Can FreeSurfer Compete with Manual Volumetric Measurements in Alzheimer's Disease?

Lies Clerx[1,2], Ed H.B.M. Gronenschild[1,2,*], Carmen Echavarri[3], FransVerhey[1,2], Pauline Aalten[1,2] and Heidi I.L. Jacobs[1,2]

[1]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Center, Maastricht, Alzheimer Center Limburg, The Netherlands; [2]European Graduate School of Neuroscience (EURON), Maastricht University, Maastricht, The Netherlands; [3]San Miguel Clinic, Department of Neurology, 31006 Pamplona, Navarre, Spain

Ed H.B.M. Gronenschild

: Alzheimer's disease-related pathology results in tremendous structural and functional changes in the brain. These morphological changes might lead to a less precise performance of automated brain segmentation techniques in AD-patients, which in turn could possibly lead to false allocations of gray matter, white matter or cerebrospinal fluid. FreeSurfer has been shown to operate as an accurate and reliable instrument to measure cortical thickness and volume of neuroanatomical structures. Considering the principal role of FreeSurfer in the imaging field of AD, the present study aims to investigate the robustness of FreeSurfer to capture morphological changes in the brain against varying processing variables in comparison to manual measurements (the gold standard). T1-weighted MRI scan data were used pertaining to a sample of 53 individuals (18 healthy participants, 18 patients with mild cognitive impairment, and 18 patients with mild AD). Data were analyzed with different FreeSurfer versions (v4.3.1, v4.5.0, v5.0.0, v5.1.0), on a custom-built cluster (LINUX) and a Macintosh (UNIX) workstation. Group differences across versions and workstations were most consistent for both the hippocampus and posterior cingulate, regions known to be affected in the earliest stages of the disease. The results showed that later versions of FreeSurfer were more sensitive to identify group differences and corresponded best with the results of gold standard manual volumetric methods. In conclusion, later versions of FreeSurfer were more accurate than earlier versions, especially in medial temporal and posterior parietal regions. This development is very promising for future applications of FreeSurfer in research studies and encourages the future role of FreeSurfer output as a candidate marker in clinical practice.

: Alzheimer's disease, mild cognitive impairment, MRI, imaging, automated segmentation, FreeSurfer.

INTRODUCTION

Considering the increase of the aging population in our society and age being the greatest risk factor for the development of dementia, there is a growing interest in understanding and treating dementia. Currently, Alzheimer's disease (AD) is estimated to affect 35 million patients worldwide (or 0.5% of the global population) and this number is estimated to increase to 115 million by 2050 [1]. On a brain level, AD pathology results in excessive structural and functional damage, secondary to processes such as accumulation of amyloid-beta and tau proteins, neuroinflammation and neuronal death [2]. Structural and functional imaging measurements are currently evaluated for clinical use in predicting or diagnosing AD [3, 4]. An indispensable part of this effort is the development of a robust method to measure morphological and pathological changes in the brain [5]. Manual volumetric measurements are still regarded gold standard for evaluation of local brain atrophy [5, 6], however, clinical settings require diagnostic instruments that are quick, reliable and easy to implement. FreeSurfer (Athinoula A. Martinos Center for Biomedical Imaging, Boston) comprises a popular and freely available set of tools for deriving neuroanatomical volume and cortical thickness measurements by means of automated brain segmentation (http://surfer.nmr.mgh.harvard.edu). At present, more than 80[1] studies are published using FreeSurfer to investigate structural changes in the brain of (early) AD patients (Source: PubMed, http://www.ncbi.nlm.nih.gov/pubmed/).

Due to the variation in software and hardware environments, both in research and clinical practice, an equally important question related to measuring brain atrophy concerns the power and robustness of automated techniques to capture (small) morphological and pathological changes in the brain against these varying processing conditions. Since the pathological events seen in AD affect the morphology of the brain, it is conceivable that with such fundamental changes in brain

*Address correspondence to this author at the Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Alzheimer Centrum Limburg, Maastricht University Medical Center, P.O. Box 616, 6200 MD Maastricht, The Netherlands; Tel: +31 43 ;
Fax: +31 43 ; E-mail: ed.gronenschild@maastrichtuniversity.nl

[1] Pubmed search terms: (freesurfer) AND ((alzheimer*)OR(mild cognitive impairment))

1875-5828/15 $58.00+.00 © 2015 Bentham Science Publishers

structure, minor changes in processing conditions could affect the segmentation process and thus observed group differences.

