NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 24 (2005) 163–173

Cortical thickness analysis examined through power analysis and a population simulation

Jason P. Lerch and Alan C. Evans*

McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada H3A 2B4

Received 23 March 2004; revised 27 June 2004; accepted 12 July 2004
Available online 18 November 2004

We have previously developed a procedure for measuring the thickness of cerebral cortex over the whole brain using 3-D MRI data and a fully automated surface-extraction (ASP) algorithm. This paper examines the precision of this algorithm, its optimal performance parameters, and the sensitivity of the method to subtle, focal changes in cortical thickness.

The precision of cortical thickness measurements was studied using a simulated population study and single subject reproducibility metrics. Cortical thickness was shown to be a reliable method, reaching a sensitivity (probability of a true-positive) of 0.93. Six different cortical thickness metrics were compared. The simplest and most precise method measures the distance between corresponding vertices from the white matter to the gray matter surface. Given two groups of 25 subjects, a 0.6-mm (15%) change in thickness can be recovered after blurring with a 3-D Gaussian kernel (full-width half max = 30 mm). Smoothing across the 2-D surface manifold also improves precision; in this experiment, the optimal kernel size was 30 mm.

Introduction

The measurement of cortical thickness has long been of interest to the neurosciences, starting with the early reconstructions of Brodmann (1909) and von Economo and Koskinas (1925). Recent advances in image processing and image acquisition has allowed for the automatic extraction of cortical thickness from MRI (Fischl and Dale, 2000; MacDonald, 1997; MacDonald et al., 2000). This paper investigates and summarizes current methodology and evaluates the power and sensitivity of the different techniques.

The study of the morphometry of the cerebral cortex at the macroscopic level visible in current MRI provides the neurosciences with an opportunity to investigate both normal and abnormal change. Most such investigations use a combination of semiautomatic techniques, usually focusing on the manual delineation of structures of interest, followed by statistical comparisons of volumes (cf. Pruessner et al., 2001). This approach, while clearly quite capable of providing important information about the population under investigation, has several disadvantages. It is very labor intensive, it suffers from intra- and interrater reliability issues, and most importantly, it restricts the analysis to predetermined regions of interest.

Several fully automated approaches have also been developed; the most widely used of these is voxel-based morphometry (VBM) (Ashburner and Friston, 2000). At its most generic, VBM is the comparison of voxels in a series of linear models. Most methods (cf. Ashburner and Friston, 2000; Baron et al., 2001; Paus et al., 1999) employ a standard set of image processing steps involving linear registration, tissue classification, and creation of "voxel density" maps representing tissue concentration in a local neighborhood. The usual end result is an image that contains regions that have significantly increasing or decreasing signal that correlates with some independent neurobiological parameter. This latter parameter may be just a categorical difference between two groups, for example, separated by disease status or gender, or more generally, will be a continuous variable, such as age or behavioral performance, in which case a regression of image signal against that variable is plotted at each voxel (Paus et al., 1999; Wright et al., 1995).

Cortical thickness analysis is similar to VBM, albeit the analysis is performed at the nodes of a three-dimensional polygonal mesh rather than on a 3-D voxel grid, but it has the advantage of providing a direct quantitative index of cortical morphology. The metric captures the distance between the white matter surface and the gray CSF intersection according to some geometric definition; the output is a scalar value measured in millimeters. The regression slope at each vertex across the cortex in a statistical analysis is meaningful: not only can one determine that cortical thickness is significantly different between groups, but one can also measure that difference. This naturally leads to the ability to define clinical as well as statistical significance.

The use of cortical thickness analysis in MRI studies is relatively new, with only a small number of studies published on the methodology (Fischl and Dale, 2000; Jones et al., 2000; Kabani

* Corresponding author. McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Webster 2B, 3801 University Street, Montreal, QC, Canada H3A 2B4. Fax: +1 514 398 8948.
E-mail address: alan@bic.mni.mcgill.ca (A.C. Evans).

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2004.07.045

et al., 2001; MacDonald, 1997; MacDonald et al., 2000; Meyer et al., 1996; Miller et al., 2000; Tosun et al., 2001; Yezzi and Prince, 2003; Zeng et al., 1999) and even fewer on normal or abnormal populations (Fischl and Dale, 2000; MacDonald et al., 2000; Rosas et al., 2002). This is due to the difficult nature of extracting the inner and outer surfaces of the cerebral cortex at the limited resolution provided by today's MRI machines (usually 1 mm^3), where the fine details of sulcal anatomy are often obscured by the partial volume effect. Moreover, manual delineation of cortical thickness is very difficult (whether from MRI or postmortem samples) due to the necessity of creating a correct cut or slice plane perpendicular to the surfaces.

Defining cortical thickness, even when models of the inner and outer surfaces are present, is not trivial. Cortical thickness is a distance metric but there are multiple ways of defining corresponding points on the two surfaces between which that distance is to be measured. Moreover, the distance need not be measured in a straight line but can be the result of a more complicated equation, such as fluid flow lines.

This paper examines the power of cortical thickness as an analysis tool; it compares the various definitions of cortical thickness proposed in the literature; the effect of different size blurring kernels; and analyzes the effect of correcting for multiple comparisons. These studies were performed using a simulated population study where the true difference between the two groups is artificially induced and therefore known. Furthermore, repeat scans of a single subject will be used to examine the variability inherent in the different cortical thickness metrics and make a first attempt at defining the power of the method.

Rather than addressing accuracy we focus on the question of precision. The distinction between the two is subtle but crucial:

Accuracy: The ability of a metric to capture the correct distance between the pial and white matter surfaces, as defined by anatomical criteria and validated through manual measurements or accurate MR simulations.

Precision: The ability of a metric to provide reproducible results from repeated estimations and thereby differentiate between two measures known to be different.

A metric can therefore be declared most accurate through the comparison of automated and manual measurements, or through directly simulating a cortical sheet with a known thickness and validating each metric against such a construct. Each of these methods has to overcome significant challenges. Manual measurements of cortical thickness are tremendously difficult to undertake, being highly dependent on a perfectly perpendicular cutting angle. Moreover, even using the exact same postmortem slice, individual raters can easily differ by over 0.5 mm at any one location due to the blurred cortical boundary at the white matter surface (von Economo and Koskinas, 1925). Furthermore, the traditional thickness measurements derived from postmortem slices are dependent on a straight-line measurement of cortical thickness, as these measurements are always carried out in two dimensions. Accurate validation of MR measurements of cortical thickness thus requires a three-dimensional reconstruction of high-resolution postmortem data. The alternative of validation through construction of a cortical sheet with known thickness is very attractive but difficult to use in comparing thickness metrics. The reason is that the construction of such a cortex would be dependent on a preexisting definition of cortical thickness and would therefore be biased towards that metric from the beginning. Furthermore, accurately simulating MRI from polygonal models of the cortex has to first address the issue of correctly incorporating partial volume into the tissue model. We plan to address the question of accuracy both through the use of a simulator as well as high-resolution postmortem reconstructions; that, however, is the subject of future work. This paper examines the precision of cortical thickness analysis through the use of repeated acquisitions of the same subject as well as a population simulation.

