A Comparison Of Voxel And Surface Based Cortical Thickness Estimation Methods

Matthew J. Clarkson[a,b,*], M. Jorge Cardoso[a], Gerard R. Ridgway[b,a], Marc Modat[a], Kelvin K. Leung[b,a], Jonathan D. Rohrer[b], Nick C. Fox[b], Sébastien Ourselin[a,b]

[a] *Centre for Medical Image Computing (CMIC), The Engineering Front Building, University College London, London, WC1E 6BT, UK*
[b] *Dementia Research Centre, UCL Institute of Neurology, University College London, London, WC1N 3BG, UK*

Abstract

Cortical thickness estimation performed in-vivo via magnetic resonance imaging is an important technique for the diagnosis and understanding of the progression of neurodegenerative diseases. Currently, two different computational paradigms exist, with methods generally classified as either surface or voxel-based. This paper provides a much needed comparison of the surface-based method FreeSurfer and two voxel-based methods using clinical data. We test the effects of computing regional statistics using two different atlases and demonstrate that this makes a significant difference to the cortical thickness results. We assess reproducibility, and show that FreeSurfer has a regional standard deviation of thickness difference on same day scans that is significantly lower than either a Laplacian or Registration based method and discuss the trade off between reproducibility and segmentation accuracy caused by bending energy constraints. We demonstrate that voxel-based methods can detect similar patterns of group-wise differences as well as FreeSurfer in typical applications such as producing group-wise maps of statistically significant thickness change, but that regional statistics can vary between methods. We use a Support Vector Machine to classify patients against controls and did not find statistically significantly different results with voxel based methods compared to FreeSurfer. Finally we assessed longitudinal performance and concluded that currently FreeSurfer provides the most plausible measure of change over time, with further work required for voxel based methods.

Keywords: Cortical thickness estimation, Laplacian, Registration, FreeSurfer, Alzheimer's disease, atrophy

[*] Corresponding author. M. J. Clarkson. Dementia Research Centre, Institute of Neurology, University College London, 8-11 Queen Square, London, WC1N 3BG, UK. Fax: +44 207 676 2066.
Email address: m.clarkson@ucl.ac.uk (Matthew J. Clarkson)

1. Introduction

The human cerebral cortex is a highly folded layer or ribbon of interconnected neurons, with an average thickness of around 2.5mm - varying between 1 and 4.5mm in different parts of the brain (Fischl and Dale, 2000; von Economo, 1929). There is significant variability between individuals in disease and in health. The cortex plays a key role in most cognitive processes and demonstrates regional specification such that visual function, language, calculation, executive function and so on, have relatively localised cortical representation in different parts of the brain. The thickness of the cortex is of interest as it develops, follows the normal ageing process and changes under a wide variety of neurodegenerative diseases. Recently, imaging studies of cortical thickness have compared the group-wise differences between healthy control subjects and patients with conditions such as sporadic and familial Alzheimer's disease (AD) (Lerch et al., 2005; Gutierrez-Galve et al., 2009; Knight et al., 2009), fronto-temporal lobar degeneration (FTLD) (Du et al., 2007; Rohrer et al., 2009), posterior cortical atrophy (Lehmann et al., 2009), multiple sclerosis (Sailer et al., 2003), Huntington's disease (Rosas et al., 2008), and the changes that occur in healthy controls under normal ageing (Salat et al., 2004).

The methods for estimating cortical thickness from magnetic resonance (MR) images can be broadly categorised as surface based, or voxel based. Both of these methods require an initial segmentation to separate grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). In this paper the WM/GM boundary is referred to as the *WM boundary*, and the GM/CSF boundary as the *pial boundary*.

Surface based methods typically construct a triangulated mesh based on either the WM boundary (Dale et al., 1999; Fischl et al., 1999; Fischl and Dale, 2000; Shattuck and Leahy, 2002; Xu et al., 1999; Han et al., 2004), or the pial boundary (Davatzikos and Bryan, 1996), which is then deformed to find the opposing boundary. Alternatively, with WM and pial boundaries defined, both boundaries can be deformed simultaneously using either snake like deformable models (MacDonald et al., 2000; Kim et al., 2005) or level sets (Zheng et al., 1999), thereby utilising distance constraints to ensure a realistic coupling of the two surfaces. The use of explicit surface models enables sub-voxel accuracy (Fischl and Dale, 2000), high sensitivity (Lerch and Evans, 2005), and robustness to different field strengths, scanner upgrade and scanner manufacturer (Han et al., 2006). With the cortex closed at the brain stem, the resultant surface is topologically equivalent to a sphere. Surface based cortical thickness methods try to ensure correct topology of the surface after initial segmentation of the WM boundary (Shattuck and Leahy, 2001; Xu et al., 1999; Han et al., 2004), using smoothness and self intersection constraints (Dale et al., 1999; MacDonald et al., 2000), by correcting topological defects as they occur (Fischl et al., 2001; Segonne et al., 2005), or using a Laplacian function (Kim et al.,

2005). Ensuring correct topology or surface regularity massively increases computational cost (Fischl et al., 2001; Han et al., 2004), may require a difficult balance of parameter weights (Kim et al., 2005; Scott et al., 2009), and reduces the model's ability to follow areas of high curvature such as extremely thin gyral stalks (Lohmann et al., 2003) or opposing sides of sulci with no clear CSF between, which can produce bias and error in thickness measurements (Scott et al., 2009).

In contrast, voxel based methods (Lohmann et al., 2003; Hutton et al., 2008; Acosta et al., 2009; Aganj et al., 2009; Das et al., 2009; Cardoso et al., 2011; Scott et al., 2009) work directly on the voxel grid and are computationally very efficient. However, they are considered to be less accurate due to the limited resolution of the voxel grid, less robust to noise and mis-segmentation and significantly affected by partial volume (PV) effects at the boundaries of convoluted structures such as deep sulci (Acosta et al., 2009). Methods include morphological (Lohmann et al., 2003), line integral (Aganj et al., 2009; Scott et al., 2009), Laplacian (Jones et al., 2000) and registration (Das et al., 2009) based approaches. Laplacian approaches (Hutton et al., 2008; Acosta et al., 2009; Cardoso et al., 2011), solve the Laplace equation (Jones et al., 2000) using boundary value relaxation (Press et al., 1991) or matrix methods (Haidar et al., 2005), calculate thickness by integrating the tangent to the Laplacian scalar field (Jones et al., 2000), summing the Euclidean distance from neighbouring voxels on the same streamline, or using a partial differential equation (Yezzi and Prince, 2003) with boundaries set to zero (Yezzi and Prince, 2003), half the mean voxel dimension (Diep et al., 2007) or using Lagrangian initialisation (Bourgeat et al., 2008; Acosta et al., 2009). In contrast, the registration based approach of Das et al. (2009) uses a greedy diffeomorphic registration algorithm to warp the WM segment to match the GM+WM segment. The thickness is then calculated as the distance that the WM/GM boundary moved during the registration. A potential advantage for voxel based methods may be in the fact that the runtimes can be significantly less than the surface based methods which may enable new application areas.

Thus far, surface based methods have been more widely used than voxel based methods. This may be partly due to long running software efforts, producing accessible software packages such as BrainSuite[1](Shattuck and Leahy, 2001, 2002), BrainVISA[2](Mangin et al., 1995) and FreeSurfer[3](Dale et al., 1999; Fischl et al., 1999; Fischl and Dale, 2000). Of these, FreeSurfer is the most widely used (Nakamura et al., 2010), and the FreeSurfer wiki lists many references on both the methodology and clinical studies.

