjects and 20 OASIS subjects) were used to determine the repeatability of regional cortical thickness measurements, T. Similar to the reproducibility assessment given in [46], we demonstrate this in terms of the percent variability error:

$$\varepsilon = \frac{|T_{scan} - T_{rescan}|}{0.5 \times (T_{scan} + T_{rescan})}. \quad (2)$$

Comparison of the ANTs and FreeSurfer percent variability errors for the 62 DKT regions for both the OASIS and MMRR scan-rescan data sets are given in Figure 5. Although the variance is slightly greater for the set of ANTs measurements, statistical testing per cortical region (two-tailed paired t-test, corrected using false discovery rate) did not indicate non-zero mean differences for either approach for any region.

We also calculated the intraclass correlation coefficient ("ICC(2,1)" in the notation of [85]) to assess scan/rescan reliability. The ANTs thickness pipeline produced an ICC value of 0.98 and the FreeSurfer thickness pipeline yielded an ICC value of 0.97, indicating good scan/rescan reliability for both ANTs and FreeSurfer.

3.2. Age prediction assessment

Despite good repeatability with both ANTs and FreeSurfer, such measures do not provide an assessment of accuracy or even relative utility. For example, strong priors can yield good repeatability measures but potentially at the expense of data fidelity thus compromising the quality of models (statistical or otherwise) built from such results. Given that ground truth is not available for these data nor for the many studies looking at brain morphology, an indirect method (or set of methods) is required for determining the quality of thickness estimation.

Previous research has used predictive modeling for comparing cortical thickness algorithms. For example, in [16], classification of healthy, semantic dementia, and progressive non-

information collected during the course of the imaging studies is presumably self-reported, we assume that most self-identify in terms of gender and, therefore, use the term "gender" in data descriptions.

fluent aphasia categories using regional cortical thickness values was used to determine the predictive modeling capabilities of different cortical thickness processing protocols in 101 subjects. However, differential diagnosis of dementia [68] is not as straightforward as obtaining a subject's age or gender and regressing that against cortical thickness; the latter constitute biological relationships that have been well-studied and reported in the literature.

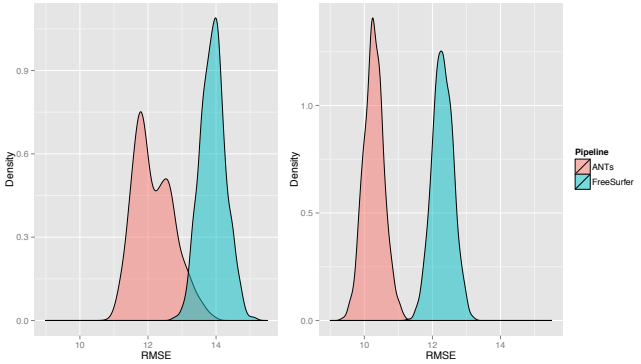

Figure 6: Age prediction RMSE distributions of linear (left) and random forest (right) models for the ANTs- and FreeSurfer-derived thickness values. For both prediction models ANTs RMSE error is lower.

For our first assessment, we modeled age versus regional cortical thickness values to determine which framework produces better predictive thickness estimates. We first subdivided the thickness data into training and testing subsets with an even split between the two subsets.[24] We then used the training data to create two models for each pipeline: 1) standard linear regression and 2) random forests (a non-parametric machine learning technique) [12], for estimating age from both ANTs and FreeSurfer thickness values in the testing data.

The formula (in the notation of [101]) for the linear model is

$$AGE \sim VOLUME + \sum_{i=1}^{62} T(DKT_i) * GENDER \qquad (3)$$

where $T(DKT_i)$ is the average thickness value in region DKT_i. Similarly, the random forest model was specified as a combination of all terms using the `randomForest`[25] package in R with the default settings and 200 trees.

In order to ensure a fair comparison, the procedure described above consisting of training and testing steps was performed for $n = 1000$ permutations to elicit a performance distribution which we measure using the relative mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum \left(AGE_{true} - AGE_{predicted}\right)^2}{N}}. \qquad (4)$$

[24] We tried various training proportions between 10 and 90% (in increments of 10%) to see if that had an effect on relative performance for both age and gender prediction comparisons. Although age predictive capabilities for both pipelines showed improvement (gender prediction was mostly unaffected), the relative outcomes were the same.

[25] http://cran.r-project.org/package=randomForest

Table 2: Mean RMSE for age prediction (in years).

	Linear	**Random Forest**
ANTs	12.2	10.3
FreeSurfer	13.9	12.3

The resulting distributions are illustrated in Figure 6 with the linear model results displayed on the left and random forest results on the right. The RMSE value was lower with ANTs thickness values for both models with the ANTs-based random forest predictions performing the best. Mean RMSE values are provided in Table 2.