Methodology

Measuring cortical thickness

Measuring cortical thickness is a complex process involving multiple image processing steps. The native data, usually consisting of a T1 MRI per subject but optionally includes any number of modalities. These one or more images of the brain parenchyma are used to provide an anatomic label for each voxel (typically this means classification into gray matter, white matter, CSF, and nonbrain classes). Prior to this classification step, intensity and spatial normalization must be performed. Intensity correction for nonuniformity is obtained using the N3 algorithm (Sled et al., 1998); spatial normalization is done to the ICBM 152 average using a nine parameter linear registration (Collins et al., 1994). If there is more than one image per subject, any additional MRIs are registered to the first MRI using mutual information registration (Collins et al., 1994).

Each subject's brain is classified into white matter, gray matter, CSF, and background using all available imaging modalities and a classifier trained by stereotaxic space probability maps (Kollokian, 1996; Zijdenbos et al., 2002). These probability maps were created from 305 classified samples; 1000 points per tissue class were randomly chosen from areas having a greater than 90% chance of being of the correct tissue type in that location in stereotaxic space. Prior to classification, the training tag points are pruned for each individual subject to remove any outliers.

The inner and outer cortical surfaces are then extracted using the automated surface-extraction (ASP) algorithm (MacDonald et al., 2000). The essence of ASP is the creation of simple (non-self-intersecting) surfaces with spherical topologies using deformable models. The classified volume is taken as input, and the process begins with the deformation towards the white matter surface. Along with the image information, $T_{\text{boundary-dist}}$, several model terms are used to constrain the fit, and self-intersection is explicitly prohibited. The model terms are: T_{stretch}, constraining distances between neighboring vertices; T_{bend}, constraining deviation from model shape; and $T_{\text{self-proximity}}$, constraining the proximity of pairs of nonadjacent polygons. The gray matter surface is obtained using the same constrains as listed above along with $T_{\text{surface–surface}}$, preventing the two surface from coming within a certain distance of each other, and $T_{\text{vertex–vertex}}$, which penalizes corresponding vertices as they deviate from an ideal distance. This last constraint allows for sulcal penetration of the gray matter surface even when the sulcus in question has been obscured due to partial volume blurring of the CSF space.

The creation of the two surfaces then allows for the measuring of cortical thickness using various distance metrics. They are summarized in Table 1 and described in more detail below.

The first method is t_{link}. It is conceptually very simple, measuring the distance between linked nodes on the inner and outer surface. The correspondence between such nodes is created

Table 1
Description of cortical thickness metrics used in this study

Name	Description	Citation
t_{link}	Distance between linked nodes	MacDonald et al. (2000)
t_{near}	Distance to nearest node	MacDonald et al. (2000)
t_{normal}	Distance along surface normal	MacDonald et al. (2000)
$t_{\text{layered-normal}}$	Distance along iteratively computed normal	NA
$t_{\text{average-near}}$	Distance to nearest node computed twice, averaged	Fischl and Dale (2000)
t_{laplace}	Distance solved using Laplace's equation	Jones et al. (2000)

by the expansion of the outer surface from the inner surface, each polyhedron having the same topology and number of vertices. This method is inherently very robust: the model constraints that govern the expansion will guarantee low variability, minimizing large errors and outliers. However, there is no guarantee that the linked method will produce a distance measure corresponding to what an anatomist would chose.

The t_{near} method performs a simple search across the opposite surface and picks the vertex that is the shortest (Euclidian) distance away. While intuitive, this method has the potential for gross errors, such as jumping across gyri, as there is no guarantee that the nearest point is the anatomically most sensible one. The t_{normal} constrains the point that can be found to lie along the intersection of the surface normal. The $t_{\text{layered-normal}}$ creates a series of nested surfaces first by a process of weighted averages of the inner and outer surfaces evaluated between corresponding nodes on each surface. The surface normal is computed at each of these nested surfaces and then averaged, thereby producing a constraint for finding the corresponding point on the opposite surface that is less prone to producing outliers than the simple surface normal. The $t_{\text{average-near}}$, first published in Fischl and Dale (2000), computes t_{near} twice, once from the outside to the inside surface and once from the inside to the outside surface. These two values are then averaged to produce a thickness value at that node.

The last method is t_{laplace}, first published in Jones et al. (2000) and reimplemented locally. Two boundaries, the white matter volume and the extracortical volume as defined by the two surfaces extracted using the ASP algorithm, are defined and fixed. Laplace's equation, shown in Eq. (1), is then iteratively solved across the entire volume using the Jacobi method. Iterations continue until the change across each iteration becomes smaller than a preset threshold. Gradients are then computed using two point differences and the gradient vectors normalized to produce tangent vector fields. Streamlines are computed at every voxel in the cortical volume by integrating towards each of the boundaries using Euler's method, and the two path-lengths added together to produce a thickness value at that point.

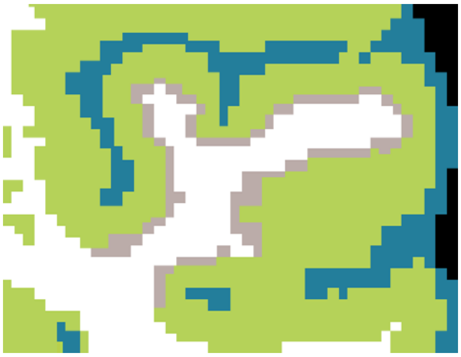

Fig. 2. An illustration of the rSTG thinning procedure. Green is gray matter, white is white matter, and gray is the gray matter that was removed in the "patient" population.

$$\Delta^2 \Psi = \frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} \qquad (1)$$

Eq. (1): Laplace's equation

The final step before analysis is blurring the thickness data. A surface-based diffusion smoothing kernel is used, which generalizes Gaussian kernel smoothing and makes it applicable to any arbitrary curved surface (Chung et al., 2002). It has to be remembered that this blurring kernel used on the surface has a different meaning from the standard volumetric kernels since surface curvature is followed as illustrated in Fig. 1.

The arguments for blurring are fourfold:

1. By the central limit theorem, smoothing has the effect of rendering the data more normally distributed, thereby increasing the validity of statistical tests.
2. It reduces the impact of imperfect alignment between cortices by replacing individual vertex values with neighborhood averages.

