NeuroImage xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements☆

Nicholas J. Tustison [a,*], 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,d]

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

ARTICLE INFO

Article history:
Accepted 15 May 2014
Available online xxxx

Keywords:
Advanced normalization tool
Age prediction
MRI
Gender prediction
Open science
Scientific reproducibility

ABSTRACT

Many studies of the human brain have explored the relationship between cortical thickness and cognition, phenotype, or disease. Due to the subjectivity and time requirements in manual measurement of cortical thickness, scientists have relied on robust software tools for automation which facilitate the testing and refinement of neuroscientific hypotheses. The most widely used tool for cortical thickness studies is the publicly available, surface-based FreeSurfer package. Critical to the adoption of such tools is a demonstration of their reproducibility, validity, and the documentation of specific implementations that are robust across large, diverse imaging datasets. 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 1205 images from four open data sets (IXI, MMRR, NKI, and OASIS), with parcellation based on 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 an ability to make statistical inferences, we further tested the neurobiological validity of these approaches by evaluating thickness-based prediction of age and gender. ANTs is shown to have a higher predictive performance than FreeSurfer for both of these measures. In promotion of open science, we make all of our scripts, data, and results publicly available which complements the use of open image data sets and the open source availability of the proposed ANTs cortical thickness pipeline.

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. Such research has yielded insight concerning cortical variability and its developmental correlates including those associated with normal aging (Walhovd et al., 2013) and gender differences (Luders et al., 2006). Conditional abnormalities from Alzheimer's disease and frontotemporal dementia (Dickerson et al., 2009; Du et al., 2007) to Parkinson's (Jubault et al., 2011) and Huntington's disease (Rosas et al., 2005) also demonstrate sensitivity to cortical thickness assessments. Additional explorations have included such topics of interest as autism (Chung et al., 2005), athletic ability (Wei et al., 2011), male-to-female transsexuality (Luders et al., 2012), obesity (Raji et al., 2010), and Tetris-playing ability in female adolescents (Haier et al., 2009). Although these findings are subject to debate and interpretation (Gernsbacher, 2007), the availability of quantitative computational methods for extracting cortical thickness measures 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 (Clarkson et al., 2011; Scott et al., 2009). Representative of the former is the FreeSurfer cortical modeling software package (Dale et al., 1999; Fischl and Dale, 2000; Fischl et al., 1999, 2002, 2004) which owes its popularity to public availability, excellent documentation, good performance, and integration with other toolkits, such as the extensive FMRIB software library (Smith et al., 2004). Similar to other surface-based cortical thickness estimation approaches (e.g., Davatzikos and Bryan, 1996; Kim et al., 2005;

☆ Partial support was provided by the Defense Health Program through the U.S. Army Medical Research Acquisition Activity, Grant Number W81XWH-09-2-0055.
* Corresponding author at: PO Box 801339, Charlottesville, VA 22908, USA.
E-mail address: ntustison@virginia.edu (N.J. Tustison).

http://dx.doi.org/10.1016/j.neuroimage.2014.05.044
1053-8119/

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

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	

```