3.3. Gender prediction assessment

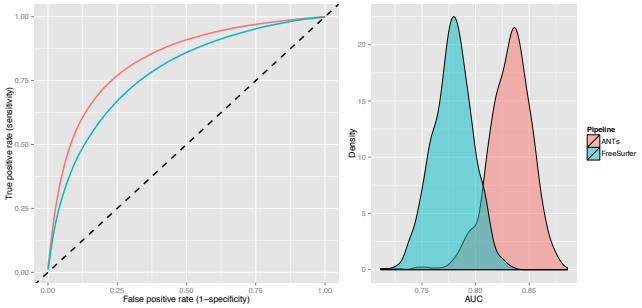

Figure 7: Average ROC curve and corresponding AUC distributions for gender prediction using ANTs and FreeSurfer thickness values. Values were averaged for 1000 permutations resulting in mean values of $\text{ANTs}_{AUC} = 0.83$ and $\text{FreeSurfer}_{AUC} = 0.78$ ($p < 2.2 \times 10^{-16}$).

We also performed a similar prediction assessment using gender as the regressand. The binomial generalized linear model is

$$GENDER \sim VOLUME + \sum_{i=1}^{62} T(DKT_i) * AGE \qquad (5)$$

where $T(DKT_i)$ is the average thickness value in region DKT_i. We then characterized performance using a ROC curve for both methods (see Figure 7) where we averaged over 1000 permutations. The mean area under the curve (AUC) for both methods was also quantified with values of $\text{ANTs}_{AUC} = 0.83$ and $\text{FreeSurfer}_{AUC} = 0.78$.

3.4. Computation time

All images underwent the ANTs and FreeSurfer pipeline processing using the computational cluster at the University of Virginia. Processing times varied approximately between 10–20 hours per subject for both pipelines for the entire cortical thickness estimation procedure although ANTs processing, on average, took slightly longer (cf Figure 8). Averaged over all cohorts, ANTs required 15.7 ± 2.0 hours per subject and FreeSurfer required 14.1 ± 2.9 hours per subject.

The propagation of the DKT labels to each subject using label fusion as described earlier was performed in parallel and

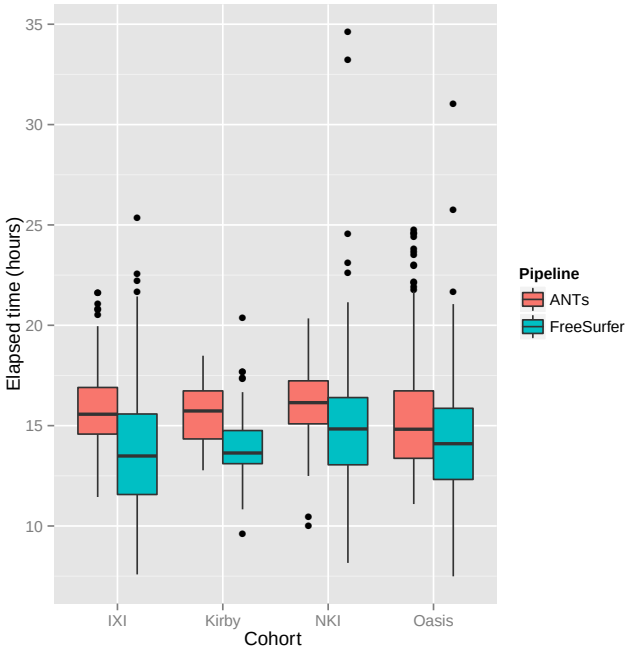

Figure 8: Elapsed time across data sets for ANTs and FreeSurfer processing. Averaged over all cohorts, ANTs required 15.7 ± 2.0 hours per subject and FreeSurfer required 14.1 ± 2.9 hours per subject.

took anywhere between 40 and 80 hours per subject for 16 serial image registrations and application of the joint label fusion algorithm [97]. For each subject, 20 atlas registrations are used to generate the labeling for that subject. Therefore to do the MALF labeling for the entire cohort, approximately $1200 \times 20 = 24000$ registrations were required. The `antsMalfLabeling.sh` script mentioned earlier parallelizes the registration component which decreases the time for parallel computation platforms.

4. Discussion

In the absence of ground truth, we used repeatability and prediction of demographic variables to compare the ANTs and FreeSurfer cortical thickness pipelines. One very important issue that was not discussed in this work is quality control for ensuring proper pipeline processing. The time required to go through approximately 1200 sets of results (×2 for both pipelines) would be enormous (not to mention the tedium). However, the first author did do this for the brain extraction step to ensure that both pipelines were achieving expected intermediate results. The only major failure was the FreeSurfer brain extraction of a single IXI subject (IXI430-IOP-0990). Also, three NKI subjects were not processed to completion with FreeSurfer (1713515, 18755434, and 2674565) and were not included in the analysis. Although researchers might quibble over processing minutiae such as the inclusion of too much (or not enough) of the meninges, we approached our evaluation using more objective criteria which concern all those engaged in this type of research. We are currently trying to develop methods to facilitate data inspection for quick quality assurance/control.