[1]http://www.loni.ucla.edu/Software/BrainSuite
[2]http://brainvisa.info/
[3]http://surfer.nmr.mgh.harvard.edu/

Recently there has been significant interest in the development of voxel based methods (Hutton et al., 2008; Scott et al., 2009; Acosta et al., 2009; Aganj et al., 2009; Cardoso et al., 2011; Das et al., 2009). In addition, voxel based methods have featured in a comparison with voxel based morphometry (Hutton et al., 2009), been used to correlate changes of cortical thickness with diffusion measures using sparse canonical correlation analysis (Avants et al., 2010) and used in clinical studies (Querbes et al., 2009). However, evaluation of these approaches has been limited by a lack of studies comparing voxel based and surface based methods. This paper aims to provide such a comparison, comparing the freely available surface based **FreeSurfer** (version 4.5.0) method with our implementations of two voxel based methods. We call these voxel-based methods a **Laplacian** based method and a **Registration** based method, and describe both of these below. We chose FreeSurfer as it is the most widely used of the surface based methods (Nakamura et al., 2010). Of the voxel based methods, we chose a Laplacian method similar to Acosta et al. (2009) as many of the references above are Laplacian based, and a registration method similar to Das et al. (2009) as there is current interest in diffeomorphic registration algorithms, many of which could be applied to this application. The methods are compared in terms of reproducibility, disease differentiation and the ability to detect changes of cortical thickness in longitudinal imaging studies.

2. Materials and methods

2.1. The FreeSurfer Method

The FreeSurfer cortical thickness pipeline has been described and validated in previous publications (Dale et al., 1999; Fischl et al., 1999; Fischl and Dale, 2000; Han et al., 2006). Briefly, processing involves intensity normalisation, registration to Talairach space, skull stripping, segmentation of white matter, tesselation of the WM boundary, smoothing of the tesselated surface and automatic topology correction. The tesselated surface is used as the starting point for a deformable surface algorithm to find the WM and then the pial boundary. For each point on the tesselated WM surface, the cortical thickness is calculated as the average of the distance from the WM surface to the closest point on the pial surface and from that point back to the closest point on the WM surface (Fischl and Dale, 2000).

2.2. A Laplacian Based Method

There are several Laplacian based methods implemented in the literature, originating from the paper of Jones et al. (2000). A processing pipeline was implemented consisting of the following steps: an initial probabilistic segmentation (Cardoso et al., 2011) of GM, WM and CSF is performed on a T1 weighted (T1w) image, resulting in probability maps for each tissue type. These probability images are resampled to 0.5mm

iso-tropic voxels using linear interpolation as in Hutton et al. (2008) and then a three label image is formed by labelling voxels as GM where $p(GM) >= 0.80$ and otherwise choosing the tissue type with the highest probability, ignoring $p(GM)$. The boundary is corrected as in section 2.3 of (Acosta et al., 2009) to make sure the GM is at least one voxel wide. The Laplace equation is solved over the GM region (Jones et al., 2000), then thickness calculated via a PDE based approach (Yezzi and Prince, 2003) using Lagrangian initialisation (Acosta et al., 2009). The thickness measurement is capped at 6mm (discussed below). Note that the GM mask is constructed where $p(GM) >= 0.80$, which picks voxels that are classified as being highly likely to be GM, resulting in a relatively thin GM region. The Lagrangian initialisation starts from these high probability of GM voxels and ray casts through the GM probability map, searching for the $p(GM) = 0.5$ boundary, stopping when the probability indicates that another tissue type is more likely. We found this to be more reliable than thresholding the GM probability map directly at $p(GM) = 0.5$ in areas where the GM from opposing sides of a sulci touch. The choice of 0.5mm voxels was made to increase the number of voxels in the GM which improves the convergence of the relaxation methods used to solve the Laplacian and thickness PDEs. This method is comparable to (Acosta et al., 2009), with a different segmentation algorithm (Cardoso et al., 2011) at the start.

2.3. A Registration Based Method

A registration based method was implemented based on Das et al. (2009) and consisted of the following steps: an initial probabilistic segmentation of GM, WM and CSF is performed (Cardoso et al., 2011) and a greedy diffeomorphic registration algorithm was used to expand the WM segment, to match the GM+WM segment or until a maximum of 6mm displacement was reached. From the three probability maps, a three label image is formed by picking the tissue type with the highest probability at each voxel. For each boundary voxel on the GM/WM boundary, the thickness is calculated as the distance moved under the registration transformation, and this thickness value is then propogated across the GM mask. In comparison to the Laplacian method, where we selected $p(GM) >= 0.8$ for the grey matter mask, the registration method is less dependent on this factor. The algorithm relies on having a good WM/GM boundary, so the WM boundary is determined where $p(WM) > p(GM)$, and this is evolved outwards to the GM/CSF boundary, thereby identifying cases where opposite sites of a sulci touch. In the Laplacian method, the segmented probability images are resampled to 0.5mm isotropic voxels. This step is not necessary for the Registration method, as the Registration method, based on (Avants et al., 2006) is performing subvoxel registration anyway, and resampling to smaller voxels would add unnecessary computational and memory overhead. This method is a re-implementation of (Das et al., 2009), with a different segmentation algorithm (Cardoso et al., 2011).

2.4. Voxel Based Processing

Both voxel based methods are capped at 6mm. For the Laplacian method, the Lagrangian initialisation (Bourgeat et al., 2008; Acosta et al., 2009) can suffer due to noisy estimates of the surface normal, leading to erroneously high initialisation estimates. For the registration method, the WM mask is deformed outwards to match the WM+GM mask, and in (Das et al., 2009) a fixed thickness prior (τ in step 3) is applied to stop the registration. In practice, few voxels will reach this limit as the thickness is known to vary between 1 and 4.5mm in different parts of the brain (Fischl and Dale, 2000; von Economo, 1929).

To calculate region based statistics for both voxel based methods, a region of interest must be defined from either an atlas, or a parcellation. In these experiments, we register the AAL atlas (Tzourio-Mazoyer et al., 2002) to each subject using block matching (Ourselin et al., 2000) followed by a spline based non-linear deformation (Modat et al., 2009; Rueckert et al., 1999) both implemented in NiftyReg[4], or alternatively we use the FreeSurfer parcellation to define the regions. For each subject the GM mask is assigned region labels based on the closest atlas or parcellation label. The midline of the GM is extracted by selecting the closest voxel to the midline of the Laplacian field. For each voxel along the midline, the inter-quartile mean of the thickness values within a 3mm radius and within the region of interest was calculated and assigned to the midline voxel. Region based statistics are calculated over the thickness values in the midline voxels.

3. Experiments

Our four experiments were chosen to help inform the reader in a manner that was most relevant to the existing literature, and to clinical research studies. Cortical thickness studies may use different atlases to provide regional based statistics. The first experiment tests the hypothesis that there is no difference in regional statistics when using different atlases. In the absence of a gold-standard, the second experiment assesses the reproducibility of each method. The third and fourth experiments are motivated by the increasing number of cross-sectional and longitudinal studies in the literature.

3.1. Subjects And Scan Selection

In this paper two clinical patient cohorts and matched controls were studied. The clinical subjects were recruited from the Specialist Cognitive Disorders Clinic of the National Hospital of Neurology and Neurosurgery, London, UK. The control subjects were recruited from patient spouses or other healthy age matched

[4] http://sourceforge.net/projects/niftyreg/

volunteers. Informed consent was obtained from all subjects and these studies had local ethics committee approval.

3.1.1. Cohort 1

Cohort 1 consisted of 49 subjects (see Table 1): 33 patients with probable AD, and 16 healthy controls included in a longitudinal clinical and MRI study, details of which are provided in previous publications (Schott et al., 2005, 2006; Barnes et al., 2007, 2008; Gutierrez-Galve et al., 2009). The diagnosis of probable AD was made according to the National Institute of Neurologic, Communicative Disorder and Stroke-Alzheimer disease and Related Disorders Association (NINCDS-ARDA) criteria (McKhann et al., 1984). All subjects had volumetric MRI acquired on a single 1.5T GE Signa scanner (General Electric, Milwaukee, WI). T1-weighted volumetric images were obtained using a spoiled fast GRASS sequence with a 24-cm field of view and a 256×256 field of view to provide 124 contiguous 1.5-mm-thick slices in the coronal plane. The scan acquisition parameters were as follows; TR = 15ms, TE = 5.4ms, Flip angle = 15°, TI=650ms. This dataset was chosen because for each of the 49 subjects, two same-day baseline scans and a single one year repeat image had been obtained.