# 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.

MacDonald et al., 2000; Magnotta et al., 1999), 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 Jones et al. (2000) 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 Yezzi and Prince (2003), a Eulerian partial differential equation approach was proposed to facilitate the computation of correspondence paths. Extending the surface-based work of MacDonald et al. (2000), the hybrid approach of Kim et al. (2005) uses the discrete Laplacian field to deform the white matter surface mesh towards the outer cortical surface. Other volumetric algorithms employ coupled level sets (Zeng et al., 1999), model-free intelligent search strategies either normal to the gray–white matter interface (Scott et al., 2009), or using a min–max rule (Vachet et al., 2011). Most relevant to this work is the DiReCT (Diffeomorphic Registration-based Cortical Thickness) algorithm proposed in Das et al. (2009) where generated diffeomorphic mappings between the gray/white matter and exterior cortical surfaces are used to propagate thickness values through the cortical gray matter.

The general lack of availability of published algorithms (Kovacevic, 2006) (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 (Clarkson et al., 2011) compared FreeSurfer (a surface-based method) with two volumetric methods, viz., Das et al. (2009) and Jones et al. (2000). Whereas the entire FreeSurfer processing pipeline has been made publicly available, refined by the original authors and other contributors, and described in

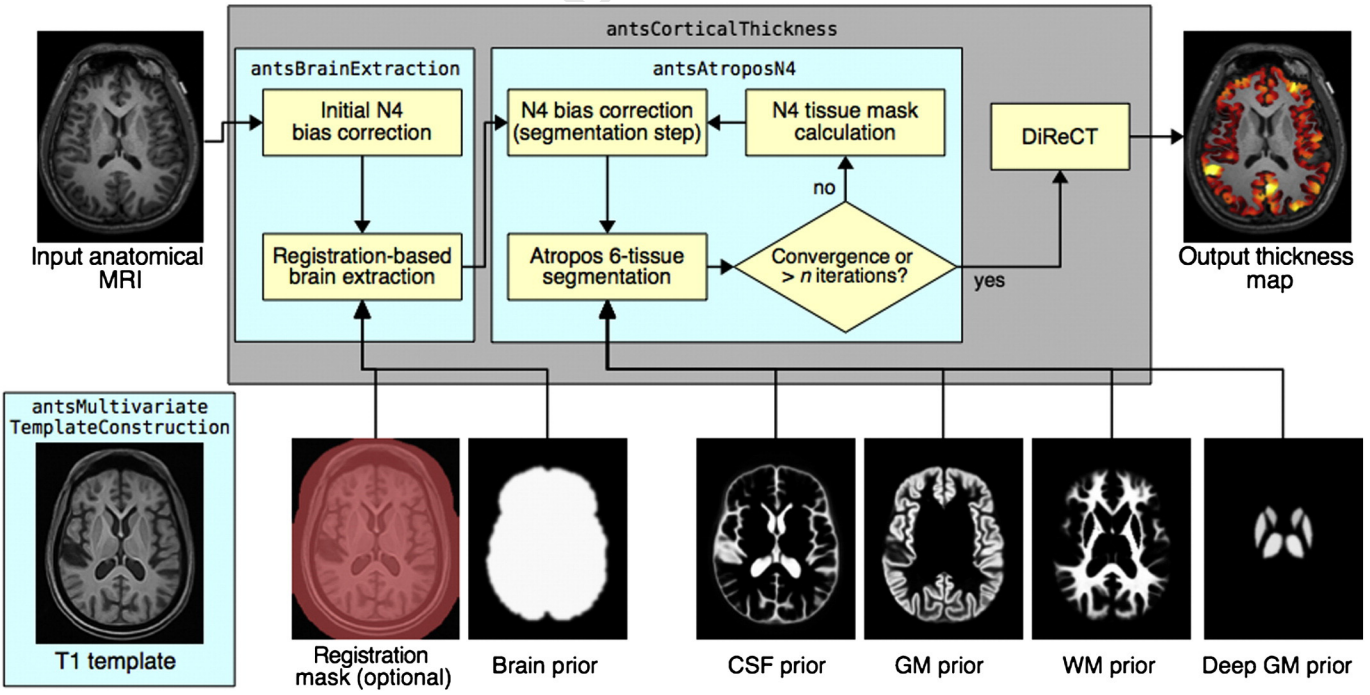


Fig. 1. Illustration of the main components of the ANTs processing workflow containing all elements for determining cortical thickness. We also included the domain of operations for the selected scripts. Not shown are the probability maps for the brain stem and cerebellum priors. All template-based prior probability maps are generated prior to pipeline processing of each individual subject.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

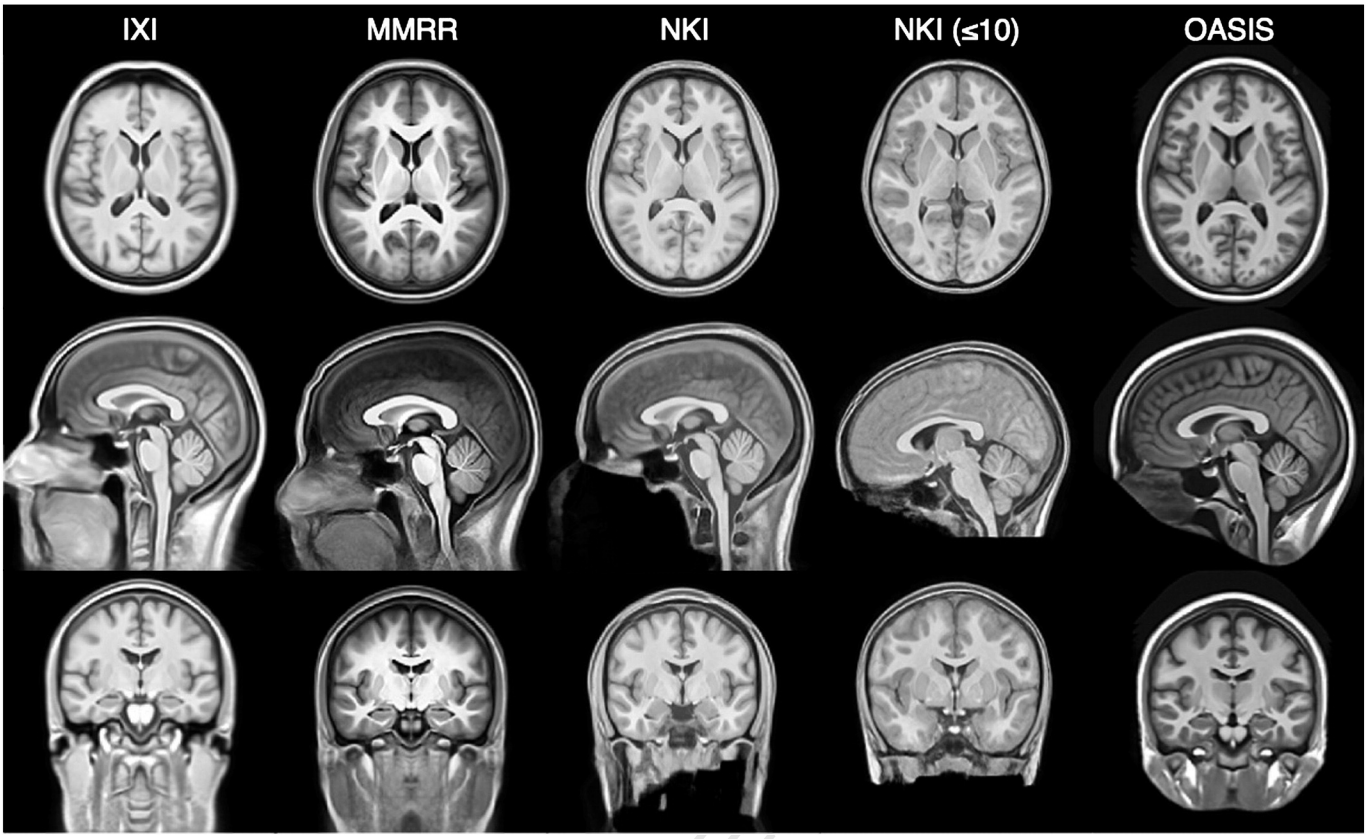


Fig. 2. Population-specific templates for each of the four public data sets used for cortical thickness estimation. The benefit of using such population-specific templates is obvious when one sees the variability in acquisition and data preparation (e.g., defacing protocols).

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 Tustison et al., 2013 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 (Klein et al., 2010).

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 (Tustison et al., 2010), brain extraction (Avants et al., 2010a), *n*-tissue segmentation (Avants et al., 2011b), template construction (Avants et al., 2010b), and image normalization (Avants et al., 2011a). More importantly, we provide explicit coordination among these components within a set of well-documented shell scripts 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., Clarkson et al., 2011), 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 the determination of measurement quality. Although FreeSurfer validation has included histological (Rosas et al., 2002) and image-drawn (Kuperberg et al., 2003) 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 or discussion of quality assurance. Alternatively, without ground truth, other forms of evidence can be adduced (e.g., Bouix et al., 2007) in making comparative inferences. In this work we use demographic-based assessments (based on well-studied relationships between cortical thickness and age/gender) to show that ANTs outperforms FreeSurfer-based thickness estimation for these data in terms of prediction.

Methods and materials

Public data resources

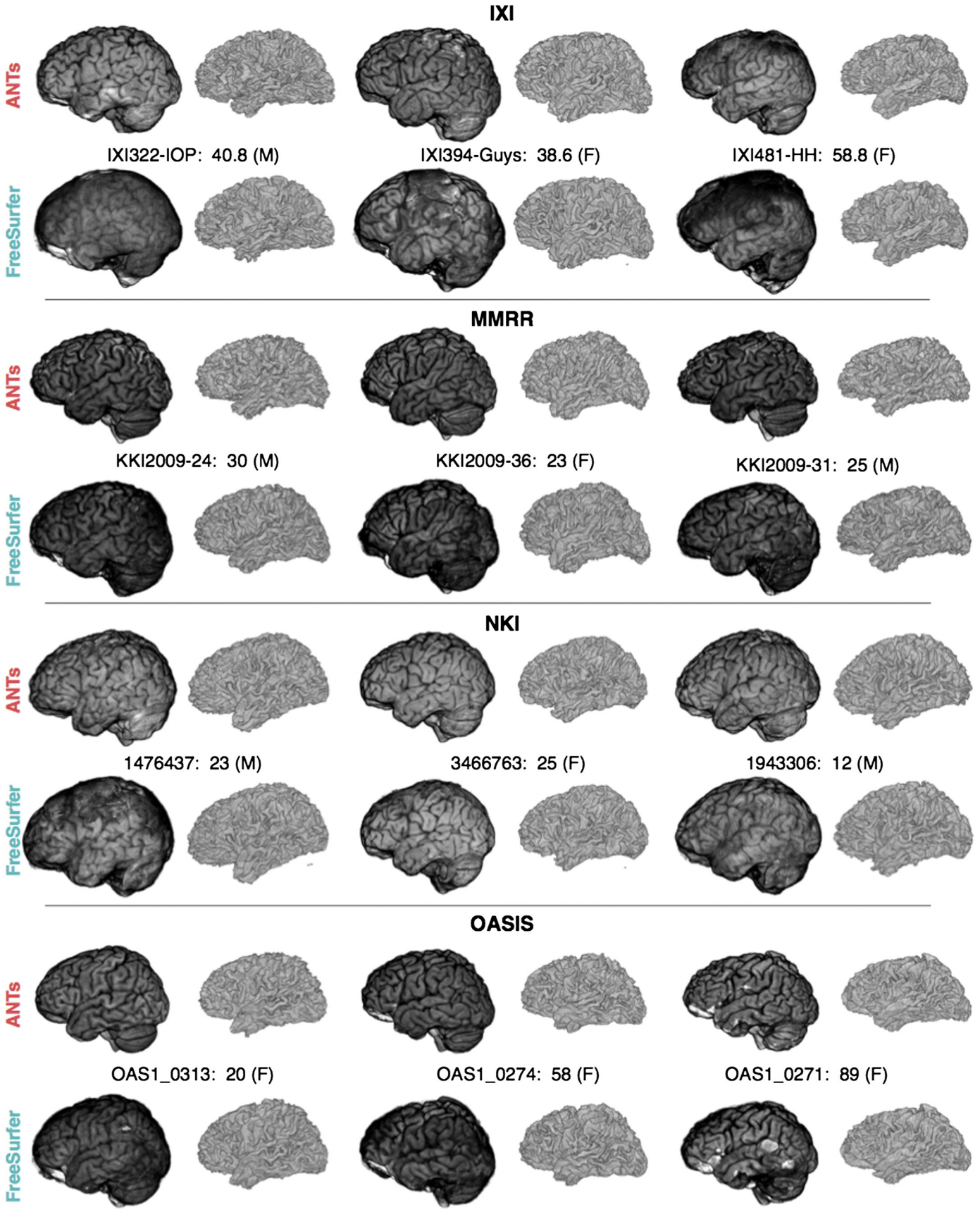
A comparative evaluation between FreeSurfer and ANTs was run on four publicly available data sets: IXI, MMRR, NKI, and OASIS. In addition to these data, we used a subset of the MindBoggle-101 data labeled using the Desikan–Killiany–Tourville (DKT) protocol (Klein and Tourville, 2012) to define the regions of interest (ROI) in the analysis. This latter data set was not included in the thickness analysis. All five data sets are described below.