In a previous study we systematically evaluated how the morphometric results derived from FreeSurfer may be affected by the following processing variables: FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation type (Macintosh and Hewlett-Packard), and Macintosch Operating System version (OSX 10.5 and OSX 10.6) [7]. Results revealed significant differences between FreeSurfer version v5.0.0 and the two earlier versions, ranging between 8.8 ± 6.6% (range 1.3 – 64.0%) (volume) and 2.8 ± 1.3% (1.1 – 7.7%) (cortical thickness). About twice as small differences were found between either the two workstation types or between OSX 10.5 and OSX 10.6. Even though our previous study investigated changes in healthy young individuals and psychiatric patients suffering from a psychotic disorder, these measurement differences were almost equal to the effect sizes reported in neurodegenerative studies [7]. Brain changes in these groups are less pronounced than those found in aging and neurodegenerative diseases. In view of the fact that FreeSurfer is extensively used in studies of age- ing, a next step is a validation of FreeSurfer in a neurode- generative population. The novelty in this study is that a comparison with manual volumetry is included as reference method, which is required in order to validate the accuracy of the segmentation results and thus subsequently detect which abnormalities are due to neurodegeneration and which can be classified as errors related to the process of automated segmentation.

Such validation studies require a comparison with manual segmentations or even post-mortem assessments (see [8] for a discussion on the limitations of both reference methods). The accuracy of the cortical thickness measures is more difficult to validate and requires mainly post-mortem (histological) measurements [9]. Only few studies directly compared FreeSurfer with manual volumetric measurements, mainly focusing on the medial temporal lobe in AD [10, 11], major depressive disorder [12], and temporal lobe epilepsy [13, 14]. These studies generally suggest that manual volumetry is slightly superior or equally sensitive to FreeSurfer automated volumetric measurements.

The aim of the present study is to investigate whether the sensitivity of FreeSurfer to detect group differences is consistent over different software versions and operating systems despite tremendous morphological changes typically seen in AD-patients. Stability of group differences across various processing conditions is investigated, and in order to understand which processing condition fits best with the 'reference' group differences, FreeSurfer's group differences are compared with manual volumetry, the gold standard in research practice [6, 15]. To the best of our knowledge, this is the first study assessing group differences across various processing conditions and validating these findings against gold standard measurements. Based on the literature, six well-established AD signature regions were chosen. Hippo- campal atrophy is known to play a major role in the devel- opment of AD [16], but is however not specific for AD [17]. Since other medial temporal lobe (e.g., parahippocampal gyrus)[18], prefrontal (e.g., inferior prefrontal and orbitofrontal cortex) [19] and posterior parietal regions (e.g. posterior cingulate and precuneus) [20] have shown to be altered during the disease process, these regions are additionally evaluated.

METHODS

Participants

Three groups of male participants were included: 18 healthy participants without any subjective memory impairment (CON), 18 patients with amnestic MCI (aMCI), and 18 patients with mild AD [18, 21]. Patients with MCI and mild AD were recruited from the Memory Clinic of the Maas- tricht University Hospital. Diagnosis was made according to the Petersen criteria for MCI (with at least an impairment in the memory domain) [22, 23], and the DSM-IV [24] and NINCDS-ADRDA criteria for AD [25]. The study was approved by the ethics committee of the Maastricht University Medical Center and all participants gave written informed consent in accordance with the committee's guidelines and with the Declaration of Helsinki [26].

MRI Acquisition

MRI scans were acquired with a 3T whole-body MR system release 2.0 (Philips Achieva, Philips Medical Systems, Best, The Netherlands) equipped with an eight-element head coil (SENSE, factor 2). Anatomical T1 images were acquired with a gradient echo sequence with TR = 8.0 ms, TE = 3.7 ms, FA = 8°, FOV = 240 x 240 mm^2, matrix = 240 x 240, number of slices = 180, voxel size = 1.0 x 1.0 x 1.0 mm^3.