3.1.2. Cohort 2

Cohort 2 consisted of 101 subjects (see Table 2): 73 patients with clinically diagnosed frontotemporal dementia (FTD) and 28 healthy controls. The FTD patients included 30 patients with progressive non-fluent aphasia (PNFA), 43 patients with semantic dementia (SemD). A clinical diagnosis of SemD was made according to modified Neary criteria as per (Adlam et al., 2006) with patients having fluent speech, marked anomia, impaired word comprehension and deficits in non-verbal semantic domains. A diagnosis of PNFA was made based on modified Neary criteria with patients having a speech production impairment characterised by apraxia of speech and agrammatism. Some of these subjects' data have been used in previous studies (Rohrer et al., 2009; Lehmann et al., 2010b,a). All subjects had volumetric MRI acquired on four different 1.5T GE Signa scanners (General Electric, Milwaukee, WI). T1 weighted volumetric images were obtained using an IR-prepared fast SPGR sequence with a 24-cm field of view and 256×256 matrix, to provide 124 1.5-mm-thick slices in the coronal plane.

3.2. Comparison of Different Atlases

Surface based methods such as FreeSurfer and voxel based methods such as the Laplacian and Registration based methods used for these experiments often assess thickness measures by calculating statistics over regions defined on an anatomical atlas. FreeSurfer (Fischl and Dale, 2000) uses their own atlas, Acosta et al.

(2009) and Cardoso et al. (2011) used the AAL atlas (Tzourio-Mazoyer et al., 2002), whereas Hutton et al. (2009) used the IBASPM atlas (Aleman-Gomez et al., 2006). The voxel based methods were first tested using the FreeSurfer parcellation and the AAL atlas to determine whether different regions in each atlas produced significantly different results. In preparation for the next experiment, FreeSurfer was run on the first baseline scan of each subject in cohort 1, with default settings and no manual editing. For each subject, FreeSurfer resamples the original T1-weighted image to isotropic 1mm voxels. This resampled image was used as the input to the Laplacian based cortical thickness algorithm described above. This is purely a convenience, to make comparison easier, as the input to the voxel based methods can be considered to be in the same coordinate system as the FreeSurfer results. The output is an image where each voxel in the GM contains the thickness at that point. Nine anatomical regions of interest were chosen in advance: the parahippocampal gyrus (PHG), fusiform (FUS), superior temporal gyrus (STG), precuneus (PRE), superior parietal gyrus (SPG), supramarginal gyrus (SMG), lateral occipital sulcus (LO), lingual (L) and the superior frontal gyrus (SFG). These were chosen as they are available in both the FreeSurfer and AAL atlases, and of interest in these neurodegenerative diseases. The AAL atlas was registered to the T1w volume using block matching (Ourselin et al., 2000) followed by a spline based non-linear registration (Modat et al., 2009; Rueckert et al., 1999). For each of the 49 subjects in cohort 1, and each atlas, the mean cortical thickness of over each atlas region was calculated as described in section 2.4. The FreeSurfer and AAL atlases were compared by using paired samples two-tailed t-tests on the mean regional cortical thickness, and Pitman's test to compare the variance for each of the nine regions.

3.3. Results of Comparing Different Atlases

Table 3 shows the mean (standard deviation) of the cortical thickness computed over the regions contained within the FreeSurfer and AAL atlas. Left and right hemispheres have been averaged together. Note that the thickness data remains constant, for rows 1 and 2 in table 3, as it is only the choice of atlas that changes. The third row shows p-values from the paired two-tailed t-tests and the Pitman's tests in brackets. In 7 out of 9 t-tests, there is a significant ($p < 0.05$) difference in mean cortical thickness. We did not find statistically significant evidence of a difference in mean cortical thickness using the two different atlases in the precuneus and lingual regions. In contrast, 7 out of 9 tests of variance showed no statistically significant evidence of a difference in variance, with only the superior temporal gyrus and superior parietal gyrus being statistically significant at the $p < 0.05$ level.

3.4. Comparison Of Reproducibility

FreeSurfer was run on the first and second baseline scan of each of the 49 subjects of cohort 1, with default settings and no manual editing. The FreeSurfer resampled 1mm isotropic T1w image was again used as input to the Laplacian and Registration voxel based methods. For this and all subsequent experiments we selected the FreeSurfer parcellation as the atlas over which to compute regional statistics.

To assess the reproducibility of each method, the standard deviation over all subjects of the difference in regional cortical thickness between the two same day scans was calculated for each region and method. To visualise the results, a single FreeSurfer brain surface was chosen at random, and for each method, the standard deviation of each region was colour coded onto the surface and rendered using Paraview[5]. Pitman's test was used to assess whether there was a significant difference in variance between the three methods, for each of the nine regions.

3.5. Results of Reproducibility Comparison

Figure 1 shows a visual representation of the standard deviation over the 49 subjects of the difference in mean regional cortical thickness between the two same day scans. The FreeSurfer result has a lower standard deviation than the Laplacian method for all meaningful regions[6], and a lower standard deviation than the Registration method for all meaningful regions except the left temporal pole. In 33 out of 70 regions, the Registration method had a lower standard deviation than the Laplacian method. Table 4 shows the mean and standard deviation of the difference in cortical thickness for each of the nine regions and for each of the three methods, again with left and right sides averaged together. FreeSurfer had a statistically significantly ($p < 0.05$) lower variance than either the Laplacian or Registration method for all of the 9 tested regions. The Laplacian method had a statistically significantly ($p < 0.05$) lower variance than the Registration method in the superior frontal gyrus, but we did not find significant differences for the other 8 regions.

3.6. Comparison of Cross Sectional Disease Differentiation

The complete FreeSurfer cortical thickness pipeline was run on cohort 2, and the results edited as described on the FreeSurfer wiki by an experienced neurologist (JR). Using FreeSurfer tools, an average pial surface was created, and a vertex-by-vertex analysis using a general linear model (Worsley et al., 2009) was used to assess differences in cortical thickness between the control subjects and either SemD or PNFA

[5] http://www.paraview.org/

[6] ignoring the FreeSurfer "unknown" region, and the corpus callosum which is set to zero thickness

patients. Cortical thickness C was modelled as a function of group, controlling for age, sex and total intercranial volume (TIV) by including them as nuisance covariates. $C = \beta_1$ SemD $+\beta_2$ PNFA $+\beta_3$ controls $+\beta_4$ age $+\beta_5$ sex $+\beta_6$ TIV $+\mu + \epsilon$ (where μ is a constant, and ϵ is error), with contrasts of interest being the two-tailed t-tests between the estimates of the group parameters, i.e. β_1 and β_3, β_2 and β_3. Two-tailed unpaired t-tests were computed at each vertex, with significance assessed at the $p = 0.05$ level, when corrected for multiple comparisons using the False Discovery Rate (FDR) (Genovese et al., 2002).

In addition, the full Laplacian and Registration based methods were run on cohort 2, again using the T1w image produced by FreeSurfer. The average of the FreeSurfer WM and pial surface was created for each subject. This surface was used to sample the thickness data produced by each voxel based method by finding the closest non-zero thickness voxel to each vertex. This thickness data was projected onto the FreeSurfer average pial surface created above, and the same linear model re-run for both the Laplacian and Registration based methods. The per-vertex p-values of the average surface were visualised for each of the methods and visually assessed for similarity.