Public data for thickness estimation evaluation

Diverse and publicly available data sets collected from multiple sites with a mixture of 3 T and 1.5 T T1-weighted brain images were analyzed using the ANTs and FreeSurfer pipelines. 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, field strength, and the broad variation in extracerebral tissue. Failure can occur in initial

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

brain extraction, segmentation, registration, or bias correction, any of which can lead to an inaccurate cortical thickness measurement. In total, we processed 1205 T1-weighted images from four different public data sets to obtain cortical thickness values for both cortical thickness analysis software. Below we describe the four data sets.



Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

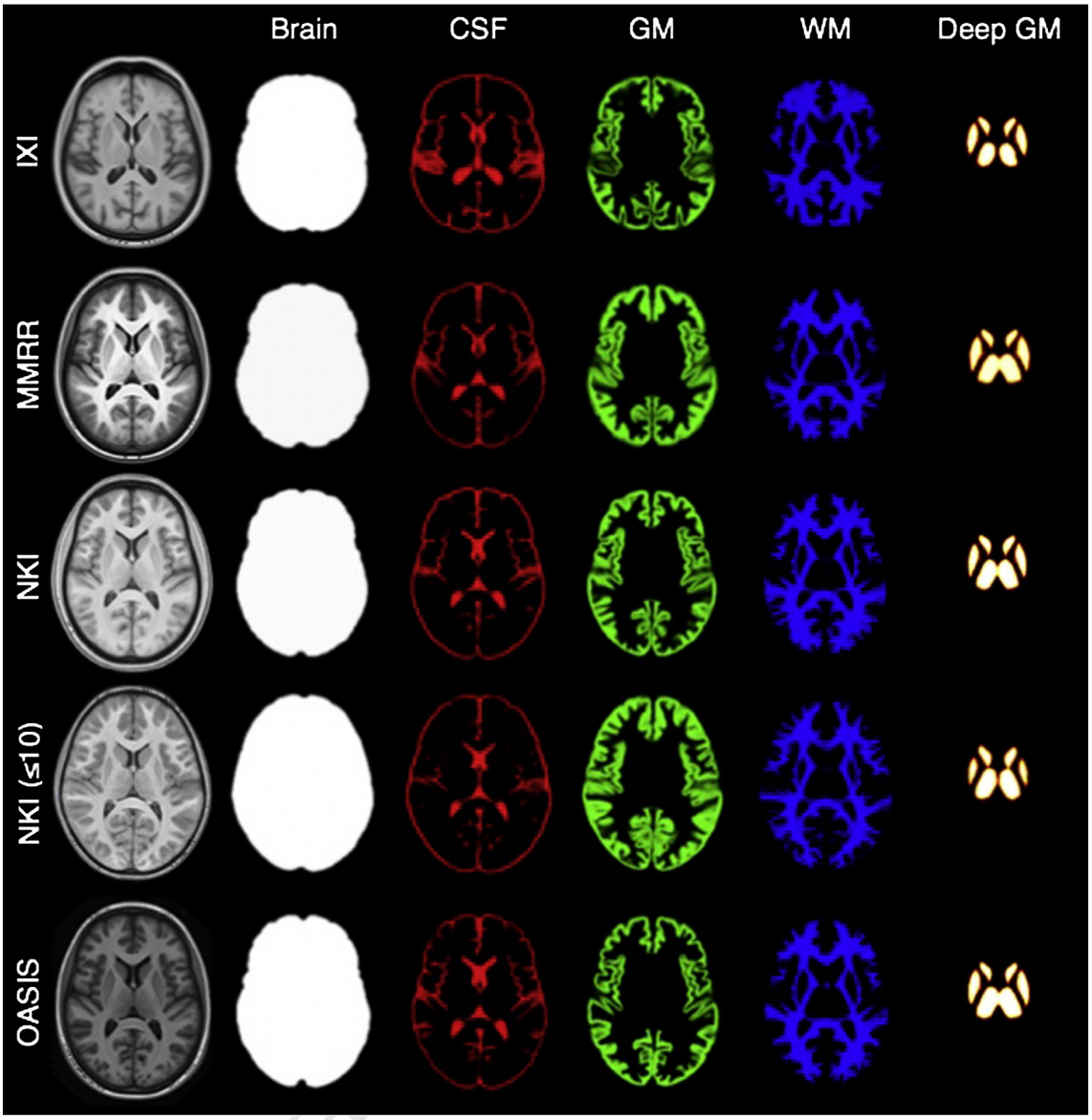


Fig. 4. Axial slices from each of the five T1 templates including the corresponding probability masks used for brain extraction and brain tissue segmentation. Not shown are the prior probability maps for brain stem and cerebellum regions.

IXI. Initially, we processed 581 T1-weighted images from the IXI 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.

MMRR. The Multi-Modal MRI Reproducibility Resource (MMRR) data set, was originally described in Landman et al. (2011) consisting of 21 subjects (10 females, 11 males) and features a rich set of modalities, as well as repeated scans.

NKI. In support of open science, the 1000 Functional Connectomes Project 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).

OASIS. The initial Open Access Series of Imaging Studies (OASIS) data set consisted of 433 T1-weighted images. We processed all of these data, 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.

Fig. 3. 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.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

MindBoggle-101 data for ROI definitions

In Klein and Tourville (2012) the authors proposed the DKT cortical labeling protocol—a modification of the popular Desikan–Killiany protocol (Desikan et al., 2006) to improve cortical labeling consistency and to improve FreeSurfer's cortical classification of 31 cortical regions per hemisphere, listed in Table 1. 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 Klein and Tourville (2012) 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.