FreeSurfer

Automated Volumetry

Cortical reconstruction and volumetric segmentation was performed with FreeSurfer, which is freely available (http://surfer.nmr.mgh.harvard.edu). The technical details of these procedures have been described previously (for a recent overview see: [9]) . Briefly, in this approach, brain areas are segmented using a nonlinear template matching [27]. After linearly registering the test data to the template, the algorithm estimates the nonlinear transformation between a given MRI and a probabilistic atlas of the selected brain structure constructed from a cohort of 40 subjects aged 19- 87 years using a maximum likelihood criterion [28]. Prob- abilistic labels are warped back to the individual MRI using the inverse of this transform. The final segmentation is ac- complished by maximizing the a posteriori probability in the Bayes formula at each voxel. Voxel-wise probabilistic labels and their predicted image intensities serve as the prior term, while the intensity similarity between the target image and the template serves as the likelihood term. In this study, both voxel and tabulated volumes (corrected for partial volume effects) were used. Important to note is that voxel volumes are most suited to obtain a proper comparison with manual volumetry, because of the absence of partial volume correc- tion.

Cortical Thickness (CT)

The FreeSurfer CT pipeline has been described and validated in previous publications [29-32]. To summarize, processing involved intensity normalisation, registration to Talairach space, skull stripping, segmentation of white matter (WM), tesselation of the WM boundary, smoothing of the tesselated surface, and automatic topology correction. The tesselated surface was 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 CT was 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 [33]. The cortex of the brain was automatically subdivided into gyral-based regions of interest (ROIs) [32]. To accomplish this, a registration procedure was used that aligns the cortical folding patterns and probabilistically assigns every point on the cortical surface to one of the 34 ROIs. For the purposes of this study, we focused on 6 ROIs bilaterally. For each ROI the mean cortical thickness was extracted for subse- quent statistical analysis. This technique is referred to as CT-parcellation.

For our second approach, a vertex-wise analysis, the thickness measures were mapped on a spherically inflated surface of each participant's reconstructed brain. This allows visualization across the surface without interference from cortical folding. By means of a combination of linear and non-linear transformations, the spherical cortical folding pat- terns were aligned to a spherical template provided by Free- Surfer. This technique, called "surface-based intersubject registration" [34], provides an accurate matching of morphologically homologous cortical locations across participants on the basis of each individual's anatomy while minimizing metric distortions. The resulting CT map was smoothed by a circularly symmetric Gaussian filter with a full width half maximum (FWHM) set to 20 mm in order 1) to compensate for residual misalignments; 2) to increase the signal-to-noise ratio; and 3) to make the data more normally distributed. This technique is referred to as CT-vertex henceforth.

Quality control was performed after each step of the automated FreeSurfer pipeline (volumetry and cortical thickness) to account for possible errors (e.g., misregistrations, outliers). No manual editing was carried out to ensure a valid analysis.

Manually Defined ROIs

For manual tracing of the ROIs we used GIANT (General Image Analysis Tools; [34]), a software program which allows ROI-tracing in a triplanar and rotatable 3D surface-rendered view, and hence calculation of GM volumes of interest. Boundaries of the selected frontal and temporal structures were set according to criteria described in a previous publication [35]. Boundaries of the posterior cingulate and precuneus cortex were adapted from [36] and [37], respectively. Both raters (LC, CE) were blind to the demographic and cognitive characteristics of the participants. Intra-rater reliability was determined by the Intraclass Correlation Coefficient (ICC) [38]. ICC's for each region can be found in a previous publication [39] (Supplementary Table **S1**).

ROI's Selected from the Desikan Atlas

For the purpose of this study, we used the Desikan atlas and focused on one subcortical ROI and five cortical ROIs bilaterally: the hippocampus (HIPPO), the parahippocampal gyrus (PhG), the inferior prefrontal cortex (IPFC), the orbital frontal cortex (OPFC), the posterior cingulate cortex (PCC), and the precuneus (PC). See Appendix for more details on which FreeSurfer ROIs [28]) were selected and merged. FreeSurfer ROIs were chosen in accordance with the anatomical borders of the manually defined ROIs: e.g., FreeSurfer's definition of the isthmus was most consistent with our definition of the posterior cingulate cortex as adopted for the manual segmentation and was subsequently used. The posterior cingulate cortex ROI included in the FreeSurfer automated measurement is more rostral compared to our definition.