4.1. Repeatability of thickness measurements

The OASIS data set and the MMRR data set allow us to test whether the same thickness values emerge from T1-weighted MRI collected on the same subject but at different times of the day or over a time separation within a few weeks. Although the ANTs cortical thickness pipeline produced similar repeatability assessments as FreeSurfer in these data, there are many additional issues to explore with the ANTs-based framework. Pre-analysis confounds such as short-term alterations in cortical morphology due to the T1-weighted susceptibility to blood flow [34, 78, 103] and MRI acquisition parameters such as field strength, site, resolution, scanner, longitudinal variation in scanner conditions, and pulse sequence [40, 62, 46] have been evaluated with FreeSurfer which has shown good reliability under various permutations of these conditions. Although we did not explicitly investigate the repeatability performance of the ANTs framework under such effects, the relatively good performance on the large and varied data (in terms of site, field strength, scanner, and acquisition sequence) used in this study provides confidence in its robustness to a variety of imaging conditions.

However, as we mentioned previously, good repeatability does not necessarily translate to accurate measurements. A crucial component in estimating cortical thickness is accurate gray matter segmentation for which several algorithms are available. A recent study compared FreeSurfer, FSL's FAST [106], SPM8 [3], and VBM8 (an extension of SPM8)[26] in terms of both reliability and accuracy [26]. Good accuracy was achieved with the latter three methods with Dice coefficients of > 0.93 for all regions in both simulated and real T1-weighted data. In contrast, FreeSurfer yielded slightly less accurate results (mean Dice > 0.88) on simulated data and even less accurate results for real data (mean Dice > 0.58) Despite the relatively poor accuracy, FreeSurfer was extremely reliable, demonstrating the least variability in calculated gray matter volumes for single-subject, 10 scan data and the lowest average volume differences in the OASIS scan-rescan data.

4.2. Age and gender prediction

Although repeatability between ANTs and FreeSurfer is comparable, such measures are not as useful in determining the utility of the measuring software. That is the reason we used a training and testing paradigm to evaluate how well both frameworks produce measurements capable of predicting demographics which are well-known to correlate with cortical thickness. Additionally, these demographic measures are probably some of the easiest and most reliably obtained of all possible demographic measures used for this type of assessment. For age prediction, we used both a linear model (due to its general ubiquity) and a random forest model (a non-parametric model to contrast with the linear approach) which showed overall good

[26] http://dbm.neuro.uni-jena.de/vbm/

performance. Also, the linear and random forest models have the advantage of being interpretable. That is, the models reveal the specific predictors that are most valuable which will be explored in future work. Another interesting result from the age prediction study was that predictive performance for the linear model degenerated towards the 10% level, i.e. when using approximately 10% of the total number of subjects for training and the remaining 90% to test model predictive performance (see supplementary material). This translates into approximately 100 subjects being used for training, raising concerns about the use of smaller cohorts for performance comparisons with these data.

This study was limited to a cross-sectional investigation thus limiting extrapolations of ANTs performance to longitudinal data unlike recent FreeSurfer extensions which accommodate longitudinal data [75, 46]. Also, some users may choose to segment and register with ANTs and subsequently employ any alternative (e.g., surface-based) method for thickness estimation. Further work is needed by independent authors working on established pipelines to better compare surface-based and volume-based thickness reliability and accuracy across different populations, age ranges, and with longitudinal protocols.

4.3. Computation time

Computation time for the registration and segmentation components of the ANTs pipeline are substantial. It is likely that nearly as reliable results can be obtained in much less time for many of the subjects in this study. However, our interest in maximizing reliability and quality led us to employ parameters in the registration, segmentation, and bias correction that are as robust as possible to differences in head position, the presence of large deformations between template and target brains and substantial inhomogeneity or other artifacts in the image content itself. Several subjects (e.g., NKI: 1898228, 1875434) provide examples of more difficult data from which we are able to extract meaningful segmentations and registrations, despite the presence of a "garbage-in/garbage-out" problem. A subject of future study is determining an exact cut-off for the inclusion of such data. We do not investigate this issue here, which has concerned statisticians for over half a century [39].