ANTs volume-based cortical thickness estimation pipeline

The ANTs cortical thickness estimation workflow is illustrated in Fig. 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. Although other preprocessing components are possible (e.g., noise reduction as in Smith, 1996), the major steps constituting the ANTs pipeline are limited to those enumerated above.

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".

Anatomical template construction

Certain preprocessing steps, such as brain extraction and segmentation, rely on templates and corresponding spatial priors. In addition, normalizing images to a standard coordinate system reduces intersubject variability in population studies. Various approaches exist for determining an optimal template, such as the selection of a preexisting template based on a single individual (e.g., the Talairach atlas; Talairach and Tournoux, 1988) or a publicly available average of multiple individuals (e.g., the MNI template by Collins et al., 1994 or the ICBM template by Mazziotta et al., 1995), or an average template constructed from the individuals under study. The work of Avants et al. (2010b) explicitly models the geometric component of the normalized space during optimization to produce such mean templates. Coupling the intrinsic symmetry of SyN pairwise registration (Avants et al., 2011a) 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 five templates used in this evaluation study are represented in Fig. 2.

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 (Sled et al., 1998), as part of their standard protocol (Boyes et al., 2008).

In Tustison et al. (2010) 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 Fig. 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 the algorithm of Otsu (1979). 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).

Brain extraction

Brain extraction using ANTs combines template building, high-performance brain image registration, and Atropos segmentation with topological refinements. An optimal template (Avants et al., 2010b), 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 Fig. 2). 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 voxels. 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.

In previous work (Avants et al., 2010a) we compared an earlier version of our extraction method with publicly available brain extraction algorithms, including AFNI's 3dIntracranial (Ward, 1999), FSL's BET2 (Smith, 2002), FreeSurfer's mri_watershed (Ségonne et al., 2004), and BrainSuite (Dogdas et al., 2005). Our hybrid registration/segmentation approach performed with an accuracy comparable to FreeSurfer and a parameter-tuned version of BrainSuite. Fig. 3 presents a visual comparison of results derived with the current ANTs brain extraction method and results obtained using FreeSurfer.

Atropos six-tissue segmentation

In Avants et al. (2011b) 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., Ashburner and Friston, 2005; Zhang et al., 2001). 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.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

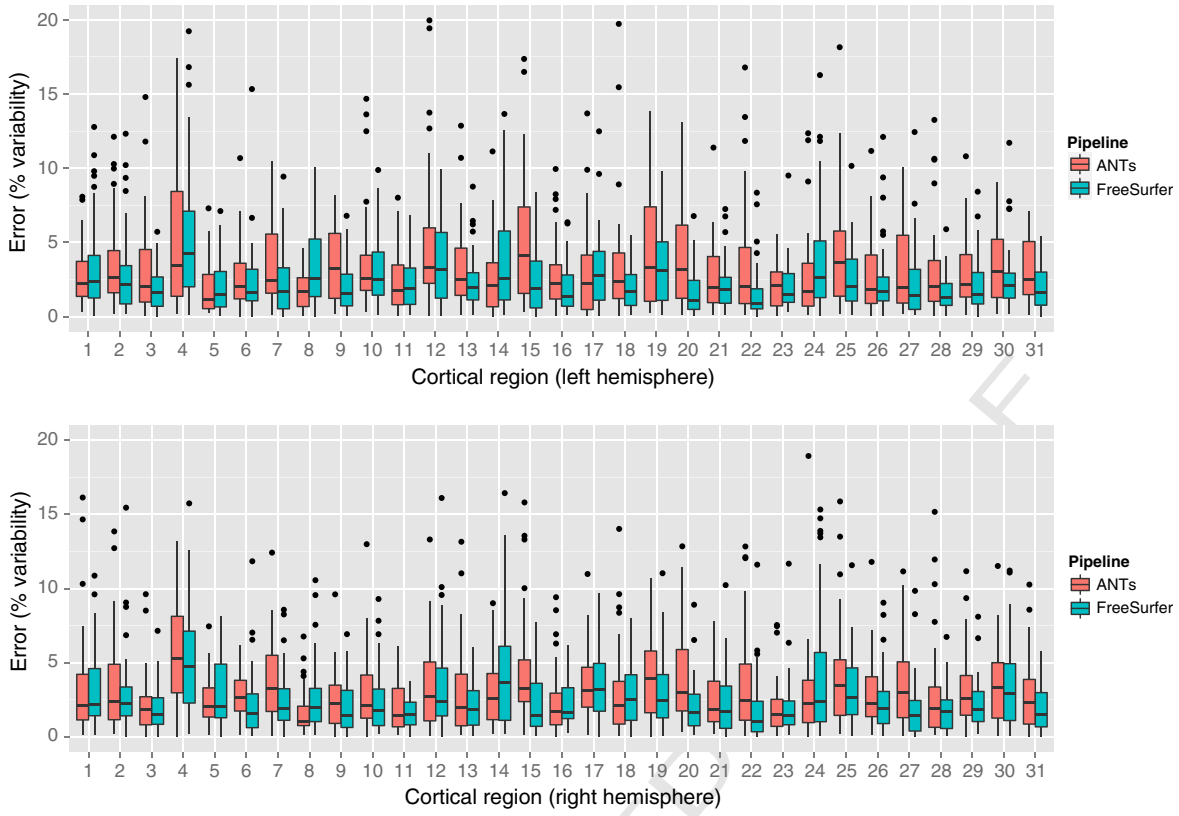


Fig. 5. Percent error variability for both ANTs and FreeSurfer pipelines over the left and right hemispheres of both the MMRR and OASIS data subsets within the 62 regions defined by the Desikan–Killiany–Tourville atlas. Both methods demonstrate good repeatability qualities.

Additional capabilities include handling of multivariate data, partial volume modeling (Shattuck et al., 2001), a memory-minimization mode, label propagation, a plug-and-play architecture for incorporation of novel likelihood models which include 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 ⇌ Atropos iterations. 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 ⇌ Atropos iteration from

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

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 is used in N4 to weight more heavily the influence of 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 $P(x|y) \propto P(y|x)^{1-\alpha} P(x)^{\alpha}$ where α is a user-selected parameter which weights the tradeoff 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).

To generate the priors for each T1 template, we used the multi-atlas label fusion (MALF) algorithm of Wang et al. (2013) in conjunction with a labeled subset of the OASIS data set.[1] First, we normalized the labeled OASIS subset to the template. We then performed MALF on the template using the normalized labeled data as input. This resulted in a labeled, parcellated template consisting of 100+ labels defining the different brain regions. We then condensed this template-specific labeling to the six needed for our analysis, viz., cerebrospinal fluid (CSF), gray matter (GM), white matter (WM), deep gray matter, brain stem, and cerebellum. For example, all cortical regions were assigned a single label representing the gray matter.

These binary masks were then smoothed using Gaussian convolution with a one voxel-width kernel. 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. This new CSF prior was then subtracted from each of the other five prior probability images and limited to the probability range of [0,1]. The prior probabilities for the five templates used in this evaluation are given in Fig. 4.

[1] These data were originally acquired by the first and last authors to aid in the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling. The data was released under the Creative Commons Attribution-NonCommercial license. Labelings were provided by Neuromorphometrics, Inc. (http://Neuromorphometrics.com/) under academic subscription.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

Table 2
Mean repeatability error over all regions.