Subsequently, the same nine regions used in sections 3.2 and 3.4 were used to compare statistics. For each of the 9 regions the mean cortical thickness was calculated over all vertices (FreeSurfer) or voxels (Laplacian and Registration methods) for each subject. Unpaired samples two-tailed t-tests were performed to test for significant differences, and Cohen's d to test for effect size, comparing the control group with both the SemD and PNFA groups for each region and for each method.

Finally, a linear Support Vector Machine (SVM) was used to classify subjects (Vapnik, 1995, 1998), implemented with LIBSVM version 2.89 (Chang and Lin, 2001) under MATLAB version 7.2.0. The comparison of interest is how well the classifier can separate the three groups, using the thickness data produced by the three methods. Subjects were classified in an n-dimensional space, where n is the total number of vertices in both hemispheres, excluding the medial wall. SVMs identify an optimal separating hyperplane, such that subjects from each group lie as far as possible from the hyperplane, on opposite sides. We use the C-SVM formulation, employing a two-level nested cross-validation to optimise the mis-classification penalty parameter C using a leave one out procedure within the main leave one out loop (Wilson et al., 2009). This ensures an unbiased estimation of genaralisation accuracy by leaving each scan out entirely from the training procedure. A direct comparison of the classification accuracy was performed, by calculating 95% confidence intervals for the difference in accuracy (Newcombe, 1998).

3.7. Results of Cross Sectional Comparison

Figure 2 shows a visual comparison of the three methods, comparing SemD and PNFA patients' cortical thickness with control subjects. Tables 5 and 6 show the t-test p-values and Cohen's d in brackets for each method comparing the mean difference of cortical thickness between control subjects and either SemD or PNFA patients. Table 7 shows the SVM scores in terms of classification accuracy and confidence intervals. The direct comparison of the difference in accuracy rates, gave 95% confidence intervals spanning zero for all pairwise combinations.

3.8. Assessment Of Longitudinal Change

The FreeSurfer longitudinal pipeline was run on the 49 subjects of cohort 1, using the first baseline scan, and the one year repeat scan. The FreeSurfer longitudinal pipeline (version 4.5.0) takes the T1w image at n-timepoints, creates an average T1w image and on this average image creates the WM and pial boundary as described above. These initial surfaces are used as a starting point for a deformable model algorithm at all n-timepoints. In this case $n = 2$. The rationale is to provide a starting point that is unbiased to the order of the images. Both voxel based methods were applied to the FreeSurfer resampled T1w isotropic image for both the baseline and repeat scan independently. Using the FreeSurfer atlas, the mean cortical thickness was calculated for each of the 9 regions and each method, and an anualised percentage change computed as in (Holland et al., 2009).

For the control ($n = 16$) and AD groups ($n = 33$), the mean and standard deviation of cortical thickness was calculated for each region, and Cohen's d was calculated as a measure of effect size.

3.9. Results of Longitudinal Comparison

Table 8 shows the mean (standard deviation) of the regional cortical thickness for each method, for each subject group, and for each of the 9 regions, and the value for Cohen's d for each method. FreeSurfer results in an anualised percentage change that for control subjects ranges from +0.53% (PHG) to -2.14% (SPG), and for AD subjects a percentage change of -2.22% (SPG) to -3.70% (STG), and for all of the 9 regions, the annualised percentage change for AD subjects has consistently higher magnitude (more atrophy) than control subjects. For both the Laplacian and Registration methods 7 out of 9 cases show AD subjects having more atrophy than control subjects. In general it can be seen that the standard deviation of the annualised percentage change for the voxel based methods is higher than for FreeSurfer.

4. Discussion

In this paper we have compared the surface based cortical thickness method FreeSurfer with two voxel based methods. This is a challenging task as the methodologies are significantly different, and we must err on the side of caution in the interpretation of the results. Furthermore, to add to the challenge, it is difficult to obtain a gold standard. Previous authors have used simulated MRI phantoms (Lee et al., 2006) at one time point, or simulations of atrophy (Camara et al., 2008; Lerch and Evans, 2005) for longitudinal studies, however providing a physiologically plausible simulation of atrophy is itself a difficult task. For this reason, we chose to compare the performance of the algorithm according to reproducibility and both cross-sectional and longitudinal group differentiation, which are common applications within the literature.

We assessed the influence of the atlases used to define anatomical regions: atlas creation is an extensive topic within the literature, with each atlas dependent on the quantity and quality of data, the segmentation and registration algorithms used, and the demographics of the subjects themselves. For these reasons, the borders of anatomical regions in different atlases are expected to be different. We show that regional means and standard deviations of cortical thickness, calculated using an identical method, differ significantly depending on which atlas is used - with up to 10% difference in certain regions assigned the same label. This result is important for this paper, as it indicates that for a fair comparison, we must use the same atlas for all three methods, but furthermore, it has implications when interpreting results from other papers. Simply put, caution is advised when comparing the results of different studies, whether the comparison is at a methodological or clinical level, whenever the underlying atlas is different.

Subsequently we assessed the reproducibility of the thickness measurements in experiment 3.4. The surface and voxel based methods are fundamentally different. The FreeSurfer surface based method creates a WM segmentation, then a tesselated surface mesh, and deforms that mesh to find both surfaces. This means that reproducibility will be affected by the consistency of the segmentation and also the performance of the deformable model process, whereby the evolving mesh will have a good opportunity to correct for any segmentation differences. The surface will deform and converge to a consistent local minima on two different scans and be guided or restricted by the bending energy constraints of the mesh. Although these bending energy constraints may themselves cause the segmentation to be incorrect, such as in thin gyral stalks (Lohmann et al., 2003), or buried sulci, at least the results will be consistent. On the other hand, voxel based methods create an initial segmentation, and then measure the thickness directly. Any errors, or differences between scans that result in a single voxel being differently classified may impact the thickness results. Figure 1 shows a visual representation of the consistency of the algorithms by projecting onto a

randomly chosen single subject brain surface the standard deviation over each region of the difference in cortical thickness for two same day scans. The FreeSurfer results have a lower standard deviation than the Laplacian method for all regions, and a lower standard deviation than the Registration method for all regions apart from the left temporal pole. For the Laplacian method, we tried both 1mm iso-tropic and 0.5mm iso-tropic voxels. The Laplacian method uses a grid based relaxation process (Press et al., 1991) to solve the Laplace equation and the thickness PDE. The cortex varies between 1 and 4.5mm in different parts of the brain (Fischl and Dale, 2000; von Economo, 1929), which means that with 1mm iso-tropic voxels the grey matter might be only 1 - 4 voxels wide. This may lead to a poor convergence of the relaxation process, and additionally poor estimation of surface normals. Simply by sub-sampling to 0.5mm helps alleviate these problems, and this approach can be seen in the work of Hutton et. al. (Hutton et al., 2008, 2009). Subsampling further may improve results, but becomes prohibitively expensive in terms of memory and computational cost. It can be seen that both the Laplacian and Registration methods produce very visually similar results and in 8 out of 9 tested regions, we found no significant difference between the regional variance in thickness (Table 4). Furthermore, the mean difference shows negligible bias for all three methods.