Intracranial Volume

The ICV was calculated from the visually checked inner skull contours produced by the FSL Brain Extraction Tool [36, 40-42].

Processing Conditions

Workstations

Several workstations and corresponding operating systems (OS) were used: an iMac with OSX 10.5, a MacPro with OSX 10.6 (called "iMac2" and "MacPro2" in Gronen- schild *et al.* 2012, respectively), and a custom-built cluster equipped with Intel i7 quad-cores (3.20 GHz) running under Scientific Linux 6.2 (called "RadCluster" henceforth). Both Macintosh (Mac) workstations were used in 32 bits mode and the RadCluster in 64 bits mode.

Software Versions

Four versions of FreeSurfer were used: v4.3.1, released on 19 May 2009; version v4.5.0, released on 11 August 2009; version v5.0.0, released on 16 August 2010; version v5.1.0, released on 24 May 2011. For the Mac workstations these are 32 bits versions, whereas for the RadCluster these are 64 bits versions. An additional remark with respect to v5.1.0 should be made: we used an intermediate version of the processing pipeline in order to resolve issues around the order of the correction for intensity non-uniformity stage and Talairach stage in the pipeline (see also https://surfer.nmr.mgh.harvard.edu/fswiki/TalFailV5.1).

Statistical Analysis

Group analysis of the segmentation results was performed in two ways. In the first approach, analysis was performed with IBM SPSS Mac version 19 (Chicago, IL, USA). FreeSurfer-based volumes, manual volumes and CT-parcellations were compared between the three groups by means of univariate pair-wise ANCOVA with either volume or CT-parcellation as dependent variable, group as independent variable and centered age as covariate. For both volumetric measurements, intracranial volume (ICV) was taken as an additional covariate. To correct for multiple comparisons, we applied the false discovery rate (FDR) controlling procedure [43].

Our second approach was a vertex-wise analysis of CT using FreeSurfer tools. Statistical comparisons between the CT maps were generated by computing a general linear model (GLM) of the CT group differences (corrected for centered age) at each vertex in the cortical mantle, with a statistical threshold set to $p = 0.05$. A cluster-wise procedure was performed to correct for multiple comparisons [44]. This method utilizes a simulation to get a measure of the distribution of the maximum cluster size under the null hypothesis. Z-maps are synthesized and smoothed using a residual FWHM, and then thresholded at p = 0.05. Next, areas of maximum clusters are recorded, under these specifications, and the procedure is repeated for 5000 iterations. Once the distributions of the maximum cluster size are obtained, correction for multiple comparisons is achieved by finding clusters in the statistical maps using the same threshold as was set during the simulation procedure. For each cluster, the p- value is the probability of perceiving a maximum cluster of that size, or larger, during the simulation. Clusters remaining in similar areas of significance as in the original cortical thickness maps would imply that the result is not likely due to chance. For each cluster, maximum, minimum, mean, and standard deviation of the p-values were extracted[2].

To quantify the differences of the results of the above vertex-wise analyses across FreeSurfer versions or workstations, the measure of spatial overlap (Dice coefficient, [45]) of the respective corrected clusters was computed. Its range is between 0 (no overlap) and 1 (complete overlap, i.e., exactly similar). It is generally accepted that a value larger than 0.7 indicates a good agreement [46].

RESULTS

One individual in the mild AD group was excluded because of FreeSurfer processing errors. The three groups significantly differed with respect to age, Mini-Mental state examination (MMSE) score, and score on the delayed recall task, but not with respect to educational level (Table **1**).

A complete overview of the comparison CON vs. AD for each selected ROI is illustrated in Fig. (**1**) (the group comparisons CON vs. MCI and MCI vs. AD are shown in the supplemental material, Figs. (**S1-S2**) respectively). Each cell is color-coded according to its *p*-value after correction for multiple comparisons. In case of CT-vertex we have taken the minimal *p*-value in each ROI.