	MMRR	OASIS
ANTs	3.2%	3.3%
FreeSurfer	2.5%	2.8%

DiReCT cortical thickness estimation

DiReCT was introduced in Das et al. (2009) and was made available in ANTs as the program KellySlater. Since then several improvements have been made and incorporated into the program KellyKapowski.[2] The more recent implementation has made numerous advances including multi-threading, written in rigorous ITK coding style, and has been made publicly available through ANTs, complete with a unique command line interface design developed specifically for ANTs tools.

Processing miscellany

Given the documented variability in FreeSurfer results with version and operating system (Gronenschild et al., 2012) (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 (http://www.uvacse.virginia.edu) 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). The brain extraction and segmentation results from both pipelines were visually inspected to screen for major problems. No manual changes were made for any component of either pipeline and no change was made to the settings of either processing pipeline.

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,[3] and age and how its performance 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.

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 Jovicich et al. (2013), 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 is given in Fig. 5. Mean values are given in Table 2. 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 Shrout and Fleiss, 1979) 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.

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 or for the many studies looking at brain morphology, an indirect method (or set of methods) is required for determining the quality of thickness estimation.

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.[4] 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) (Breiman, 2001), for estimating age from both ANTs and FreeSurfer thickness values in the testing data.

The formula (in the notation of Wilkinson and Rogers, 1973) for the linear model is

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

where $T(DKT_i)$ is the average thickness value in region DKT_i and $VOLUME$ is total intracranial volume. Similarly, the random forest model was specified as a combination of all terms using the randomForest 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):

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

Due to the limited range in ages across data sets, we restricted training and testing to the age range [20,80]. The resulting distributions are illustrated in Fig. 6. In addition to a combined assessment, we also

[2] 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.

[3] We recognize the distinction often made between "sex" and "gender" (cf http://www.who.int/gender/whatisgender/en/). As the demographic 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.

[4] 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.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

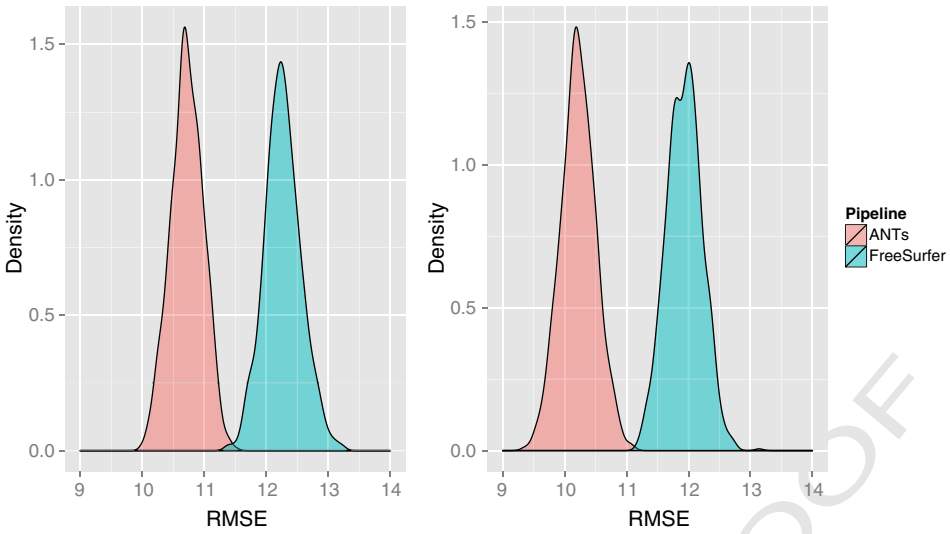


Fig. 6. Age prediction RMSE distributions of linear (left) and random forest (right) models for the ANTs- and FreeSurfer-derived thickness values over the combined four cohorts. For both prediction models ANTs RMSE error is lower.

perform separate model prediction for each of the three larger data sets (i.e., IXI, NKI, and OASIS).

ANTs-based RMSE values were lower for both models and each of the four different subset comparisons except for the random forest model constructed from the OASIS data set. All mean RMSE values are provided in Table 3.

To further elucidate the regional differences in predictive power specifically in the random forest model, we provide variable importance plots for both pipelines using the mean decrease in accuracy measure in Fig. 7. During random forest model construction (specifically the out-of-bag error calculation stage), the decrease in prediction accuracy with the omission of a single feature or variable is tracked and averaged. Thus, those features which have the greatest decrease in mean accuracy are considered to be the most discriminative. It should be noted that correlative effects are not considered in the rankings.

Gender prediction assessment

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

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

where $T(DKT_i)$ is the average thickness value in region DKT_i and $VOLUME$ is the total intracranial volume. We then characterized performance using a ROC curve for both methods (see Fig. 9) where we averaged over 1000 permutations. The mean area under the curve (AUC) for both methods was also quantified with values of $\text{ANTs}_{AUC} = 0.87$ and $\text{FreeSurfer}_{AUC} = 0.83$. (See Fig. 8.)

Table 3
Mean RMSE for age prediction in years.

	Linear model	Random forest
ANTs (combined)	10.7	10.2
FreeSurfer (combined)	12.3	11.9
ANTs (IXI)	9.3	8.6
FreeSurfer (IXI)	12.3	11.7
ANTs (NKI)	NA[a]	10.9
FreeSurfer (NKI)	NA[a]	13.3
ANTs (OASIS)	15.0	12.4
FreeSurfer (OASIS)	15.0	11.4

[a] Fitting error.

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 and 20 h per subject for both pipelines for the entire cortical thickness estimation procedure although ANTs processing, on average, took slightly longer. Averaged over all cohorts, ANTs required 15.7 ± 2.0 h per subject and FreeSurfer required 14.1 ± 2.9 h per subject.

The propagation of the DKT labels to each subject using label fusion as described earlier was performed in parallel and took anywhere between 40 and 80 h per subject for 16 serial image registrations and the application of the joint label fusion algorithm (Wang et al., 2013). 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 $1200 \times 20 = 24{,}000$ registrations were performed. The antsMalfLabeling.sh script mentioned earlier parallelizes the registration component which decreases the time for parallel computation platforms.

Discussion

In the absence of ground truth, we used repeatability and prediction of demographic variables to compare the ANTs and FreeSurfer cortical thickness pipelines. 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.

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

Fig. 7. Regional importance random forest plots for (left) ANTs and (right) FreeSurfer using "MeanDecreaseAccuracy" ranking all model variables specified by Eq. (3).

framework. Pre-analysis confounds such as short-term alterations in cortical morphology due to the T1-weighted susceptibility to blood flow (Franklin et al., 2013; Salgado-Pineda et al., 2006; Yamasue et al., 2007) and MRI acquisition parameters such as field strength, site, resolution, scanner, longitudinal variation in scanner conditions, and pulse sequence (Han et al., 2006; Jovicich et al., 2013; Lüsebrink et al., 2013) 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.