We compared the three algorithms in terms of the ability to detect group wise differences (experiment 3.6). This is a typical application found in the literature, with conclusions typically drawn based on visual inspection. Figure 2 shows an average brain, colour coded with regions where there is statistically significant evidence ($p < 0.05$), when corrected for multiple comparisons using the FDR method (Genovese et al., 2002), of SemD patients (figure 2a) or PNFA patients (figure 2b) having thinner cortex than control subjects. In figure 2, the areas where $p > 0.05$ are all grey, so all coloured areas are deemed to show statistically significant evidence of thinning (red to yellow), or thickening (blue to light blue), relative to control subjects. The three columns in each sub-figure show the results for each method. Referring to figure (a), for SemD patients, all 3 methods are suitable for detecting group-wise differences, displaying qualitatively similar results. All 3 methods display atrophy on the left more than right side, and in concurrence with (Rohrer et al., 2009), we see evidence of atrophy in the left temporal lobe, in particular the temporal pole, entorhinal cortex, parahippocampus, and inferior temporal gyri for all three methods. There is also evidence of atrophy in the right temporal lobe, in particular the entorhinal cortex, temporal pole and parahippocampus for all three methods. However, FreeSurfer additionally found evidence of atrophy in the fusiform, an area known to be very atrophic in SemD (Chan et al., 2001). For PNFA patients the FreeSurfer method produces evidence of atrophy in the left superior temporal lobe, banks of the superior temporal sulcus and some evidence in the

left inferior frontal lobes. By contrast, both voxel based methods find a more extensive spread of atrophy in the left temporal lobe, with the Laplacian method extending to the inferior midbrain. Additionally, both voxel based methods find evidence of atrophy in the right temporal lobe.

When region based averages were derived (Table 5 and 6), the regions that differed between cases and controls varied between the methods. Nonetheless for some regions all methods showed significant atrophy. For example for SemD patients, the parahippocampal gyrus, supramarginal gyrus, lingual and left superior frontal gyrus have significant evidence of atrophy for all three methods. However, the fact that these results do differ for each method suggests that care should be taken at every stage of processing in any cortical thickness pipeline, and cohorts should be as large as possible. Furthermore, the p-values and Cohen's d values combined demonstrate that there are cases where voxel based methods can show larger effect sizes than FreeSurfer, and vice versa. Voxel based methods in particular would benefit from improvements that drive down the standard deviation of thickness measurements. In Tables 5 and 6 we can see that effect size provides additional information to significance tests. As with p-values, the results vary, with both FreeSurfer and the Laplacian method more consistently producing larger negative (atrophy) values than the Registration method.

We did not find any statistically significant evidence of a difference between methods when using an SVM to try and classify controls from SemD patients or controls from PNFA patients. This fits with other studies that suggest that voxel based methods are capable of finding similar group-wise differences when applied to a cross sectional study (Hutton et al., 2008; Acosta et al., 2009; Querbes et al., 2009).

Longitudinal cortical thickness measurement has been proposed as a potential bio-marker (Desikan et al., 2009) however the available methods are still under active development. The FreeSurfer longitudinal pipeline was released with version 4.5.0 (Aug 2009), and provides an unbiased methodology whereby the WM and pial surfaces are created on an average volume and deformed to match each timepoint. Voxel based longitudinal methods have been proposed such as CLADA (Nakamura et al., 2010) and also Das et al. (2009) segment a baseline scan and measure thickness on the baseline scan, then use registration to warp the baseline image to the follow-up image (Das et al., 2009). For the experiments in this paper, we wanted to simply test the capability of applying the thickness calculations to two timepoints, as each method has been more widely used in a cross sectional sense. For all three methods, thickness was calculated at two points and an annualised percentage change calculated for each region as in (Holland et al., 2009). Whilst no gold standard exists, one would expect AD patients to have greater atrophy than control subjects, and for neither group to have increasing cortical thickness. FreeSurfer is most consistent with this hypothesis, with only the

parahippocampal gyrus showing an increase in thickness for control subjects, all other longitudinal changes being a reduction in thickness, and AD patients showing greater reduction in thickness than controls. The Laplacian method has 3 regions showing increasing cortical thickness and in 7 out of 9 regions AD patients show greater reduction in thickness than controls. Similar results can be seen for the Registration method. It can be seen that in general FreeSurfer provides a larger effect size than both the Laplacian and Registration methods for 8 regions, with the exception being the superior parietal gyrus. This may also be a consequence of the improved reproducibility seen on the two same-day scans. In the voxel based methods, even small change around the borders of an object can influence the thickness results, making it difficult to detect small changes in cortical thickness. For example, a 2% change in a 4mm thick region is only 0.08mm. Future work should include a comparison of true longitudinal methods, using 2 or preferably more timepoints.

5. Conclusions

This paper is the first to compare voxel and surface based cortical thickness estimation methods. The choice of atlas produces a significant effect on regional based statistics, suggesting that the comparison of cortical thickness results across different papers, where the authors have used different atlases should proceed with caution. FreeSurfer produced more reproducible results on same day scans than both the Laplacian and Registration methods in all but one cortical regions, with the Laplacian and Registration methods performing similarly. FreeSurfer benefits from the deformable model settling to a consistent boundary, and the smoothness constraints therein enforcing consistent results. Furthermore, this consistency plays a part in a more convincing measure of longitudinal change, that currently the voxel-based Laplacian and Registration methods reviewed here do not possess. We also conclude that for group-wise studies where the aim is to produce maps of statistically significant changes in thickness for visual comparison, both surface and voxel based methods produce comparable results. Furthermore, using and SVM we did not find statistically significant evidence of a difference in methods when performing a classification task. Comparisons of methods such as this will hopefully stimulate efforts to improve different cortical thickness measures.

6. Acknowledgments

This work was undertaken at UCL/UCLH who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. The Dementia Research Centre is an Alzheimer's Research Trust Co-ordinating centre and has also received equipment funded by the Alzheimer's

Research Trust. MC and KL were supported by TSB grant M1638A. MC was additionally funded by CBRC grant 168. NCF was funded by the MRC (UK).

Acosta, O., Bourgeat, P., Zuluaga, M. A., Fripp, J., Salvado, O., Ourselin, S., The Alzheimer's Disease Neuroimaging Initiative, Oct 2009. Automated voxel-based 3D cortical thickness measurement in a combined Lagrangian-Eulerian PDE approach using partial volume maps. Med Image Anal 13 (5), 730–743.

Adlam, A.-L. R., Patterson, K., Rogers, T. T., Nestor, P. J., Salmond, C. H., Acosta-Cabronero, J., Hodges, J. R., Nov 2006. Semantic dementia and fluent primary progressive aphasia: two sides of the same coin? Brain 129 (Pt 11), 3066–3080.

Aganj, I., Sapiro, G., Parikshak, N., Madsen, S. K., Thompson, P. M., Oct 2009. Measurement of cortical thickness from MRI by minimum line integrals on soft-classified tissue. Hum Brain Mapp 30 (10), 3188–3199.

Aleman-Gomez, Y., Melie-Garcia, L., Valdes-Hernandez, P., June 2006. IBASPM: Toolbox for the automatic parcellation of brain structures. In: The 12th Annual Meeting of the Organization for Human Brain Mapping. Florence, Italy.

Avants, B. B., Cook, P. A., Ungar, L., Gee, J. C., Grossman, M., Apr 2010. Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. Neuroimage 50 (3), 1004–1016.

Avants, B. B., Schoenemann, P. T., Gee, J. C., Jun 2006. Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. Med Image Anal 10 (3), 397–412.

Barnes, J., Boyes, R. G., Lewis, E. B., Schott, J. M., Frost, C., Scahill, R. I., Fox, N. C., Nov 2007. Automatic calculation of hippocampal atrophy rates using a hippocampal template and the boundary shift integral. Neurobiol Aging 28 (11), 1657–1663.

Barnes, J., Scahill, R. I., Frost, C., Schott, J. M., Rossor, M. N., Fox, N. C., Aug 2008. Increased hippocampal atrophy rates in AD over 6 months using serial MR imaging. Neurobiol Aging 29 (8), 1199–1203.

Bourgeat, P., Acosta, O., Zuluaga, M., Fripp, J., Salvado, O., Ourselin, S., May 2008. Improved cortical thickness measurement from MR images using partial volume estimation. In: Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on. IEEE, Paris, pp. 205 – 208.

Camara, O., Schnabel, J. A., Ridgway, G. R., Crum, W. R., Douiri, A., Scahill, R. I., Hill, D. L. G., Fox, N. C., Aug 2008. Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal Alzheimer's disease images. Neuroimage 42 (2), 696–709.