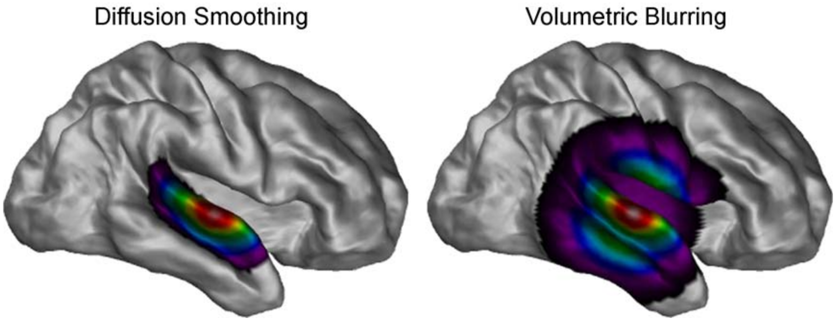

Fig. 1. An illustration of the difference between the geometry preserving diffusion smoothing blurring over a 2-D surface manifold and the more commonly employed 3-D volumetric blurring kernels. The FWHM was set at 30 mm in both cases. One can see how anatomically disparate areas such as the inferior motor and sensorimotor areas are influenced by the volumetric kernel but not by diffusion smoothing.

Table 2
Epidemiological statistics used in this study

		Test	
		+	–
True state	+	a	b
	–	c	d

Sensitivity: a / a + b (probability of a true-positive). Specificity: d / c + d (probability of a true-negative). True-positive: positive test where true state is positive. True-negative: negative test where true state is negative. Sensitivity: probability of a true-positive. Specificity: probability of a true-negative.

3. It reduces noise in the measurement of cortical thickness. The fact that the average cortex is only a few voxels thick leads to some variability in thickness measures due to the inadequate MRI sampling. By averaging neighboring vertices in the diffusion smoothing operation, this noise is reduced.
4. Since blurring increases the interdependence of the neighboring vertices, it also reduces the number of comparisons to be controlled for using random field theory (see Statistical analysis section).

These improvements in signal-to-noise and statistical normality are of course obtained at the cost of a degradation in image resolution in the classical image analysis trade-off. The choice of optimal blurring kernel width is discussed below.

Statistical analysis

Once the thickness maps have been generated and optionally smoothed for each subject, statistical tests can be performed. A linear model is applied separately at each vertex t: $Y(t) = \mathbf{X}\beta(t) + \varepsilon(t)$, where $Y(t)$ is the measure of cortical thickness, $\mathbf{X}$ is the matrix of explanatory variables, β represents the slope to be estimated for each explanatory variables, and $\varepsilon(t)$ is the normally distributed error. A series of statistical tests, such as a t, F, or adjusted R^2 values, can be applied. The regression slope, β, can also be plotted at every vertex. The ability to derive meaning out of the regression slope is one of the key strengths of cortical thickness analysis since that slope can be expressed as millimeters change. Accurate estimation and interpretation of the slope will be influenced by the kernel used since blurring causes thickness to be estimated across areas of cortex rather than individual vertices. There is thus potential for underestimation of local change should the kernel size be too large.

The challenge, also faced by VBM and functional imaging techniques, is to correct for the multiple comparisons that are undertaken in the analysis. For the cortical thickness analysis experiments described here, the number of nodes in the cortical mesh resulting from ASP is 40962. The most common method to control for multiple testing is to adjust the required significance threshold such that type I error is controlled for across the entire brain. The by now default method for such correction uses random field theory thresholding (Worsley et al., 1992, 1996), which takes the smoothness of the data into account in determining the number of resels, thereby reducing the effective number of tests to be controlled for. The resulting threshold states that in an area where the null hypothesis is true, the chance of rejecting one or more of the tests is less than or equal to α, the preset level of confidence. This type of control is quite stringent, providing the benefit that vertices where the null hypothesis is rejected are highly likely to be true-positives but also leaves a high likelihood of false-negatives. Implementation of random field thresholding on the surface is made more complex by the nonisotropic nature of the images. The solution is to estimate the effective FWHM (eFWHM) (determined through the normalized residuals of the fitted model) along the edge of each vertex and to warp (in a statistical sense) the coordinates of each vertex so that the eFWHM is approximately constant (Worsley et al., 1999).

Analyzing the variance

In order to quantify the normal variance expected for cortical thickness estimation, 19 different T1 MRIs with 1 mm isotropic sampling were acquired from the same subject over a short period of time (Holmes et al., 1998). The cortical thickness pipeline as described above was run on each acquisition. All the different metrics were used, each blurred with different sized kernels. Means and standard deviations were computed both at every vertex along with each subject's mean thickness. A normalized standard deviation map was produced by dividing the standard deviation by the mean at each vertex. The procedure was then repeated for 25 normal subjects taken from the ICBM database (Mazziotta et al., 2001) in order to capture the variance inherent in a normal population.

Power calculations were performed on both sets of data. The standard deviation of cortical thickness was modeled at every vertex to answer the questions:

1. What n is needed in order to recapture a change of x millimeters?
2. Given two equal groups of n subjects each, what change can be recaptured?

Table 3
Means, standard deviations, and power analyses for cortical thickness and metrics

	19 scans of same subject					25 normal subjects				
	Mean	SD	Nrm. SD	δ (%)	n	Mean	SD	Nrm. SD	δ (%)	n
t_{near}	2.46	0.21	0.09	21	20	2.38	0.36	0.15	38	52
t_{normal}	4.39	0.50	0.11	28	31	4.30	0.70	0.16	41	59
$t_{\text{layered-normal}}$	3.99	0.28	0.07	18	15	3.95	0.47	0.12	30	34
$t_{\text{average-near}}$	2.53	0.18	0.07	18	15	2.48	0.31	0.13	31	37
t_{laplace}	3.71	0.27	0.07	18	15	3.63	0.44	0.12	30	35
t_{link}	3.93	0.22	0.06	14	11	3.88	0.35	0.09	22	21

Where δ is the minimum percentage change that can be recaptured when n = 25, and n is the change that can be recaptured when δ = 25%. Nrm. Std is the normalized standard deviation. All comparisons made using 30 mm blurring kernel, P = 0.05 after correction for multiple comparisons using random field theory (t = 4.67).

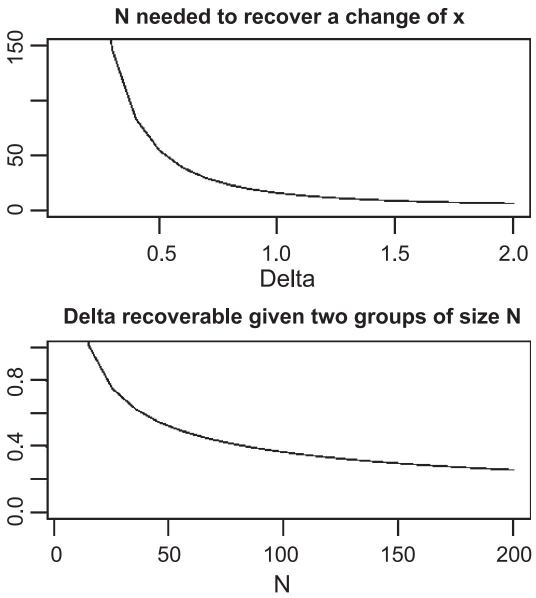

Fig. 3. Normalized standard deviation across different thickness metrics and blurring kernels. The two leftmost columns show the normalized standard deviation across different metrics, first in 19 scans taken of the same subject, second across 25 normal subjects. The last column shows the change in normalized standard deviation across blurring kernels, using the t_{link} method and 19 scans of the same subject.