5. Conclusions

Imaging biomarkers such as cortical thickness play an important role in neuroscience research. Extremely useful to researchers are robust software tools for generating such biomarkers. In this work we detailed our open source offering for estimating cortical thickness directly from T1 images and demonstrated its utility on a large collection of public brain data from multiple databases acquired at multiple sites. To our knowledge, this study constitutes the largest collection of cortical thickness data processed in a single study. We anticipate that public availability of our tools and extensive tuning on the specified cohorts will prove useful to the larger research community. In this work, we only explored a portion of the potentially interesting investigations possible with these data. Since all of the data are publicly available, further work can be easily pursued by us or by other interested groups.

References

[1] Almeida Montes, L. G., Prado Alcántara, H., Martínez García, R. B., De La Torre, L. B., Avila Acosta, D., Duarte, M. G., Mar 2012. Brain cortical thickness in ADHD: Age, sex, and clinical correlations. J Atten Disord.

[2] Amunts, K., Armstrong, E., Malikovic, A., Hömke, L., Mohlberg, H., Schleicher, A., Zilles, K., Feb 2007. Gender-specific left-right asymmetries in human visual cortex. J Neurosci 27 (6), 1356–64.

[3] Ashburner, J., Friston, K. J., Jul 2005. Unified segmentation. Neuroimage 26 (3), 839–51.

[4] Avants, B. B., Klein, A., Tustison, N. J., Woo, J., Gee, J. C., 2010. Evaluation of an open-access, automated brain extraction method on multi-site multi-disorder data. Human Brain Mapping.

[5] Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., Gee, J. C., Feb 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–44.

[6] Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., Gee, J. C., Dec 2011. An open source multivariate framework for *n*-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400.

[7] Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J. C., Feb 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49 (3), 2457–66.

[8] Ballmaier, M., Sowell, E. R., Thompson, P. M., Kumar, A., Narr, K. L., Lavretsky, H., Welcome, S. E., DeLuca, H., Toga, A. W., Feb 2004. Mapping brain size and cortical gray matter changes in elderly depression. Biol Psychiatry 55 (4), 382–9.

[9] Bermudez, P., Lerch, J. P., Evans, A. C., Zatorre, R. J., Jul 2009. Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. Cereb Cortex 19 (7), 1583–96.

[10] Bouix, S., Martin-Fernandez, M., Ungar, L., Nakamura, M., Koo, M.-S., McCarley, R. W., Shenton, M. E., Jul 2007. On evaluating brain tissue classifiers without a ground truth. Neuroimage 36 (4), 1207–24.

[11] Boyes, R. G., Gunter, J. L., Frost, C., Janke, A. L., Yeatman, T., Hill, D. L. G., Bernstein, M. A., Thompson, P. M., Weiner, M. W., Schuff, N., Alexander, G. E., Killiany, R. J., DeCarli, C., Jack, C. R., Fox, N. C., ADNI Study, Feb 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. Neuroimage 39 (4), 1752–62.

[12] Breiman, L., 2001. Random forests. In: Machine Learning. pp. 5–32.

[13] Chen, J. J., Rosas, H. D., Salat, D. H., Mar 2011. Age-associated reductions in cerebral blood flow are independent from regional atrophy. Neuroimage 55 (2), 468–478.
URL `http://dx.doi.org/10.1016/j.neuroimage.2010.12.032`

[14] Chen, Z. J., He, Y., Rosa-Neto, P., Germann, J., Evans, A. C., Oct 2008. Revealing modular architecture of human brain structural networks by using cortical thickness from MRI. Cereb Cortex 18 (10), 2374–81.

[15] Chung, M. K., Robbins, S. M., Dalton, K. M., Davidson, R. J., Alexander, A. L., Evans, A. C., May 2005. Cortical thickness analysis in autism with heat kernel smoothing. Neuroimage 25 (4), 1256–65.

[16] Clarkson, M. J., Cardoso, M. J., Ridgway, G. R., Modat, M., Leung, K. K., Rohrer, J. D., Fox, N. C., Ourselin, S., Aug 2011. A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage 57 (3), 856–65.

[17] Collins, D. L., Neelin, P., Peters, T. M., Evans, A. C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr 18 (2), 192–205.

[18] Dale, A. M., Fischl, B., Sereno, M. I., Feb 1999. Cortical surface-based analysis. i. segmentation and surface reconstruction. Neuroimage 9 (2), 179–94.

[19] Das, S. R., Avants, B. B., Grossman, M., Gee, J. C., Apr 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–79.

[20] DaSilva, A. F. M., Granziera, C., Snyder, J., Hadjikhani, N., Nov 2007. Thickening in the somatosensory cortex of patients with migraine. Neurology 69 (21), 1990–5.

[21] Davatzikos, C., Bryan, N., 1996. Using a deformable surface model to obtain a shape representation of the cortex. IEEE Trans Med Imaging 15 (6), 785–95.

[22] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., Killiany, R. J., Jul 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31 (3), 968–80.