Cardoso, M. J., Clarkson, M. J., Ridgway, G. R., Modat, M., Fox, N. C., Ourselin, S., Initiative, T. A. D. N., Jun 2011. Load: A locally adaptive cortical segmentation algorithm. Neuroimage 56 (3), 1386–1397.

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., Rossor, M. N., Apr 2001. Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. Ann Neurol 49 (4), 433–442.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Dale, A. M., Fischl, B., Sereno, M. I., 1999. Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction. NeuroImage 9, 179–194.

Das, S. R., Avants, B. B., Grossman, M., Gee, J. C., Apr 2009. Registration based cortical thickness measurement. Neuroimage 45 (3), 867–879.

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.

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., Schmansky, N. J., Greve, D. N., Salat, D. H., Buckner, R. L., Fischl, B., Initiative, A. D. N., Aug 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. Brain 132 (Pt 8), 2048–2057.

Diep, T.-M., Bourgeat, P., Ourselin, S., 2007. Efficient Use of Cerebral Cortical Thickness to Correct Brain MR Segmentation. In: ISBI. pp. 592–595.

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

Fischl, B., Dale, A. M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. PNAS 97 (20), 11050–11055.

Fischl, B., Liu, A., Dale, A. M., 2001. Automated Manifold Surgery: Constructing Geometrically Accurate and Topologically Correct Models of the Human Cerebral Cortex. IEEE Transactions on Medical Imaging 20 (1), 70–80.

Fischl, B., Sereno, M. I., Dale, A. M., 1999. Cortical Surface-Based Analysis II. Inflation, Flattening and a Surface-Based Coordinate System. NeuroImage 9, 195–207.

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

Gutierrez-Galve, L., Lehmann, M., Hobbs, N. Z., Clarkson, M. J., Ridgway, G. R., Crutch, S., Ourselin, S., Schott, J. M., Fox, N. C., Barnes, J., 2009. Patterns of cortical thickness according to APOE genotype in Alzheimer's disease. Dement Geriatr Cogn Disord 28 (5), 476–485.

Haidar, H., Egorova, S., Soul, J. S., 2005. New Numerical Solution of the Laplace Equation for Tissue Thickness Measurement in Three-Dimensional MRI. Journal of Mathematical Modelling and Algorithms 4, 83–97, 10.1007/s10852-004-3524-0.

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, 180–194.

Han, X., Pham, D. L., Tosun, D., Rettmann, M. E., Xu, C., Prince, J. L., Nov 2004. CRUISE: cortical reconstruction using implicit surface evolution. Neuroimage 23 (3), 997–1012.

Holland, D., Brewer, J. B., Hagler, D. J., Fenema-Notestine, C., Dale, A. M., the Alzheimer's Disease Neuroimaging Initiative, Nov 2009. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. Proc Natl Acad Sci U S A 106, 20954–20959.

Hutton, C., De Vita, E., Ashburner, J., Deichmann, R., Turner, R., May 2008. Voxel-based cortical thickness measurements in MRI. Neuroimage 40 (4), 1701–1710.

Hutton, C., Draganski, B., Ashburner, J., Weiskopf, N., Nov 2009. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. Neuroimage 48 (2), 371–380.

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

Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., Evans, A. C., Aug 2005. Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. Neuroimage 27 (1), 210–221.

Knight, W. D., Kim, L. G., Douiri, A., Frost, C., Rossor, M. N., Fox, N. C., Dec 2009. Acceleration of cortical thinning in familial Alzheimer's disease. Neurobiol Aging.

Lee, J. K., Lee, J.-M., Kim, J. S., Kim, I. Y., Evans, A. C., Kim, S. I., Jun 2006. A novel quantitative cross-validation of different cortical surface reconstruction algorithms using MRI phantom. Neuroimage 31 (2), 572–584.

Lehmann, M., Crutch, S. J., Ridgway, G. R., Ridha, B. H., Barnes, J., Warrington, E. K., Rossor, M. N., Fox, N. C., Sep 2009. Cortical thickness and voxel-based morphometry in posterior cortical atrophy and typical Alzheimer's disease. Neurobiol Aging.

Lehmann, M., Douiri, A., Kim, L. G., Modat, M., Chan, D., Ourselin, S., Barnes, J., Fox, N. C., Feb 2010a. Atrophy patterns in Alzheimer's disease and semantic dementia: a comparison of FreeSurfer and manual volumetric measurements. Neuroimage 49 (3), 2264–2274.

Lehmann, M., Rohrer, J. D., Clarkson, M. J., Ridgway, G. R., Scahill, R. I., Modat, M., Warren, J. D., Ourselin, S., Barnes, J., Rossor, M. N., Fox, N. C., Feb 2010b. Reduced Cortical Thickness in the Posterior Cingulate Gyrus is Characteristic of Both Typical and Atypical Alzheimer's Disease. J Alzheimers Dis 20, 587–598.

Lerch, J. P., Evans, A. C., Jan 2005. Cortical thickness analysis examined through power analysis and a population simulation. Neuroimage 24 (1), 163–173.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., Evans, A. C., Jul 2005. Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. Cereb Cortex 15 (7), 995–1001.

Lohmann, G., Preul, C., Hund-Georgiadis, M., Jul 2003. Morphology-based cortical thickness estimation. Inf Process Med Imaging 18, 89–100.

MacDonald, D., Kabani, N., Avis, D., Evans, A. C., Sep 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. Neuroimage 12 (3), 340–356.

Mangin, J.-F., Frouin, V., Bloch, I., Regis, J., Lopez-Krahe, J., 1995. From 3D Magnetic Resonance Images to Structural Representations of the Cortex Topography using Topology Preserving Deformations. Journal of Methematical Imaging And Vision 5, 297–318.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E. M., Jul 1984. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 34 (7), 939–944.

Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., Fox, N. C., Ourselin, S., Oct 2009. Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine 98, 278–284.

Nakamura, K., Fox, R., Fisher, E., Jul 2010. Clada: Cortical longitudinal atrophy detection algorithm. Neuroimage 54, 278–289.

Newcombe, R. G., 1998. Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine 17, 2635–2650.

Ourselin, S., Roche, A., Prima, S., Ayache, N., 2000. Block Matching: A General Framework to Improve Robustness of Rigid Registration of Medical Images. In: Medical Image Computing and Computer Assisted Intervention. Lecture Notes in Computer Science. Springer Berlin / Heidelberg.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 1991. Numerical Recipes in C The Art of Scientific Computing. Cambridge Press.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Dmonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., Initiative, A. D. N., Aug 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132 (Pt 8), 2036–2047.

Rohrer, J. D., Warren, J. D., Modat, M., Ridgway, G. R., Douiri, A., Rossor, M. N., Ourselin, S., Fox, N. C., May 2009. Patterns of cortical thinning in the language variants of frontotemporal lobar degeneration. Neurology 72 (18), 1562–1569.

Rosas, H. D., Salat, D. H., Lee, S. Y., Zaleta, A. K., Pappu, V., Fischl, B., Greve, D., Hevelone, N., Hersch, S. M., Apr 2008. Cerebral cortex and the clinical expression of Huntington's disease: complexity and heterogeneity. Brain 131 (Pt 4), 1057–1068.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., Hawkes, D. J., 1999. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. IEEE Transactions On Medical Imaging 18 (8), 712–721.

Sailer, M., Fischl, B., Salat, D., Tempelmann, C., Schnfeld, M. A., Busa, E., Bodammer, N., Heinze, H.-J., Dale, A., Aug 2003. Focal thinning of the cerebral cortex in multiple sclerosis. Brain 126 (Pt 8), 1734–1744.

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S. R., Busa, E., Morris, J. C., Dale, A. M., Fischl, B., Jul 2004. Thinning of the cerebral cortex in aging. Cereb Cortex 14 (7), 721–730.