For both cases, the significance level (type I error probability) was set to 0.05 and the power (1—type II error probability) at 0.95 multiple comparisons corrected for with random field theory. The interpretation of the results is dependent on the blurring kernel: in the unblurred data, we are looking at the sample size needed to recapture a change of a certain magnitude, where that change is isolated from any neighbors. The addition of blurring, however, modifies the value at each vertex to reflect a weighted neighborhood average.

Population simulation

The goal of many imaging studies in neurology is to assess morphometric differences between two populations. In order to evaluate the utility of cortical thickness analysis in such a scenario, an artificial “patient” population was created through induced thinning of the cortex. Fifty subjects from the ICBM (Mazziotta et al., 2001) database were taken. All the MRIs were corrected for nonuniformity artifacts, linearly registered into stereotaxic space, and classified into their component tissue types, all as described above. Twenty-five of these subjects were randomly chosen and designated as patients. In this group, the MRIs were segmented using ANIMAL (Collins et al., 1995) and the right superior temporal gyrus (rSTG) arbitrarily chosen for thinning. Thinning was induced through a six neighbor dilation of the white matter into the rSTG as defined by ANIMAL (see Fig. 2). The cortical surfaces were then fit on all subjects and cortical thickness measured with each of the available metrics. Statistical analysis was performed at every vertex to assess if the induced change can be recaptured. Prior to the induced thinning, the two populations were compared and no statistically significant differences were found ($P > 0.3$).

In order to evaluate the performance of the different metrics and blurring kernels, the standard epidemiological terms true-positives, false-positives, true-negatives, and false-negatives were defined (see Table 2). The definition of truth for the purposes of this simulation was based on the probability map of the rSTG. Since the rSTG is the site of the induced thinning, it should also be the area exhibiting significant results. The rSTG was defined individually for each subject, however, and is not perfectly aligned in stereotaxic space. Hence, a probability map was created for the rSTG in which each vertex value represented the proportion of subjects for whom that vertex was labeled as rSTG. Truth, for the purposes of the simulation experiment, was thus defined as a statistically significant vertices that intersect the rSTG probability map.

Results

Variability

The standard deviation of cortical thickness was measured at each point on the cortex in repeated scans across one subject as well as across a normal population. The results across metrics are summarized in Table 3.

Variability differs across different cortical thickness metrics

The t_{normal} has the highest standard deviation, $t_{\mathrm{average\text{-}near}}$ the lowest. Due to the different definitions for the thickness metrics, the mean thickness is quite variable across the different methods, ranging from a high of 4.39 mm in t_{normal} to a low of 2.46 mm in t_{near} (which by definition must have the lowest value). After normalizing to account for these differences by dividing the standard deviation with mean thickness, t_{normal} once again has the worst performance, t_{link} the best. The same pattern emerges whether these metrics are investigated across repeat scans of one subject or across 25 different young normals (see Table 3).

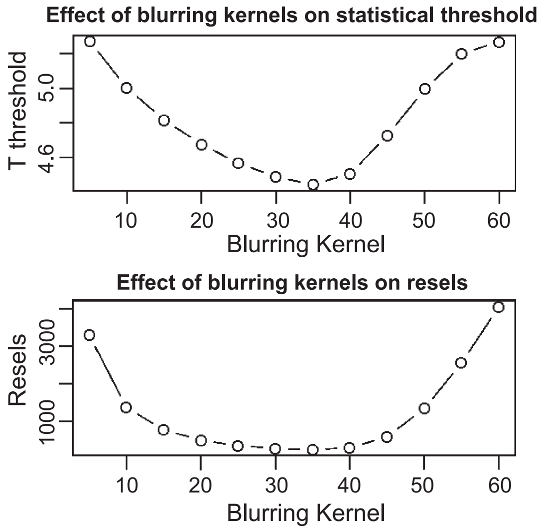

Fig. 4. Graph of normalized standard deviation across blurring kernels, using the t_{link} and t_{laplace} methods computed in 19 scans of the same subject. There is a similar pattern to the one seen in the figure—a decrease up to a kernel size of 40 mm followed by increasing standard deviations, once again indicating that the optimal blurring kernel for minimizing variance is in the 35- to 40-mm range.

Variability is not uniform across the cortex

Variance in thickness differs dependent on location in the cortex, being highest along the superior aspects of the central sulcus, lowest in the prefrontal cortex; see Fig. 3, which shows the standard deviation of cortical thickness at each vertex in both 25 normal subject as well as the 19 repeated scans of a single subject. Standard deviation is marginally related to cortical thickness at any vertex (R^2 = 0.0005). Normalizing the standard deviation increases the effect (R^2 = 0.06). Thinner cortical areas are thus more variable than their thicker counterparts, though thickness itself only explains a small part of the heterogeneity of variability. While the overall magnitude of the deviation varies across metrics, the spatial pattern is similar. As seen in the right hand column of Fig. 3, the pattern of variability remains stable across blurring kernels, even as the overall variability decreases with increased smoothing. The exception to this rule is that at high blurring kernels the representation of noncortical areas (such as the brain-stem cut) where cortical thickness measurements are meaningless and therefore highly variable begins to influence the standard deviation of their cortical neighbors.

Variability declines with increased blurring—up to a point

Normalized standard deviation (SD / mean) declines with blurring up to a 40- to 50-mm kernel, after which it increases again, as shown in Fig. 4. This increase appears to be due to an increased spatial representation of the noncortical areas. Normalized standard deviation in cortical areas spatially removed from noncortical regions continues to decline with increased blurring.

Power calculations

Power calculations for the different metrics are given in Table 3 and illustrated for the t_{link} metric in Fig. 5. Given two groups of more than 100 subjects each, a change of 0.29 mm can be recovered. Conversely, given two small groups of 20 subjects each, a change of 1 mm would reach statistical significance. These numbers assume random field theory corrections for multiple