[23] Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., Sperling, R. A., Atri, A., Growdon, J. H., Hyman, B. T., Morris, J. C., Fischl, B., Buckner, R. L., Mar 2009. The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb Cortex 19 (3), 497–510.

[24] Dogdas, B., Shattuck, D. W., Leahy, R. M., Dec 2005. Segmentation of skull and scalp in 3-D human MRI using mathematical morphology. Hum Brain Mapp 26 (4), 273–85.

[25] Du, A.-T., Schuff, N., Kramer, J. H., Rosen, H. J., Gorno-Tempini, M. L., Rankin, K., Miller, B. L., Weiner, M. W., Apr 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130 (Pt 4), 1159–66.

[26] Eggert, L. D., Sommer, J., Jansen, A., Kircher, T., Konrad, C., 2012. Accuracy and reliability of automated gray matter segmentation pathways on real and simulated structural magnetic resonance images of the human brain. PLoS One 7 (9), e45081.

[27] Evans, A. C., Janke, A. L., Collins, D. L., Baillet, S., Aug 2012. Brain templates and atlases. Neuroimage 62 (2), 911–22.

[28] Fischl, B., Dale, A. M., Sep 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc Natl Acad Sci U S A 97 (20), 11050–5.

[29] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A. M., Jan 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–55.

[30] Fischl, B., Sereno, M. I., Dale, A. M., Feb 1999. Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207.

[31] Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A. M., Jan 2004. Automatically parcellating the human cerebral cortex. Cereb Cortex 14 (1), 11–22.

[32] Fortier, C. B., Leritz, E. C., Salat, D. H., Venne, J. R., Maksimovskiy, A. L., Williams, V., Milberg, W. P., McGlinchey, R. E., Dec 2011. Reduced cortical thickness in abstinent alcoholics and association with alcoholic behavior. Alcohol Clin Exp Res 35 (12), 2193–201.

[33] Foster, N. E. V., Zatorre, R. J., Oct 2010. Cortical structure predicts success in performing musical transformation judgments. Neuroimage 53 (1), 26–36.

[34] Franklin, T. R., Wang, Z., Shin, J., Jagannathan, K., Suh, J. J., Detre, J. A., O'Brien, C. P., Childress, A. R., 2013. A VBM study demonstrating 'apparent' effects of a single dose of medication on T1-weighted MRIs. Brain Structure and Function 218 (1), 97–104.

[35] Frøkjær, J. B., Bouwense, S. A. W., Olesen, S. S., Lundager, F. H., Eskildsen, S. F., van Goor, H., Wilder-Smith, O. H. G., Drewes, A. M., Apr 2012. Reduced cortical thickness of brain areas involved in pain processing in patients with chronic pancreatitis. Clin Gastroenterol Hepatol 10 (4), 434–8.e1.

[36] Gernsbacher, M. A., Jan 2007. Presidential column: The eye of the beholder. Observer 20 (1).

[37] Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., 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 (6), e38234.

[38] Haier, R. J., Karama, S., Leyba, L., Jung, R. E., 2009. MRI assessment of cortical thickness and functional activity changes in adolescent girls following three months of practice on a visual-spatial task. BMC Res Notes 2, 174.

[39] Hampel, F. R., March 2001. Robust statistics: A brief introduction and overview. Research Report 94, Eidgenössische Technische Hochschule.

[40] Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B., Fischl, B., Aug 2006. Reliability of mri-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. Neuroimage 32 (1), 180–194.
URL `http://dx.doi.org/10.1016/j.neuroimage.2006.02.051`

[41] Hardan, A. Y., Muddasani, S., Vemulapalli, M., Keshavan, M. S., Minshew, N. J., Jul 2006. An MRI study of increased cortical thickness in autism. Am J Psychiatry 163 (7), 1290–2.

[42] He, Y., Chen, Z. J., Evans, A. C., Oct 2007. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. Cereb Cortex 17 (10), 2407–19.

[43] Heim, C. M., Mayberg, H. S., Mletzko, T., Nemeroff, C. B., Pruessner, J. C., Jun 2013. Decreased cortical representation of genital somatosensory field after childhood sexual abuse. Am J Psychiatry 170 (6), 616–23.

[44] Jiang, J., Zhu, W., Shi, F., Liu, Y., Li, J., Qin, W., Li, K., Yu, C., Jiang, T., Feb 2009. Thick visual cortex in the early blind. J Neurosci 29 (7), 2205–11.

[45] Jones, S. E., Buchbinder, B. R., Aharon, I., Sep 2000. Three-dimensional mapping of cortical thickness using Laplace's equation. Hum Brain Mapp 11 (1), 12–32.

[46] Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Tränkner, A., Schönknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J.-P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalló, N., Blin, O., Frisoni, G. B., The PharmaCog Consortium, May 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: A comparison of cross-sectional and longitudinal segmentations. Neuroimage.