Schott, J. M., Frost, C., Whitwell, J. L., Macmanus, D. G., Boyes, R. G., Rossor, M. N., Fox, N. C., Sep 2006. Combining short interval MRI in Alzheimer's disease: Implications for therapeutic trials. J Neurol 253 (9), 1147–1153.

Schott, J. M., Price, S. L., Frost, C., Whitwell, J. L., Rossor, M. N., Fox, N. C., Jul 2005. Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months. Neurology 65 (1), 119–124.

Scott, M. L. J., Bromiley, P. A., Thacker, N. A., Hutchinson, C. E., Jackson, A., Apr 2009. A fast, model-independent method for cerebral cortical thickness estimation using MRI. Med Image Anal 13 (2), 269–285.

Segonne, F., Grimson, E., Fischl, B., 2005. A genetic algorithm for the topology correction of cortical surfaces. Inf Process Med Imaging 19, 393–405.

Shattuck, D. W., Leahy, R. M., Nov 2001. Automated graph-based analysis and correction of cortical volume topology. IEEE Trans Med Imaging 20 (11), 1167–1177.

Shattuck, D. W., Leahy, R. M., Jun 2002. BrainSuite: an automated cortical surface identification tool. Med Image Anal 6 (2), 129–142.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. NeuroImage 15, 273–289.

Vapnik, V., 1995. The nature of statistical learning theory. Springer, New York.

Vapnik, V., 1998. Statistical learning theory. John Wiley and Sons, New York.

von Economo, C., 1929. The Cytoarchitectonics Of The Human Cerebral Cortex. Oxford University Press.

Wilson, S. M., Ogar, J. M., Laluz, V., Growdon, M., Jang, J., Glenn, S., Miller, B. L., Weiner, M. W., Gorno-Tempini, M. L., Oct 2009. Automated MRI-based classification of primary progressive aphasia variants. Neuroimage 47 (4), 1558–1567.

Worsley, K., Taylor, J., Carbonell, F., Chung, M., Duerden, E., Bernhardt, B., Lyttelton, O., Boucher, M., Evans, A., 2009. Surfstat: A matlab toolbox for the statistical analysis of univariate and multivariate surface and volumetric data using linear mixed effects models and random field theory. NeuroImage 47 (Supplement 1), S102 – S102, organization for Human Brain Mapping 2009 Annual Meeting.
URL http://www.nitrc.org/projects/surfstat

Xu, C., Pham, D. L., Rettmann, M. E., Yu, D. N., Prince, J. L., Jun 1999. Reconstruction of the human cerebral cortex from magnetic resonance images. IEEE Trans Med Imaging 18 (6), 467–480.

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

Zheng, X., Staib, L. H., T.Schultz, R., Duncan, J. S., 1999. Segmentation and measurement of the cortex from 3D MR images using coupled surfaces propogation. IEEE Trans Med Imaging 18, 100–111.

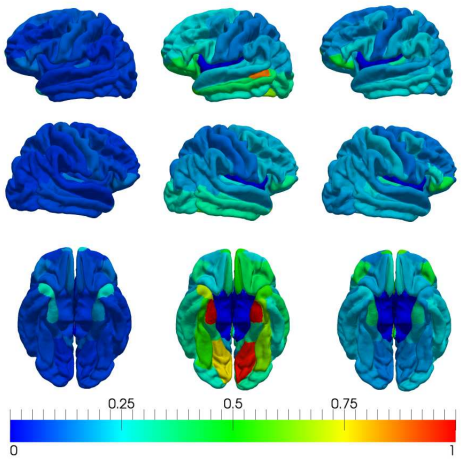
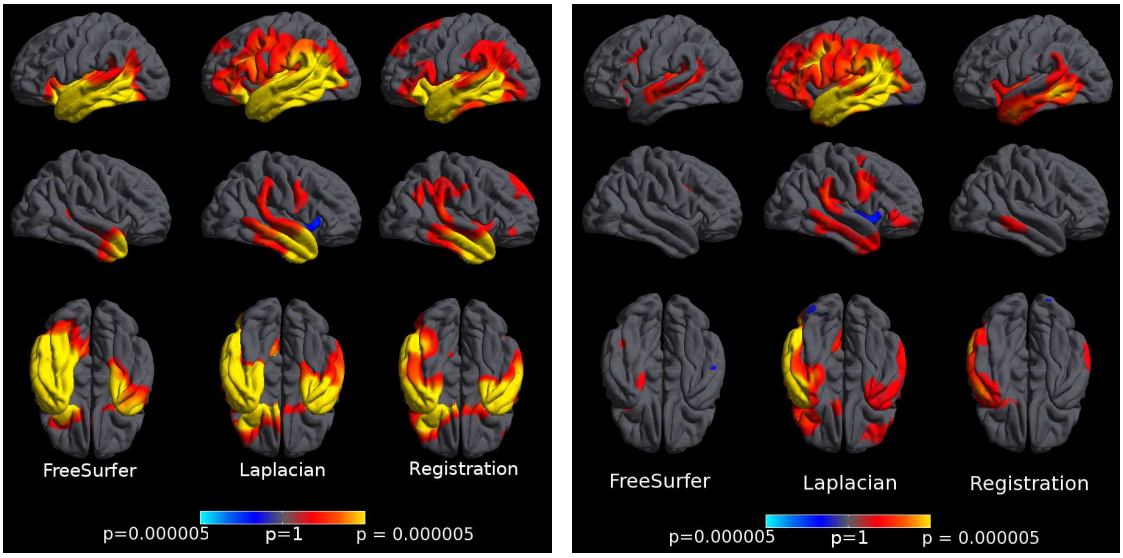


Figure 1: Reproducibility of FreeSurfer (left), Laplacian (middle) and Registration (right) based methods. The standard deviation of the difference in mean cortical thickness per region for two same day scans (n=49) is colour coded onto an average brain surface.



(a) Control versus SemD

(b) Control versus PNFA

Figure 2: A comparison of FreeSurfer, Laplacian and Registration based methods, displaying colour coded t-test p-values, comparing control subjects with SemD patients (left) and PNFA patients (right). Results are thresholded FDR corrected p-values < 0.05. Red to yellow indicates patients thinner than controls, and blue to light blue indicates patients thicker than controls.

Table 1: Subject Demographics for Cohort 1

Characteristic	Controls	AD
Number of subjects	16	33
Number of women (%)	8 (50)	14 (42)
Mean (SD) age at baseline (years)	72.5 (13.2)	72.1 (10.4)
Mean (SD) scan interval (days)	366 (6)	366 (18)

Table 2: Subject Demographics for Cohort 2

Characteristic	Controls	SD	PNFA
Number of subjects	28	43	30
Number of women (%)	17 (61)	26 (60)	21 (70)
Mean (SD) age at baseline (years)	66.4 (8.3)	63.8 (7.4)	66.2 (7.7)

Table 3: Atlas comparison: mean (standard deviation) of the regional cortical thickness in millimetres for the Laplacian method, where statistics are computed over regions defined by the FreeSurfer and also the AAL atlas. The third row shows p values of the t-tests (Pitman's tests).

Atlas	PHG	FUS	STG	PRE	SPG	SMG	LO	L	SFG
FreeSurfer	3.42 (0.39)	3.81 (0.32)	3.39 (0.32)	2.95 (0.38)	2.53 (0.31)	3.18 (0.38)	2.60 (0.41)	2.86 (0.40)	3.27 (0.30)
AAL	3.75 (0.38)	3.93 (0.31)	3.16 (0.40)	2.96 (0.37)	2.66 (0.35)	3.23 (0.41)	2.68 (0.43)	2.90 (0.40)	3.35 (0.32)
p-value	0.00 (0.78)	0.00 (0.67)	0.00 (0.00)	0.20 (0.41)	0.00 (0.00)	0.01 (0.10)	0.00 (0.26)	0.08 (0.98)	0.02 (0.52)

Table 4: Reproducibility: mean (standard deviation) of the difference in cortical thickness in millimetres per region for each of the three methods, over regions defined by the FreeSurfer atlas. An asterisk (*) indicates a Pitman's test p-value < 0.05 for that region when comparing the variance of the Laplacian method with the Registration method.