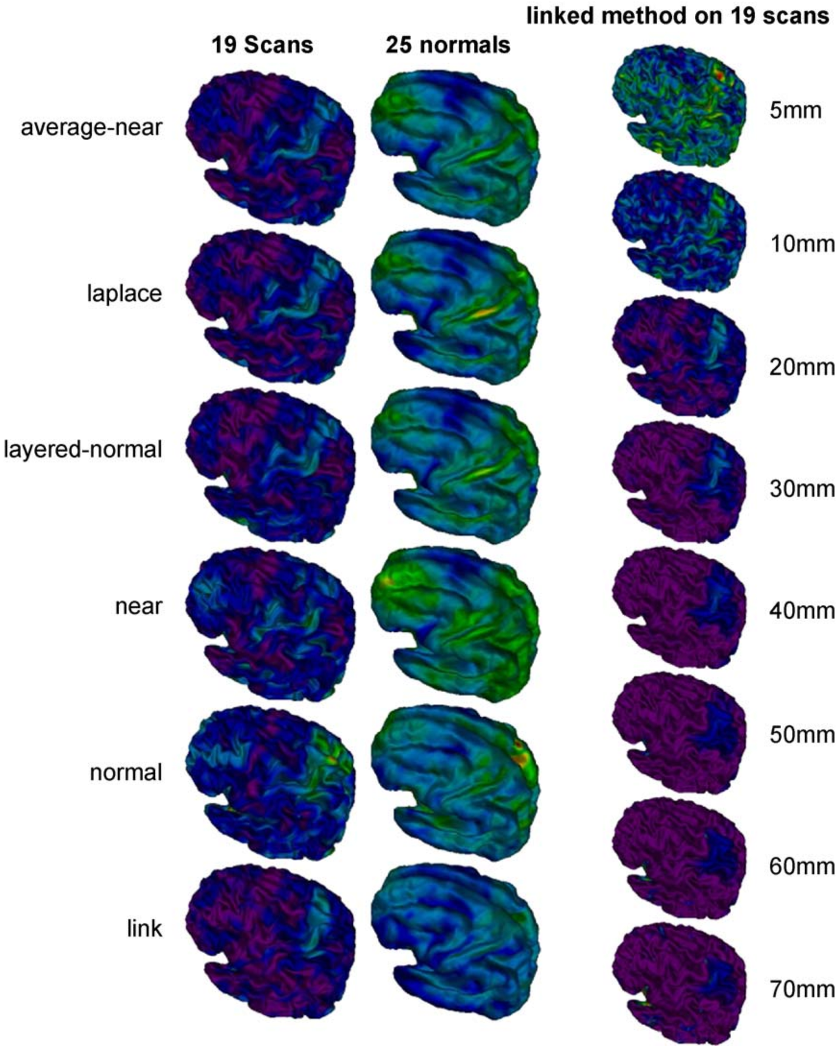

Fig. 5. An illustration of the power of the t_{link} method, standard deviation computed after application of a 30-mm blurring kernel and using a standard deviation of 0.27 mm (computed from the single subject variance shown in Table 3). The top graph shows the number of subjects needed to recover a thickness change of size *delta.* The bottom graph shows the size *delta* that can be recovered given two equal groups of subjects of size *n*. One can see that given an $n > 100$, a 0.35-mm change can be recovered; in two small groups (n = 20), a 1-mm change would reach significance.

comparisons. Since power calculations are dependent on variance, the exact change required to reach significance is not uniform across the cortex, being highest in the superior aspects of the central sulcus, lowest in the prefrontal areas (see Fig. 3). Moreover, statistical significance is dependent on the eFWHM, which in turn is influenced by the amount of blurring, and the resulting resels and statistical thresholds are plotted in Fig. 6. Threshold and resels are minimized at a kernel size of 35 mm.

Population simulation

The simulation was modeled after a comparison between two groups (“controls” and “patients”), where the patient group had their rSTG artificially thinned.

Thirty millimeters is the optimal kernel FWHM

An evaluation of different blurring kernels reveals the classic trade-off: increasing kernel size improves sensitivity but also decreases the ability to accurately estimate the regression slope.

As shown in Fig. 7, sensitivity increases up to a blurring kernel of 35 mm and then declines. The mean slope, which should approach 1 mm since one layer of voxels was removed in the rSTG, declines steadily with increasing kernel sizes. Optimizing the tradeoff between estimation of slope and sensitivity is performed with the equation t = sensitivity × mean (slope), whose maxima is found at 30 mm. The same pattern hold for both t_{link} and t_{laplace}, though the decline in the regression slope is much more noticeable for t_{link}.

The t_{link} is the most sensitive method

The six different cortical thickness metrics were all compared at 30 mm blurring. The performances were compared by evaluating the sensitivity (ability to recover true change) of the different metrics at increasing truth thresholds. Fig. 8 shows the thresholded t statistics maps superimposed onto a sphere along with the probability map of the rSTG that represents the vertices to be recaptured. Visual inspection of these maps shows that all methods show significant results in the correct anatomical region, and that t_{link} provides the most convincing overlap. Fig. 9 describes the same results by graphing sensitivity against probability map threshold and statistical threshold. The t_{link} is clearly the most sensitive method. The t_{laplace} is second, except at high thresholds where t_{normal} surpasses it. This figure also illustrates that regions of greater overlap of rSTG are easier to recover. Four of the six metrics show a decline at high probability map thresholds. The origin of this change is unknown but probably results from an interaction between the shape of the rSTG, the blurring kernel, and the geometric definitions of the thickness metrics. When the truth threshold is set at 50% and the statistical threshold is varied (as shown in the bottom part of Fig. 9), t_{link} is again the most sensitive.

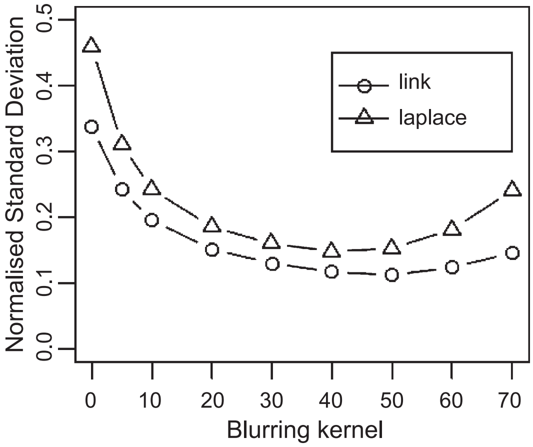

Fig. 6. The effect of blurring kernels on resels and the t statistics threshold using random field theory. The smoothness of the thickness maps as shown by the number of resels decreases up until a 35-mm kernel, increases again thereafter. This increase could be due to increasing influence of noncortical areas (such as the corpus callosum and brain stem cut) included in the mesh. The maximum kernel size that should be used is thus 35 mm.

Controlling for multiple comparisons

The results for any metric can be decomposed into its component statistical measurements such as true-positives, false-positives, and false-negatives, and their respective values evaluated across changing statistical thresholds. This is shown for the t_{link} metric in Fig. 10. False-positives (FP) decline rapidly with increasing thresholds, true-positives (TP) and false-negatives (FN) decrease and increase linearly. The relationship can be described in the thresholding index, index = TP / FP + FN, which maximizes true-positives while simultaneously minimizing false-positives and false-negatives. This function has a maxima at t = −3.3, considerably below the random field threshold of t = −4.67.

Discussion

The goal of this study was to examine a fully automated cortical thickness analysis system, to differentiate between multiple cortical thickness metrics, and to investigate the ability of cortical thickness to differentiate between different populations. Accuracy of the different metrics was never under investigation. Instead, we addressed the question of precision.

Addressing precision is important in its own right, as it can compare the relative ability of different metrics in differentiating between groups of variable thickness. For example, it is conceivable that the t_{normal} metric is the most accurate, yet its high variability would reduce its value in population studies. The precision of a metric measures its usefulness in the multiple subject studies so often undertaken in brain imaging; and while high precision with low accuracy is certainly undesirable, so is high accuracy with low precision.