[47] Jubault, T., Gagnon, J.-F., Karama, S., Ptito, A., Lafontaine, A.-L., Evans, A. C., Monchi, O., Mar 2011. Patterns of cortical thickness and surface area in early Parkinson's disease. Neuroimage 55 (2), 462–7.

[48] Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., Evans, A. C., Aug 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 (1), 210–21.

[49] Klein, A., Ghosh, S. S., Avants, B., Yeo, B. T. T., Fischl, B., Ardekani, B., Gee, J. C., Mann, J. J., Parsey, R. V., May 2010. Evaluation of volume-based and surface-based brain image registration methods. Neuroimage 51 (1), 214–20.

[50] Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front Neurosci 6, 171.

[51] Kochunov, P., Glahn, D. C., Lancaster, J., Thompson, P. M., Kochunov, V., Rogers, B., Fox, P., Blangero, J., Williamson, D. E., Sep 2011. Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan. Neuroimage 58 (1), 41–9.

[52] Kovacevic, J., 2006. From the editor-in-chief. IEEE Trans Imag Process 15 (12).

[53] Kühn, S., Schubert, F., Gallinat, J., Dec 2010. Reduced thickness of medial orbitofrontal cortex in smokers. Biol Psychiatry 68 (11), 1061–5.

[54] Kuperberg, G. R., Broome, M. R., McGuire, P. K., David, A. S., Eddy, M., Ozawa, F., Goff, D., West, W. C., Williams, S. C. R., van der Kouwe, A. J. W., Salat, D. H., Dale, A. M., Fischl, B., Sep 2003. Regionally localized thinning of the cerebral cortex in schizophrenia. Arch Gen Psychiatry 60 (9), 878–88.

[55] Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., Smith, S. A., Joel, S., Mori, S., Pekar, J. J., Barker, P. B., Prince, J. L., van Zijl, P. C. M., Feb 2011. Multi-parametric neuroimaging reproducibility: a 3-t resource study. Neuroimage 54 (4), 2854–66.

[56] Lazar, S. W., Kerr, C. E., Wasserman, R. H., Gray, J. R., Greve, D. N., Treadway, M. T., McGarvey, M., Quinn, B. T., Dusek, J. A., Benson, H., Rauch, S. L., Moore, C. I., Fischl, B., Nov 2005. Meditation experience is associated with increased cortical thickness. Neuroreport 16 (17), 1893–7.

[57] Lemaitre, H., Goldman, A. L., Sambataro, F., Verchinski, B. A., Meyer-Lindenberg, A., Weinberger, D. R., Mattay, V. S., Mar 2012. Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? Neurobiol

Aging 33 (3), 617.e1–617.e9.
URL http://dx.doi.org/10.1016/j.neurobiolaging.2010.07.013

[58] Lerch, J. P., Worsley, K., Shaw, W. P., Greenstein, D. K., Lenroot, R. K., Giedd, J., Evans, A. C., Jul 2006. Mapping anatomical correlations across cerebral cortex (MACACC) using cortical thickness from MRI. Neuroimage 31 (3), 993–1003.

[59] Luders, E., Narr, K. L., Thompson, P. M., Rex, D. E., Jancke, L., Toga, A. W., Aug 2006. Hemispheric asymmetries in cortical thickness. Cereb Cortex 16 (8), 1232–8.

[60] Luders, E., Narr, K. L., Thompson, P. M., Rex, D. E., Woods, R. P., Deluca, H., Jancke, L., Toga, A. W., Apr 2006. Gender effects on cortical thickness and the influence of scaling. Hum Brain Mapp 27 (4), 314–24.

[61] Luders, E., Sánchez, F. J., Tosun, D., Shattuck, D. W., Gaser, C., Vilain, E., Toga, A. W., Aug 2012. Increased cortical thickness in male-to-female transsexualism. J Behav Brain Sci 2 (3), 357–362.

[62] Lüsebrink, F., Wollrab, A., Speck, O., Apr 2013. Cortical thickness determination of the human brain using high resolution 3T and 7T MRI data. Neuroimage 70, 122–31.

[63] Lyoo, I. K., Sung, Y. H., Dager, S. R., Friedman, S. D., Lee, J.-Y., Kim, S. J., Kim, N., Dunner, D. L., Renshaw, P. F., Feb 2006. Regional cerebral cortical thinning in bipolar disorder. Bipolar Disord 8 (1), 65–74.

[64] MacDonald, D., Kabani, N., Avis, D., Evans, A. C., Sep 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3), 340–56.

[65] Magnotta, V. A., Andreasen, N. C., Schultz, S. K., Harris, G., Cizadlo, T., Heckel, D., Nopoulos, P., Flaum, M., Mar 1999. Quantitative in vivo measurement of gyrification in the human brain: changes associated with aging. Cereb Cortex 9 (2), 151–60.