ANTs and FreeSurfer cortical thickness mean reliability measurements are correlated across all regions (Pearson correlation = 0.44). Although our thickness reliability measurements represent the compound effect of registration, segmentation, anatomical labeling, and the thickness computation algorithm, this correlation suggests that these effects are non-random. That is, reliability measurements are influenced by characteristics intrinsic to the underlying neuroanatomy as represented in approximately one millimeter resolution volumetric T1-weighted MRI. Perhaps the least reliable region is the entorhinal cortex (region 4 in Fig. 5) which has a relatively small volume, is challenging to distinguish from surrounding structures (Price et al., 2010), and is also relatively thin. Spatial variation in segmentation accuracy is known to relate to a structure's volume and tissue characteristics and this has led to a body of research on both segmentation and acquisition protocols that are optimized for specific regions. Perhaps the most substantial work in MRI has focused on temporal lobe structures including the hippocampus. Both FreeSurfer and our own group have optimized protocols to address such concerns (http://www.hippocampalsubfields.com/). Given caveats associated with cost vs. benefit, our current results suggest that optimized protocols may be relevant for additional cortical regions.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

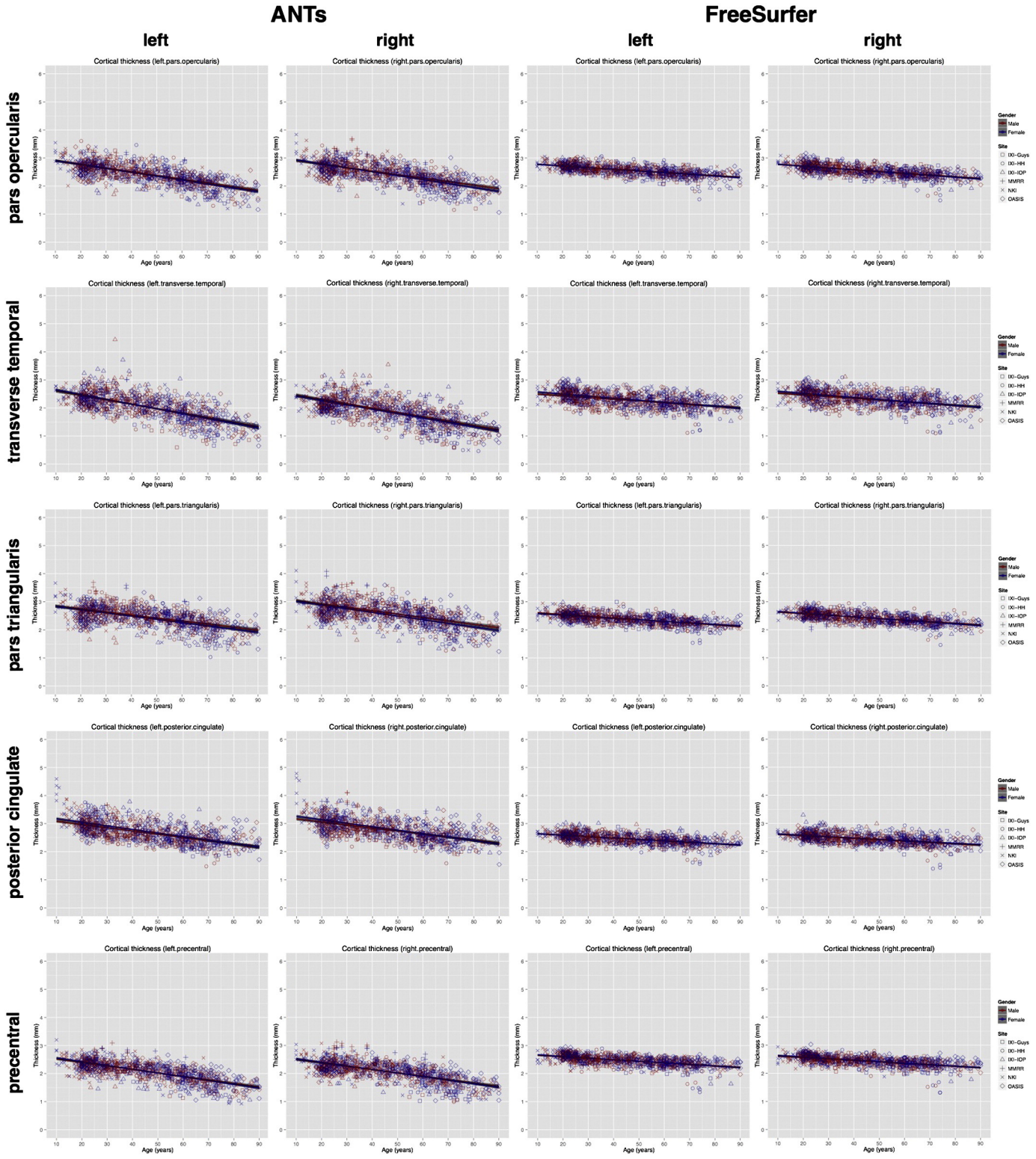


Fig. 8. Age vs. thickness plots for cortical regions that are most relevant in age prediction. These are the most discriminative regions across both methods as determined by random forest importance measurements (cf Fig. 7). Note that all regional plots for both ANTs and FreeSurfer are available online (see Appendix A).

Voxel/vertex-based analysis

One of the limitations of our evaluation was the limitation of comparative analysis to mean ROI thickness values defined by the 62 cortical regions of the DKT atlas. Quite common in the literature, however, are point-wise (vertex- or voxel-based) analyses (e.g., Chung et al., 2005). The ANTs pipeline described in this work is equally applicable to such studies. The only additional requirement is the specification of the normalization template. For this work we opted for the ROI analysis to avoid potential bias issues when navigating between surface and volume representations (Klein et al., 2010). Future work will certainly explore such analyses.

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

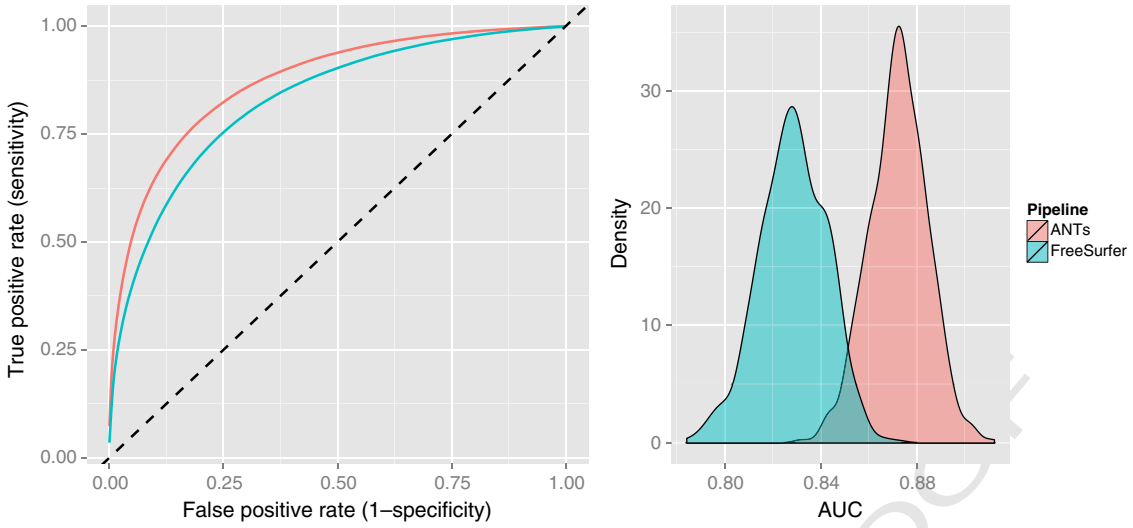


Fig. 9. Average ROC curve and corresponding AUC distributions for gender prediction using ANTs and FreeSurfer thickness values. Values were averaged for 1000 permutations resulting in mean values of $ANTs_{AUC} = 0.87$ and $FreeSurfer_{AUC} = 0.83$ ($p < 10^{-16}$).