Comparing the different metrics

We compared six different metrics from three different labs. Each metric was evaluated in terms of variance across a population or a single subject as well as performance in the population simulation. The results from these two tests are related, with normalized standard deviation proving a significant predictor of sensitivity in the population simulation (t = −3.48, P = 0.026, df = 5, R^2 = 0.75). One conclusion is that future thickness methods development should keep the goal of minimizing variability clearly in sight and if necessary increase the complexity of the algorithm to achieve that goal.

The different metrics can be ordered from best to worst by comparing their performance in the variability analysis (see Variability section) and the population simulation (see Population

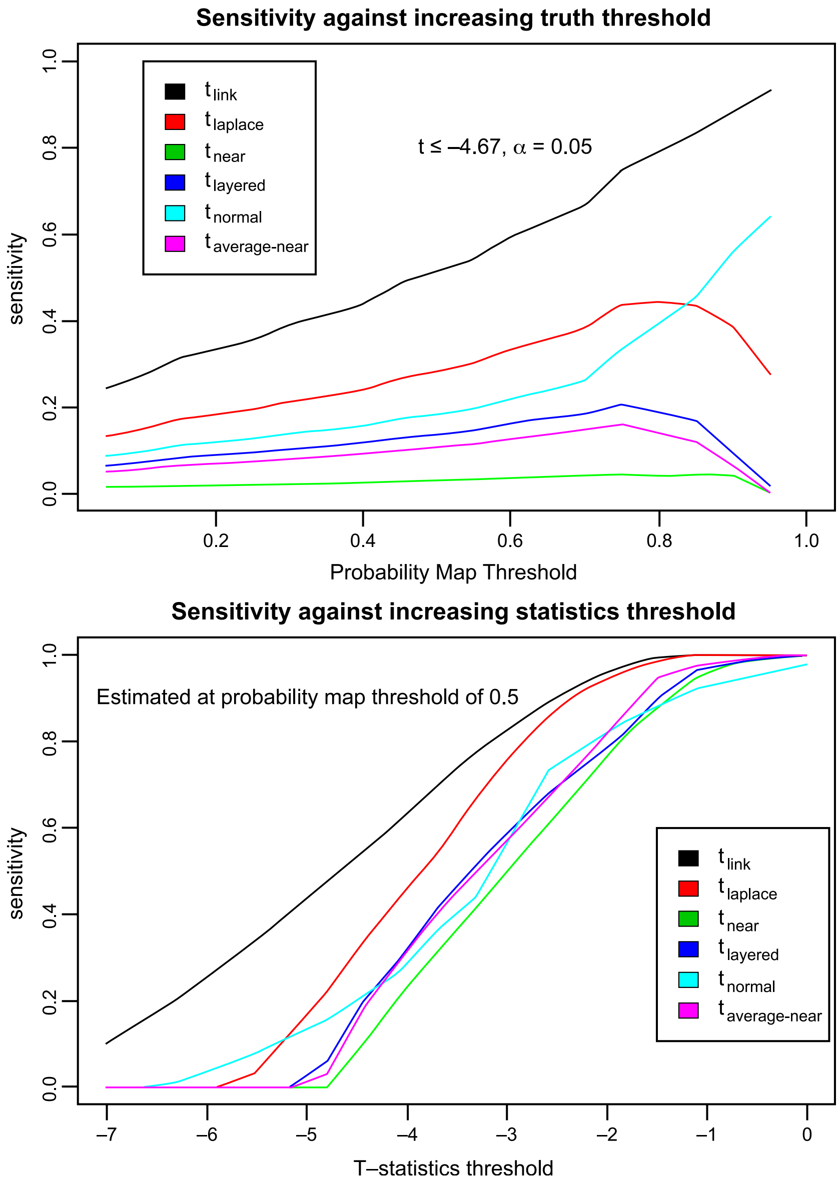

Fig. 7. The effect of increasing blurring kernels as a function of sensitivity and mean regression slope. The maxima of the function t = sensitivity × mean (slope) is to be found at 30 mm, indicating that this is the optimal blurring kernel size for this study. All calculations were made under the assumption that truth is the rSTG probability map thresholded at 1%.

section). The results of these two analyses can be summarized by dividing the sensitivity of a metric by the percentage change that can be recovered given two groups of 25. The ranking is as follows:

1. t_{link}
2. t_{laplace}
3. t_{normal}
4. $t_{\mathrm{layered\text{-}normal}}$
5. $t_{\mathrm{average\text{-}near}}$
6. t_{near}

These metrics were all compared using the default parameters; it is conceivable that tuning the image analysis pipeline in different ways will change the exact results produced in this study. An especially important topic to pursue for further study is the impact that closer spacing of vertices in the cortical meshes has on the variability of thickness metrics (i.e., increasing our vertex count).

Six methods from three different labs were studied; there are, however, others described in the literature on measuring cortical thickness from MRI. The six methods were chosen for their ability to be easily incorporated into our image processing pipeline; any methods that inherently rely on different cortical tessellations, such as Miller et al. (2000) and Zeng et al. (1999), were therefore excluded. Other methods, such as the one introduced by (Yezzi and Prince, 2003) extend methods described and tested in this paper; their improvements might very well lead to improved results in our simulation. We do believe that the population simulation framework used in this study is an elegant way to compare different cortical thickness