[66] Makris, N., Gasic, G. P., Kennedy, D. N., Hodge, S. M., Kaiser, J. R., Lee, M. J., Kim, B. W., Blood, A. J., Evins, A. E., Seidman, L. J., Iosifescu, D. V., Lee, S., Baxter, C., Perlis, R. H., Smoller, J. W., Fava, M., Breiter, H. C., Oct 2008. Cortical thickness abnormalities in cocaine addiction–a reflection of both drug use and a pre-existing disposition to drug abuse? Neuron 60 (1), 174–88.

[67] Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., Lancaster, J., Jun 1995. A probabilistic atlas of the human brain: theory and rationale for its development. the International Consortium for Brain Mapping (ICBM). Neuroimage 2 (2), 89–101.

[68] Neary, D., Snowden, J., Mann, D., Nov 2005. Frontotemporal dementia. Lancet Neurol 4 (11), 771–80.

[69] Nesvåg, R., Lawyer, G., Varnäs, K., Fjell, A. M., Walhovd, K. B., Frigessi, A., Jönsson, E. G., Agartz, I., Jan 2008. Regional thinning of the cerebral cortex in schizophrenia: effects of diagnosis, age and antipsychotic medication. Schizophr Res 98 (1-3), 16–28.

[70] Otsu, N., 1979. A threshold selection method from gray-level histograms. EEE Trans. Sys., Man., Cyber. 9 (1), 62–66.

[71] Peterson, B. S., Warner, V., Bansal, R., Zhu, H., Hao, X., Liu, J., Durkin, K., Adams, P. B., Wickramaratne, P., Weissman, M. M., Apr 2009. Cortical thinning in persons at increased familial risk for major depression. Proc Natl Acad Sci U S A 106 (15), 6273–8.

[72] Raine, A., Laufer, W. S., Yang, Y., Narr, K. L., Thompson, P., Toga, A. W., Oct 2011. Increased executive functioning, attention, and cortical thickness in white-collar criminals. Hum Brain Mapp.

[73] Raji, C. A., Ho, A. J., Parikshak, N. N., Becker, J. T., Lopez, O. L., Kuller, L. H., Hua, X., Leow, A. D., Toga, A. W., Thompson, P. M., Mar 2010. Brain structure and obesity. Hum Brain Mapp 31 (3), 353–64.

[74] Ramasamy, D. P., Benedict, R. H. B., Cox, J. L., Fritz, D., Abdelrahman, N., Hussein, S., Minagar, A., Dwyer, M. G., Zivadinov, R., Jul 2009. Extent of cerebellum, subcortical and cortical atrophy in patients with MS: a case-control study. J Neurol Sci 282 (1-2), 47–54.

[75] Reuter, M., Schmansky, N. J., Rosas, H. D., Fischl, B., Jul 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–18.

[76] Rosas, H. D., Hevelone, N. D., Zaleta, A. K., Greve, D. N., Salat, D. H., Fischl, B., Sep 2005. Regional cortical thinning in preclinical Huntington disease and its relationship to cognition. Neurology 65 (5), 745–7.

[77] Rosas, H. D., Liu, A. K., Hersch, S., Glessner, M., Ferrante, R. J., Salat, D. H., van der Kouwe, A., Jenkins, B. G., Dale, A. M., Fischl, B., Mar 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58 (5), 695–701.

[78] Salgado-Pineda, P., Delaveau, P., Falcon, C., Blin, O., 2006. Brain t1 intensity changes after levodopa administration in healthy subjects: a voxel-based morphometry study. British journal of clinical pharmacology 62 (5), 546–551.

[79] Scott, M. L. J., Bromiley, P. A., Thacker, N. A., Hutchinson, C. E., Jackson, A., Apr 2009. A fast, model-independent method for cerebral cortical thickness estimation using MRI. Med Image Anal 13 (2), 269–85.

[80] Ségonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., Fischl, B., Jul 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–75.

[81] Selemon, L. D., Rajkowska, G., Goldman-Rakic, P. S., Jan 2004. Evidence for progression in frontal cortical pathology in late-stage Huntington's disease. J Comp Neurol 468 (2), 190–204.

[82] Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., Leahy, R. M., May 2001. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13 (5), 856–76.

[83] Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., Giedd, J., Mar 2006. Intellectual ability and cortical development in children and adolescents. Nature 440 (7084), 676–9.

[84] Shin, Y.-W., Yoo, S. Y., Lee, J. K., Ha, T. H., Lee, K. J., Lee, J. M., Kim, I. Y., Kim, S. I., Kwon, J. S., Nov 2007. Cortical thinning in obsessive compulsive disorder. Hum Brain Mapp 28 (11), 1128–35.