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.

Previous research has used predictive modeling for comparing cortical thickness algorithms. For example, in Clarkson et al. (2011), classification of healthy, semantic dementia, and progressive non-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 (Neary et al., 2005) is not as straightforward as obtaining a subject's age or gender and regressing that against cortical thickness; the latter constitutes biological relationships that have been well-studied and reported in the literature.

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 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 makes comparison with previous age versus thickness assessments possible.

For example, in Hogstrom et al. (2013), 322 T1-weighted MRI of healthy adults with an age range of [20,85] were used, in part, to characterize the relationship between age and cortical thickness using FreeSurfer and a similar linear modeling approach. Significant findings for age were reported in the "precentral gyrus, medial parts of the superior frontal gyrus, DMPFC, and rostral middle frontal cortex." Based on the cortical parcellation provided by the DKT atlas, we also saw similar strong effects in the precentral gyrus (cf Fig. 7).

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 (Jovicich et al., 2013; Reuter et al., 2012). 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.

Computation time

Computation time for the registration and segmentation components of the ANTs pipeline are substantial but are not significantly worse than those of FreeSurfer. 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.

Table 4
Resources used in this work.

Packages	
ANTs	http://stnava.github.io/ANTs
FreeSurfer	http://surfer.nmr.mgh.harvard.edu
Available scripts and examples	
antsBrainExtraction.sh	https://github.com/ntustison/antsBrainExtractionExample
antsAtroposN4.sh	https://github.com/ntustison/antsAtroposN4Example
antsCorticalThickness.sh	https://github.com/ntustison/antsCorticalThicknessExample
antsMultivariateTemplateConstruction.sh	https://github.com/ntustison/TemplateBuildingExample
antsMalfLabeling.sh	https://github.com/ntustison/MalfLabelingExample
Analysis scripts	https://github.com/ntustison/KapowskiChronicles
Public data	
MindBoggle101	http://mindboggle.info/data.html
Cohort templates and priors	http://figshare.com/articles/ANTs_ANTsR_Brain_Templates/915436
IXI	http://biomedic.doc.ic.ac.uk/brain-development
MMRR	http://www.nitrc.org/projects/multimodal
NKI	http://fcon_1000.projects.nitrc.org
OASIS	http://www.oasis-brains.org
MICCAI 2012 Workshop on Multi-Atlas Labeling	https://masi.vuse.vanderbilt.edu/workshop2012/index.php

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.

Appendix A

Available resources are listed in Table 4 with their corresponding addresses. Examples and data for all scripts described in the manuscript are also available for download. This should enable interested researchers to duplicate the results in this work.

References

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851. http://dx.doi.org/10.1016/j.neuroimage.2005.02.018 (Jul).

Avants, B.B., Klein, A., Tustison, N.J., Woo, J., Gee, J.C., 2010a. Evaluation of an open-access, automated brain extraction method on multi-site multi-disorder data. Hum. Brain Mapp.

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.C., 2010b. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49 (3), 2457–2466. http://dx.doi.org/10.1016/j.neuroimage.2009.09.062 (Feb).

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011a. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54 (3), 2033–2044. http://dx.doi.org/10.1016/j.neuroimage.2010.09.025 (Feb).

Avants, B.B., Tustison, N.J., Wu, J., Cook, P.A., Gee, J.C., 2011b. An open source multivariate framework for *n*-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400. http://dx.doi.org/10.1007/s12021-011-9109-y (Dec).

Bouix, S., Martin-Fernandez, M., Ungar, L., Nakamura, M., Koo, M.S., McCarley, R.W., Shenton, M.E., 2007. On evaluating brain tissue classifiers without a ground truth. Neuroimage 36 (4), 1207–1224. http://dx.doi.org/10.1016/j.neuroimage.2007.04.031 (Jul).

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, 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. Neuroimage 39 (4), 1752–1762. http://dx.doi.org/10.1016/j.neuroimage.2007.10.026 (Feb).

Breiman, L., 2001. Random forests. Machine Learning, pp. 5–32.

Chung, M.K., Robbins, S.M., Dalton, K.M., Davidson, R.J., Alexander, A.L., Evans, A.C., 2005. Cortical thickness analysis in autism with heat kernel smoothing. Neuroimage 25 (4), 1256–1265. http://dx.doi.org/10.1016/j.neuroimage.2004.12.052 (May).

Clarkson, M.J., Cardoso, M.J., Ridgway, G.R., Modat, M., Leung, K.K., Rohrer, J.D., Fox, N.C., Ourselin, S., 2011. A comparison of voxel and surface based cortical thickness estimation methods. Neuroimage 57 (3), 856–865. http://dx.doi.org/10.1016/j.neuroimage.2011.05.053 (Aug).

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.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9 (2), 179–194. http://dx.doi.org/10.1006/nimg.1998.0395 (Feb).

Das, S.R., Avants, B.B., Grossman, M., Gee, J.C., 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–879. http://dx.doi.org/10.1016/j.neuroimage.2008.12.016 (Apr).

Davatzikos, C., Bryan, N., 1996. Using a deformable surface model to obtain a shape representation of the cortex. IEEE Trans. Med. Imaging 15 (6), 785–795. http://dx.doi.org/10.1109/42.544496.

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., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980. http://dx.doi.org/10.1016/j.neuroimage.2006.01.021 (Jul).

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., 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. http://dx.doi.org/10.1093/cercor/bhn113 (Mar).

Dogdas, B., Shattuck, D.W., Leahy, R.M., 2005. Segmentation of skull and scalp in 3-D human MRI using mathematical morphology. Hum. Brain Mapp. 26 (4), 273–285. http://dx.doi.org/10.1002/hbm.20159 (Dec).

Du, A.T., Schuff, N., Kramer, J.H., Rosen, H.J., Gorno-Tempini, M.L., Rankin, K., Miller, B.L., Weiner, M.W., 2007. Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130 (Pt 4), 1159–1166. http://dx.doi.org/10.1093/brain/awm016 (Apr).

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proc. Natl. Acad. Sci. U. S. A. 97 (20), 11050–11055. http://dx.doi.org/10.1073/pnas.200033797 (Sep).

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (1), 195–207. http://dx.doi.org/10.1006/nimg.1998.0396 (Jan).

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., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (2), 341–355 (Feb).

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., 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14 (3), 11–22 (Jan).

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 Struct. Funct. 218 (1), 97–104.

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

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. http://dx.doi.org/10.1371/journal.pone.0038234.

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. http://dx.doi.org/10.1186/1756-0500-2-174.

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., 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. http://dx.doi.org/10.1016/j.neuroimage.2006.02.051 (Aug).

Hogstrom, L.J., Westlye, L.T., Walhovd, K.B., Fjell, A.M., 2013. The structure of the cerebral cortex across adult life: age-related patterns of surface area, thickness, and gyrification. Cereb. Cortex 23 (11), 2521–2530. http://dx.doi.org/10.1093/cercor/bhs231 (Nov).

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

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, 2013. Brain morphometry reproducibility in multi-center 3 T MRI studies: a comparison of cross-sectional and longitudinal segmentations. Neuroimage. http://dx.doi.org/10.1016/j.neuroimage.2013.05.007 (May).