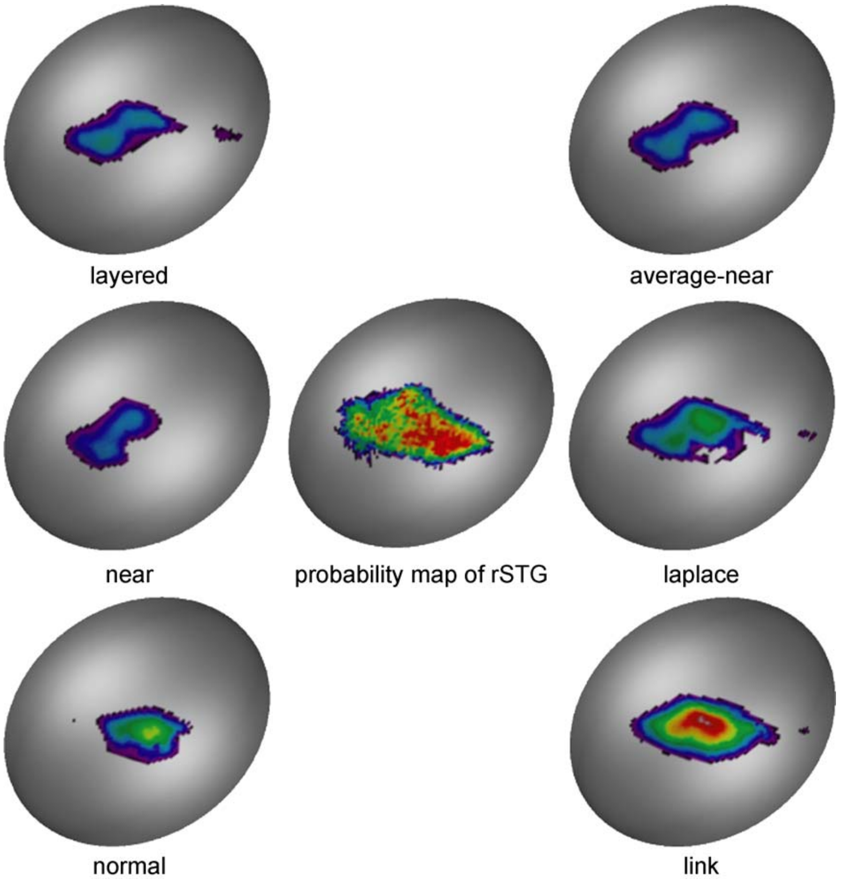

Fig. 8. Results of each of the different metrics at 30 mm blurring, thresholded at $t \geq 2.5$. The results are displayed on a sphere; the sphere in the center shows the probability map of the rSTG, and thus the area of the cortex to be recaptured. Qualitative assessment of the shapes of the t statistics maps indicate that the t_{link} and t_{laplace} produce the closest match to the probability map.

metrics in a controlled yet realistic manner. We have therefore made the volumes used in this study available in order to encourage comparisons of different methods to the six metrics described herein. The data can be found at http://www.bic.mni.mcgill.ca/thickness_population_simulation/.

Varying variability

A noticeable trend in studying the variability across the cortex is its regional variation. The most variable areas of the cortex are the pre- and postcentral gyri, the primary visual areas, and the anterior medial temporal lobes. Two explanations are likely to play a role. First, areas with the thinnest cortex (the motor and visual areas) have high variability. This can be accounted for by the variance induced by the 1-mm sampling in the data sets. Since the sampling stays uniform, but thickness varies, areas with thin cortex are likely to have the highest normalized standard deviation. The second argument relates to the difficulty in segmenting certain areas. The medial temporal lobes, for example, is one of the most challenging for the cortical fitting (Kabani et al., 2001).

Effects of blurring

An open question often asked in voxel-based morphometry applies to cortical thickness analysis as well: What amount of smoothing is desirable? Our results suggest three conclusions in this regard.

- An increase in the FWHM decreases variability, leading to improved sensitivity up until a kernel size of 35 mm (see Fig. 7).
- Increasing FWHM changes the interpretation of the regression slope, potentially underestimating the amount of localized cortical thickness change as described in the Statistical analysis section and shown in Fig. 7.
- Kernels larger than 35 mm actually decrease sensitivity (see Fig. 7).

When there is prior information about the extent of the signal (area of thinning) to be detected, the size of the blurring kernel should match the size of the putative area of change (i.e., matched filter theory; for an overview, see Pratt, 1991). However, for exploratory searches over the whole brain where there is no prior expectation of signal extent, this concept is meaningless. Tuning of blurring kernel size should thus be driven by the desire to limit the FWHM in order to allow for accurate estimation of β, while at the same time staying large enough to retain sensitivity. Ultimately, therefore, the size of the FWHM should be driven by the number of subjects in the study: a large n allows for a smaller kernel, which in turn allows for accurate estimation of the amount of local thickness change, whereas small n still needs larger

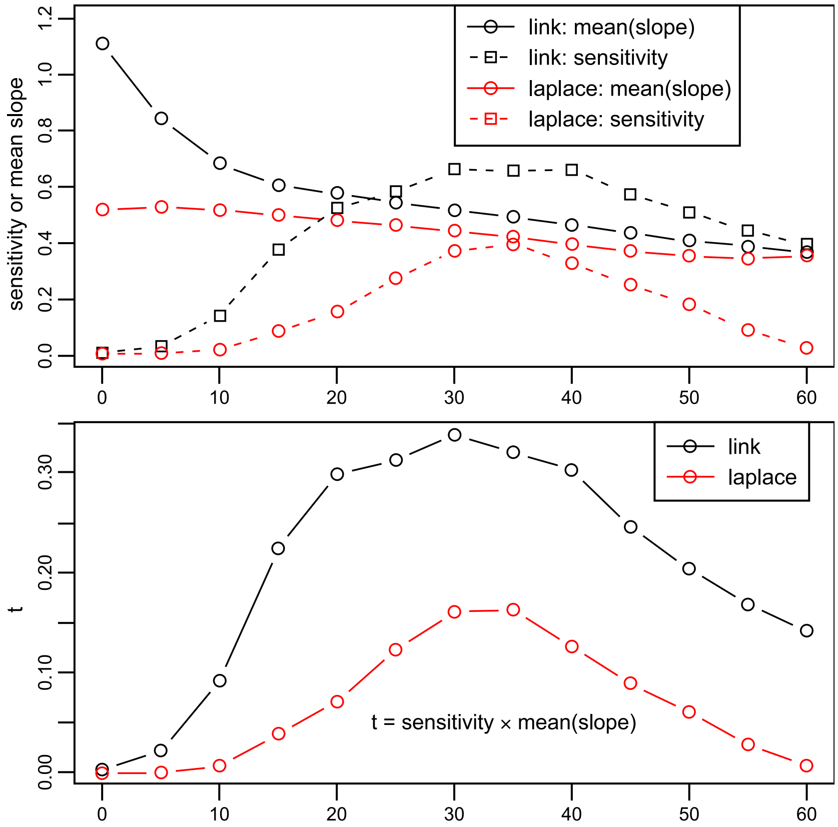

Fig. 9. Sensitivity of the six different cortical thickness metrics at 30 mm blurring, graphed against ever more stringent statistical thresholds (bottom panel) and percentage overlap of the rSTG (top panel). The superiority of the t_{link} metric is clearly noticeable, attaining both a higher sensitivity across different thresholds of the rSTG probability map as well as higher t statistics values compared to the other metrics.

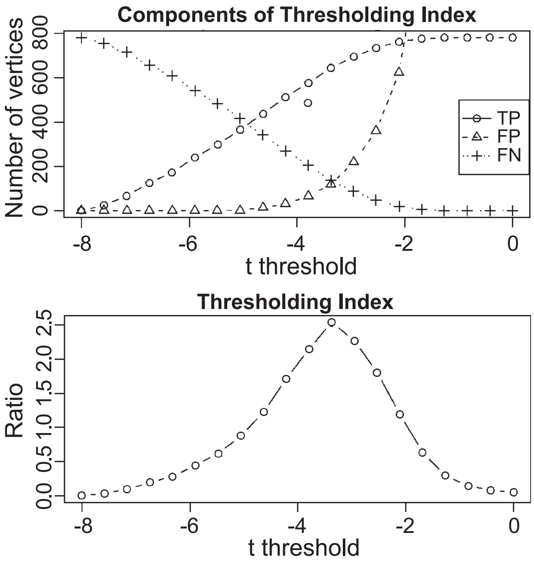

Fig. 10. The top graph shows the individual components that make up the thresholding index shown in the bottom graph. This lower graph was generated at a rSTG threshold of 0.58, chosen since that is where the ratio reaches its maximum (2.54) when searching across all possible rSTG thresholds. Definitions: TP = true-positives, FP = false-positives, FN = false-negatives.