[85] Shrout, P. E., Fleiss, J. L., Mar 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86 (2), 420–8.

[86] Sled, J. G., Zijdenbos, A. P., Evans, A. C., Feb 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17 (1), 87–97.

[87] Smith, S. M., Nov 2002. Fast robust automated brain extraction. Hum Brain Mapp 17 (3), 143–55.

[88] Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., Matthews, P. M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208–19.

[89] Sowell, E. R., Kan, E., Yoshii, J., Thompson, P. M., Bansal, R., Xu, D., Toga, A. W., Peterson, B. S., Jun 2008. Thinning of sensorimotor cortices in children with Tourette syndrome. Nat Neurosci 11 (6), 637–9.

[90] Talairach, J., Tournoux, P., 1988. Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional system—An approach to cerebral imaging. Thieme.

[91] Thompson, P. M., Lee, A. D., Dutton, R. A., Geaga, J. A., Hayashi, K. M., Eckert, M. A., Bellugi, U., Galaburda, A. M., Korenberg, J. R., Mills, D. L., Toga, A. W., Reiss, A. L., Apr 2005. Abnormal cortical complexity and thickness profiles mapped in Williams syndrome. J Neurosci 25 (16), 4146–58.

[92] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., Gee, J. C., Jun 2010. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29 (6), 1310–20.

[93] Tustison, N. J., Johnson, H. J., Rohlfing, T., Klein, A., Ghosh, S. S., Ibanez, L., Avants, B., 2013. Instrumentation bias in the use and evaluation of scientific software: Recommendations for reproducible practices in the computational sciences. Frontiers in Neuroscience 7 (162).

[94] Vachet, C., Hazlett, H. C., Niethammer, M., Oguz, I., Cates, J., Whitaker, R., Piven, J., Styner, M., February 2011. Group-wise automatic mesh-based analysis of cortical thickness. In: Benoit M. Dawant, D. R. H. (Ed.), SPIE Medical Imaging: Image Processing.

[95] Walhovd, K. B., Storsve, A. B., Westlye, L. T., Drevon, C. A., Fjell, A. M., Nov 2013. Blood markers of fatty acids and vitamin d, cardiovascular measures, body mass index, and physical activity relate to longitudinal cortical thinning in normal aging. Neurobiol Aging.
URL http://dx.doi.org/10.1016/j.neurobiolaging.2013.11.011

[96] Wang, D., Shi, L., Chu, W. C. W., Burwell, R. G., Cheng, J. C. Y., Ahuja, A. T., Jan 2012. Abnormal cerebral cortical thinning pattern in adolescent girls with idiopathic scoliosis. Neuroimage 59 (2), 935–42.

[97] Wang, H., Suh, J. W., Das, S. R., Pluta, J., Craige, C., Yushkevich, P. A., 2013. Multi-atlas segmentation with join label fusion. IEEE Trans Pattern Analysis and Machine Intelligence.

[98] Ward, B. D., 1999. Intracranial segmentation. Tech. rep., Medical College of Wisconsin,

http://afni.nimh.nih.gov/pub/dist/doc/3dIntracranial.pdf.

[99] Wei, G., Zhang, Y., Jiang, T., Luo, J., 2011. Increased cortical thickness in sports experts: a comparison of diving players with the controls. PLoS One 6 (2), e17112.

[100] Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Siuciak, J. A., Soares, H., Toga, A. W., Trojanowski, J. Q., , A. D. N. I., Feb 2012. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. Alzheimers Dement 8 (1 Suppl), S1–68.

[101] Wilkinson, G. N., Rogers, C. E., 1973. Symbolic description of factorial models for analysis of variance. Journal of the Royal Statistical Society. Series C (Applied Statistics) 22 (3), 392–399.

[102] Worsley, K. J., Chen, J.-I., Lerch, J., Evans, A. C., May 2005. Comparing functional connectivity via thresholding correlations and singular value decomposition. Philos Trans R Soc Lond B Biol Sci 360 (1457), 913–20.

[103] Yamasue, H., Abe, O., Kasai, K., Suga, M., Iwanami, A., Yamada, H., Tochigi, M., Ohtani, T., Rogers, M. A., Sasaki, T., et al., 2007. Human brain structural change related to acute single exposure to sarin. Annals of neurology 61 (1), 37–46.

[104] Yezzi, Jr, A. J., Prince, J. L., Oct 2003. An Eulerian PDE approach for computing tissue thickness. IEEE Trans Med Imaging 22 (10), 1332–9.

[105] Zeng, X., Staib, L. H., Schultz, R. T., Duncan, J. S., Oct 1999. Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation. IEEE Trans Med Imaging 18 (10), 927–37.

[106] Zhang, Y., Brady, M., Smith, S., Jan 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20 (1), 45–57.