Method	PHG	FUS	STG	PRE	SPG	SMG	LO	L	SFG
FreeSurfer	0.00 (0.07)	-0.01 (0.05)	-0.01 (0.04)	-0.01 (0.05)	-0.02 (0.07)	-0.02 (0.04)	-0.01 (0.06)	-0.01 (0.04)	-0.02 (0.08)
Laplacian	-0.03 (0.20)	-0.04 (0.20)	-0.04 (0.17)	-0.05 (0.17)	-0.03 (0.14)	-0.05 (0.18)	-0.00 (0.14)	-0.02 (0.18)	-0.02 (0.22)
Registration	-0.03 (0.20)	-0.02 (0.18)	-0.02 (0.14)	-0.03 (0.15)	-0.04(0.16)	-0.05 (0.16)	-0.00 (0.15)	-0.02 (0.17)	0.00 (0.13*)

Table 5: Group comparison: t-test p-values (Cohen's d) for 9 left hemisphere regions, contrasting the control group with either SemD or PNFA patient groups, for each of the three methods. The value 0.00 indicates a p-value < 0.005

Method	Group	PHG	FUS	STG	PRE	SPG	SMG	LO	L	SFG
FreeSurfer	SemD	0.00 (-2.47)	0.00 (-2.62)	0.00 (-3.42)	0.01 (-0.61)	0.86 (-0.04)	0.02 (-0.52)	0.44 (-0.17)	0.00 (-0.72)	0.03 (-0.46)
Laplacian	SemD	0.00 (-2.44)	0.00 (-0.86)	0.37 (-0.22)	0.00 (-2.07)	0.03 (-0.51)	0.00 (-0.65)	0.28 (-0.25)	0.00 (-3.76)	0.00 (-1.44)
Registration	SemD	0.00 (-1.11)	0.22 (-0.29)	0.07 (0.47)	0.98 (-0.01)	0.50 (-0.17)	0.00 (-0.89)	0.44 (-0.18)	0.00 (-1.83)	0.00 (-0.78)
FreeSurfer	PNFA	0.79 (-0.07)	0.01 (-0.69)	0.00 (-1.14)	0.00 (-0.82)	0.09 (-0.46)	0.00 (-0.92)	0.55 (-0.16)	0.11 (-0.43)	0.00 (-0.97)
Laplacian	PNFA	0.00 (-1.15)	0.25 (-0.31)	0.12 (-0.42)	0.36 (-0.24)	0.00 (-0.81)	0.01 (-0.72)	0.96 (-0.01)	0.00 (-1.76)	0.00 (-1.34)
Registration	PNFA	0.06 (-0.52)	0.08 (0.48)	0.90 (0.03)	0.60 (-0.14)	0.54 (-0.17)	0.71 (-0.10)	0.14 (0.40)	0.01 (-0.76)	0.05 (-0.54)

Table 6: Group comparison: t-test p-values (Cohen's d) for 9 right hemisphere regions, contrasting the control group with either SemD or PNFA patient groups, for each of the three methods. The value 0.00 indicates a p-value < 0.005

Method	Group	PHG	FUS	STG	PRE	SPG	SMG	LO	L	SFG
FreeSurfer	SemD	0.00 (-1.10)	0.00 (-0.68)	0.00 (-0.90)	0.77 (-0.07)	0.50 (0.15)	0.86 (-0.04)	0.57 (0.13)	0.51 (-0.15)	0.83 (0.05)
Laplacian	SemD	0.00 (-0.96)	0.57 (-0.15)	0.15 (0.36)	0.00 (-0.92)	0.17 (0.30)	0.36 (-0.20)	0.86 (-0.04)	0.00 (-1.40)	0.00 (-0.74)
Registration	SemD	0.06 (-0.47)	0.27 (-0.26)	0.00 (0.73)	0.76 (-0.08)	0.27 (0.25)	0.02 (-0.61)	0.73 (-0.08)	0.00 (-1.24)	0.00 (-0.94)
FreeSurfer	PNFA	0.12 (0.42)	0.58 (0.15)	0.54 (-0.16)	0.05 (-0.54)	0.58 (-0.15)	0.12 (-0.42)	0.85 (0.05)	0.28 (-0.29)	0.07 (-0.48)
Laplacian	PNFA	0.21 (-0.34)	0.22 (-0.33)	0.78 (-0.07)	0.48 (0.19)	0.02 (-0.62)	0.04 (-0.55)	0.92 (-0.03)	0.02 (-0.64)	0.00 (-1.02)
Registration	PNFA	0.68 (0.11)	0.46 (0.20)	0.69 (0.11)	0.48 (-0.19)	0.19 (-0.35)	0.39 (-0.23)	0.08 (0.48)	0.68 (-0.11)	0.18 (-0.36)

Table 7: SVM classification: Results for 3 cortical thickness methods, distinguishing control subjects from SemD and PNFA patients.

Method	Group	Accuracy (%)	-CI (%)	+CI (%)
FreeSurfer	SemD	95.8	88.1	99.1
Laplacian	SemD	97.2	90.2	99.2
Registration	SemD	95.8	88.1	99.1
FreeSurfer	PNFA	79.3	66.6	88.8
Laplacian	PNFA	84.5	72.6	92.7
Registration	PNFA	75.9	62.8	86.1

Table 8: Longitudinal results: Mean (standard deviation) and effect size (Cohen's d) of the annualised percent change in cortical thickness for control and AD subjects, for each of the three methods.

Method	Group	PHG	FUS	STG	PRE	SPG	SMG	LO	L	SFG
FreeSurfer	Control	0.53 (4.43)	-0.08 (2.85)	-1.06 (2.92)	-1.51 (3.07)	-2.14 (3.68)	-0.91 (2.17)	-1.45 (2.94)	-0.46 (2.56)	-0.98 (3.42)
	AD	-3.36 (8.37)	-3.54 (4.12)	-3.70 (2.85)	-2.74 (4.37)	-2.22 (6.58)	-2.74 (3.53)	-2.62 (3.74)	-3.06 (5.02)	-3.27 (4.79)
	Effect	**-0.54**	**-0.94**	**-0.94**	**-0.31**	**-0.01**	**-0.59**	**-0.34**	**-0.60**	**-0.53**
Laplacian	Control	2.49 (6.94)	0.77 (6.03)	0.07 (5.39)	-1.60 (5.12)	-2.10 (5.95)	-0.18 (4.03)	-1.82 (7.28)	-1.13 (4.69)	-1.03 (7.19)
	AD	-0.74 (8.51)	-2.80 (5.73)	-4.03 (5.83)	-2.45 (5.13)	-1.99 (5.68)	-2.84 (7.96)	-1.14 (7.04)	-2.50 (5.89)	-3.13 (7.31)
	Effect	**-0.41**	**-0.63**	**-0.73**	**-0.17**	**0.02**	**-0.39**	**0.10**	**-0.25**	**-0.30**
Registration	Control	-0.46 (9.28)	-0.13 (5.56)	0.06 (3.62)	-0.93 (4.09)	-1.71 (6.03)	0.79 (4.21)	-1.01 (7.34)	-2.44 (4.01)	-0.17 (5.63)
	AD	-0.31 (4.37)	-0.25 (3.37)	-1.29 (5.04)	-1.47 (4.99)	-1.94 (5.78)	-1.31 (4.32)	-1.40 (6.53)	-1.77 (4.61)	-1.71 (5.23)
	Effect	**0.03**	**-0.03**	**-0.50**	**-0.18**	**-0.06**	**-0.73**	**-0.08**	**0.24**	**-0.42**