FWHM in order to retain adequate sensitivity. This equation should be balanced by prior hypotheses about the expected area of change.

Thresholding statistical maps

In the implementation described above, 40,962 linear models are analyzed—one for each vertex—with every statistical analysis. Multiple comparisons thus have to be corrected for. The prevailing philosophy in brain imaging to date has been to provide stringent control for type I error, most commonly implemented through applications of random field theory or Bonferroni correction. This stringency, as the population simulation shows, has its costs, as it allows a high percentage of false-negatives. More liberal thresholding techniques, such as the false discovery rate (Genovese et al., 2002), might prove attractive for exploratory studies using cortical thickness analysis. Incidentally, the thresholding index maxima at $t \leq -3.3$ found in Controlling for multiple comparisons section corresponds exactly to a false discovery rate q value of 0.05, indicating that this new technique more closely approximates our ideal thresholding index than the random field theory.

Conclusions

We have shown cortical thickness to be a reliable method, reaching a sensitivity of 0.93. The most precise method is t_{link}. This is due to its ability to minimize variance leading to higher statistical sensitivity. All the metrics had a specificity of 1. While this may seem like a useless index for comparing and contrasting the different metrics, it does indicate a high degree of confidence in any results that are obtained regardless of the metric employed.

Blurring along the surface was shown to be critical, as it significantly increases the sensitivity of cortical thickness analysis. The optimal blurring kernel in our simulation was 30 mm (see Fig. 7). An optimum thresholding index, which maximizes true-positives against both false-negatives and false-positives, was found to lie at $t = 3.3$ (see Fig. 10). Given these optimal parameters and two groups of 25 subjects, a 0.6-mm (15%) change in thickness after 30 mm blurring can be recovered. Increasing the number of subjects to 100 in each group allows for a 0.29-mm (7%) change to be recovered.

In order to validate our methodology, a framework was created to capture the precision of the different thickness metrics and to test the effect of changing parameters for image blurring and statistical thresholding in the analysis pipeline. This general framework can be used to examine future advances in the entire pipeline, such as the impact of different tissue classification methodologies and nonlinear alignment techniques. More work is to be done in validating the accuracy of different metrics and possibly in creating new metrics based on higher resolution anatomical information, which should in turn be evaluated using the precision criteria illustrated in this paper. The data used for this paper has also been made available online to encourage comparisons of other cortical thickness metrics against the ones tested herein.

Acknowledgments

The authors would like to thank Drs. Alex Zijdenbos, Louis Collins, Jens Pruessner, Yasser Ad-Dab'bagh, and Jean-Francois Mangin for their suggestions and comments. Jason Lerch is funded by a K.M. Hunter/CIHR Doctoral Research Award.

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—The methods. NeuroImage 11 (6 Pt 1), 805–821.

Baron, J.C., Chetelat, G., et al., 2001. In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. NeuroImage 14 (2), 298–309.

Brodmann, K., 1909. Vergleichende Lokalisationslehre der Großhirnrinde. Leipzip, Barth.

Chung, M., Worsley, K., et al., 2002. Tensor-Based Surface Morphometry. University of Wisconsin, Madison.

Collins, D.L., Neelin, P., et al., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 18 (2), 192–205.

Collins, D.L., Holmes, C.J., et al., 1995. Automatic 3D model-based neuroanatomical segmentation. Hum. Brain Mapp. 3 (3), 190–208.

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.

Genovese, C.R., Lazar, N.A., et al., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–878.

Holmes, C.J., Hoge, R., et al., 1998. Enhancement of MR images using registration for signal averaging. J. Comput. Assist. Tomogr. 22 (2), 324–333.

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

Kabani, N., Le Goualher, G., et al., 2001. Measurement of cortical thickness using an automated 3-D algorithm: a validation study. NeuroImage 13 (2), 375–380.

Kollokian, V., 1996. Performance analysis of automatic techniques for tissue classification in magnetic resonance images of the human brain. Computer Science. McGill University, Montreal, pp. 24–106.

MacDonald, D., 1997. A method for identifying geometrically simple surfaces from three dimensional images. School of Computer Science. McGill University, Montreal, pp. 59–143.

MacDonald, D., Kabani, N., et al., 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. NeuroImage 12 (3), 340–356.

Mazziotta, J., Toga, A., et al., 2001. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos. Trans. R. Soc. London, B Biol. Sci. 356 (1412), 1293–1322.

Meyer, J.R., Roychowdhury, S., et al., 1996. Location of the central sulcus via cortical thickness of the precentral and postcentral gyri on MR. Am. J. Neuroradiol. 17 (9), 1699–1706.

Miller, M.I., Massie, A.B., et al., 2000. Bayesian construction of geometrically based cortical thickness metrics. NeuroImage 12 (6), 676–687.

Paus, T., Zijdenbos, A., et al., 1999. Structural maturation of neural pathways in children and adolescents: in vivo study. Science 283 (5409), 1908–1911.

Pratt, W.K., 1991. Digital Image Processing. John Wiley and Sons, Inc., New York.

Pruessner, J.C., Collins, D.L., et al., 2001. Age and gender predict volume decline in the anterior and posterior hippocampus in early adulthood. J. Neurosci. 21 (1), 194–200.

Rosas, H.D., Liu, A.K., et al., 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58 (5), 695–701.

Sled, J.G., Zijdenbos, A.P., et al., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17 (1), 87–97.

Tosun, D., Rettman, M.E., et al., 2001. Calculation of human cerebral cortical thickness on opposing sulcal banks. Proceedings of 7th International Conference on Functional Mapping of the Human Brain.

Von Economo, K., Koskinas, G., 1925. Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen. Julius Springer, Berlin.

Worsley, K.J., Evans, A.C., et al., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12 (6), 900–918.

Worsley, K.J., Marrett, S., et al., 1996. A unified statistical approach for determining significant signal in images of cerebral activation. Hum. Brain Mapp. 4, 58–73.

Worsley, K.J., Andermann, M., et al., 1999. Detecting changes in nonisotropic images. Hum. Brain Mapp. 8 (2–3), 98–101.

Wright, I.C., McGuire, P.K., et al., 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. NeuroImage 2 (4), 244–252.

Yezzi Jr., A.J., Prince, J.L., 2003. An Eulerian PDE approach for computing tissue thickness. IEEE Trans. Med. Imag. 22 (10), 1332–1339.

Zeng, X., Staib, L.H., et al., 1999. Segmentation and measurement of the cortex from 3-D MR images using coupled-surfaces propagation. IEEE Trans. Med. Imag. 18 (10), 927–937.

Zijdenbos, A.P., Forghani, R., et al., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imag. 21 (10), 1280–1291.