Jubault, T., Gagnon, J.F., Karama, S., Ptito, A., Lafontaine, A.L., Evans, A.C., Monchi, O., 2011. Patterns of cortical thickness and surface area in early Parkinson's disease. Neuroimage 55 (2), 462–467. http://dx.doi.org/10.1016/j.neuroimage.2010.12.043 (Mar).

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., 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), 210–221. http://dx.doi.org/10.1016/j.neuroimage.2005.03.036 (Aug).

Klein, A., Tourville, J., 2012. 101 labeled brain images and a consistent human cortical labeling protocol. Front. Neurosci. 6, 171. http://dx.doi.org/10.3389/fnins.2012.00171.

Klein, A., Ghosh, S.S., Avants, B., Yeo, B.T.T., Fischl, B., Ardekani, B., Gee, J.C., Mann, J.J., Parsey, R.V., 2010. Evaluation of volume-based and surface-based brain image registration methods. Neuroimage 51 (1), 214–220. http://dx.doi.org/10.1016/j.neuroimage.2010.01.091 (May).

Kovacevic, J., 2006. From the editor-in-chief. IEEE Trans. Image Process. 15 (12).

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., 2003. Regionally localized thinning of the cerebral cortex in schizophrenia. Arch. Gen. Psychiatry 60 (9), 878–888. http://dx.doi.org/10.1001/archpsyc.60.9.878 (Sep).

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., 2011. Multi-parametric neuroimaging reproducibility: a 3-T resource study. Neuroimage 54 (4), 2854–2866. http://dx.doi.org/10.1016/j.neuroimage.2010.11.047 (Feb).

Luders, E., Narr, K.L., Thompson, P.M., Rex, D.E., Woods, R.P., Deluca, H., Jancke, L., Toga, A.W., 2006. Gender effects on cortical thickness and the influence of scaling. Hum. Brain Mapp. 27 (4), 314–324. http://dx.doi.org/10.1002/hbm.20187 (Apr).

Luders, E., Sánchez, F.J., Tosun, D., Shattuck, D.W., Gaser, C., Vilain, E., Toga, A.W., 2012. Increased cortical thickness in male-to-female transsexualism. J. Behav. Brain Sci. 2 (3), 357–362. http://dx.doi.org/10.4236/jbbs.2012.23040 (Aug).

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044

Lüsebrink, F., Wollrab, A., Speck, O., 2013. Cortical thickness determination of the human brain using high resolution 3 T and 7 T MRI data. Neuroimage 70, 122–131. http://dx.doi.org/10.1016/j.neuroimage.2012.12.016 (Apr).

MacDonald, D., Kabani, N., Avis, D., Evans, A.C., 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3), 340–356. http://dx.doi.org/10.1006/nimg.1999.0534 (Sep).

Magnotta, V.A., Andreasen, N.C., Schultz, S.K., Harris, G., Cizadlo, T., Heckel, D., Nopoulos, P. , Flaum, M., 1999. Quantitative in vivo measurement of gyrification in the human brain: changes associated with aging. Cereb. Cortex 9 (2), 151–160 (Mar).

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 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 (Jun).

Neary, D., Snowden, J., Mann, D., 2005. Frontotemporal dementia. Lancet Neurol. 4 (11), 771–780. http://dx.doi.org/10.1016/S1474-4422(05)70223-4 (Nov).

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9 (1), 62–66.

Price, C.C., Wood, M.F., Leonard, C.M., Towler, S., Ward, J., Montijo, H., Kellison, I., Bowers, D., Monk, T., Newcomer, J.C., Schmalfuss, I., 2010. Entorhinal cortex volume in older adults: reliability and validity considerations for three published measurement protocols. J. Int. Neuropsychol. Soc. 16 (5), 846–855. http://dx.doi.org/10.1017/S135561771000072X (Sep).

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., 2010. Brain structure and obesity. Hum. Brain Mapp. 31 (3), 353–364. http://dx.doi.org/10.1002/hbm.20870 (Mar).

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418. http://dx.doi.org/10.1016/j.neuroimage.2012.02.084 (Jul).

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., 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58 (5), 695–701 (Mar).

Rosas, H.D., Hevelone, N.D., Zaleta, A.K., Greve, D.N., Salat, D.H., Fischl, B., 2005. Regional cortical thinning in preclinical Huntington disease and its relationship to cognition. Neurology 65 (5), 745–747. http://dx.doi.org/10.1212/01.wnl.0000174432.87383.87 (Sep).

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. Br. J. Clin. Pharmacol. 62 (5), 546–551.

Scott, M.L.J., Bromiley, P.A., Thacker, N.A., Hutchinson, C.E., Jackson, A., 2009. A fast, model-independent method for cerebral cortical thickness estimation using MRI. Med. Image Anal. 13 (2), 269–285. http://dx.doi.org/10.1016/j.media.2008.10.006 (Apr).

Ségonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. Neuroimage 22 (3), 1060–1075. http://dx.doi.org/10.1016/j.neuroimage.2004.03.032 (Jul).

Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. Neuroimage 13 (5), 856–876. http://dx.doi.org/10.1006/nimg.2000.0730 (May).

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428 (Mar).

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97. http://dx.doi.org/10.1109/42.668698 (Feb).

Smith, S.M., 1996. Flexible filter neighbourhood designation. Proc. 13th Int. Conf. on Pattern Recognition, vol. 1, pp. 206–212.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155. http://dx.doi.org/10.1002/hbm.10062 (Nov).

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–S219. http://dx.doi.org/10.1016/j.neuroimage.2004.07.051.

Talairach, J., Tournoux, P., 1988. Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system—an approach to cerebral imaging. Thieme. Q13

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. http://dx.doi.org/10.1109/TMI.2010.2046908 (Jun).

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. Front. Neurosci. 7 (162). http://dx.doi.org/10.3389/fnins.2013.00162.

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

Walhovd, K.B., Storsve, A.B., Westlye, L.T., Drevon, C.A., Fjell, A.M., 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. http://dx.doi.org/10.1016/j.neurobiolaging.2013.11.011 (Nov).

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 Anal. Mach. Intell. Q14

Ward, B.D., 1999. Intracranial segmentation. Technical Report. Medical College of Wisconsin (http://afni.nimh.nih.gov/pub/dist/doc/3dIntracranial.pdf).

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. http://dx.doi.org/10.1371/journal.pone.0017112.

Wilkinson, G.N., Rogers, C.E., 1973. Symbolic description of factorial models for analysis of variance. J. R. Stat. Soc.: Ser. C: Appl. Stat. 22 (3), 392–399.

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. Ann. Neurol. 61 (1), 37–46.

Yezzi Jr., A.J., Prince, J.L., 2003. An Eulerian PDE approach for computing tissue thickness. IEEE Trans. Med. Imaging 22 (10), 1332–1339. http://dx.doi.org/10.1109/TMI.2003.817775 (Oct).

Zeng, X., Staib, L.H., Schultz, R.T., Duncan, J.S., 1999. Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation. IEEE Trans. Med. Imaging 18 (10), 927–937. http://dx.doi.org/10.1109/42.811276 (Oct).

Zhang, Y., Brady, M., Smith, S., 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. http://dx.doi.org/10.1109/42.906424 (Jan).

Please cite this article as: Tustison, N.J., et al., Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.05.044