Robustness of FreeSurfer Across Workstations and Software Versions

Generally, it can be noted that FreeSurfer derives more significant results for later versions in case of volumetric measurements (either voxel or tabulated). In addition, the CT-vertex method produces consistent results through versions as well as workstations, especially for MTL regions. With respect to cross-workstation differences, it can be observed that MacOSX 10.6 is most similar to RadCluster, in particular for v5.1.0[3]. For both voxel and tabulated volumetric measurements, group differences between workstations and software versions were most consistent in hippocampal and posterior cingulate regions (for all 3 group comparisons). Group differences in the PhG, OPFC R and PC L were only found in FreeSurfer v5.0.0 and/or v5.1.0, across all workstations. The CT-vertex results for MacOSX 10.6 and all FreeSurfer versions are illustrated in Fig. (**2**) for the left hemisphere (comparison CON vs. AD, both corrected and uncorrected results are displayed). Cortical thinning (negative effects, blue colored) was mainly observed in temporal and parietal cortical areas. Cortical thickening (positive effects, red-yellow colored) was apparent in the frontal lobe only for versions v4.5.0 and v5.1.0. For the other versions, these positive clusters did not survive the correction for multiple comparisons. For the comparison CON vs. MCI, no significant clusters were found.

Table 1. Subject characteristics.

	CON	MCI	AD
N	18	18	17
Age [a,c]	64.56 (3.4)	65.11 (4.5)	70.59 (9.1)
Educational level	4 (1.4)	4 (1.8)	4 (1.9)
MMSE score [b,d,f]	28.89 (0.9)	27.61 (2.3)	21.18 (3.9)
15 WLT learning[c,de]	37.50 (7.6)	26.06 (9.8)	23.47 (11.7)
15 WLT memory [c,d,e]	8.56 (1.9)	3.67 (2.8)	1.73 (2.4)
Fluency animals [a]	23 (5.3)	21 (5.4)	13.93 (4.7)
Manual hippocampus volume L/R (mm^3) [a]	4656 (308)/ 4758 (637)	4410 (482)/ 4308 (796)	3883 (817)/ 3807 (876)
ICV (ml)	1492 (100)	1539 (121)	1574 (125)

All volumetric measurements are corrected for intracranial volume. Values are mean (sd). MMSE: Mini-Mental state examination; WLT: wordlist; CON: controls; MCI: mild cognitive impairment; AD: Alzheimer's disease.
[a] $p < 0.05$ for difference between CON and AD; [b] $p < 0.05$ for difference between CON and MCI; [c] $p < 0.05$ for difference between MCI and AD; [d] $p < 0.001$ for difference between CON and AD; [e] $p < 0.001$ for difference between CON and MCI; [f] $p < 0.001$ for difference between MCI and AD

The Dice coefficients for the agreement of the CT surface clusters (positive and negative effects taken together) are summarized in Tables **2** and **3**. In most cases the agreement was good to excellent, and better between workstations than between FreeSurfer versions. In order to detail these findings, worst and best agreements for the cross-workstation comparisons are shown in Fig. (**3**). Vertices depicted in green indicate the presence of the clusters in the results of both workstations. Red and yellow denote clusters found in only one of the respective workstations. The corresponding Dice coefficients were 0.68 and 0.97, respectively, both related to CON vs. AD comparisons. A complete disagreement was found for the PCC (see left medial view). Similarly, the worst and best results for the cross-version comparisons are shown in Fig. (**4**), with corresponding Dice coefficients of 0.58 and 0.94, respectively, also both concerning CON vs. AD comparisons. A disagreement was found in the frontal lobe, and PCC (see left medial view). Note that both for

[2] Because of an error in the cluster correction tool of FreeSurfer v4.3.1, we applied the cluster correction tool of v4.5.0 to the results of v4.3.1, see the release notes, http://surfer.nmr.mgh.harvard.edu/fswiki/ReleaseNotes. Henceforth we refer to these results as v4.3.1*

[3] During time of writing we have processed the data with MacOSX 10.7 for FreeSurfer versions v4.3.1, v5.0.0, and v5.1.0. The results were identical to those of MacOSX 10.6.