Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation?

Heath R. Pardoe*,†, Gaby S. Pell*,†, David F. Abbott*,†, and Graeme D. Jackson*,†,‡

* Brain Research Institute, Florey Neuroscience Institutes (Austin), Melbourne, Victoria, Australia
† Department of Medicine, The University of Melbourne, Victoria, Australia
‡ Department of Radiology, The University of Melbourne, Victoria, Australia

SUMMARY

Purpose—Quantitative measurement of hippocampal volume using structural magnetic resonance imaging (MRI) is a valuable tool for detection and lateralization of mesial temporal lobe epilepsy with hippocampal sclerosis (mTLE). We compare two automated hippocampal volume methodologies and manual hippocampal volumetry to determine which technique is most sensitive for the detection of hippocampal atrophy in mTLE.

Methods—We acquired a three-dimensional (3D) volumetric sequence in 10 patients with left-lateralized mTLE and 10 age-matched controls. Hippocampal volumes were measured manually, and using the software packages Freesurfer and FSL-FIRST. The sensitivities of the techniques were compared by determining the effect size for average volume reduction in patients with mTLE compared to controls. The volumes and spatial overlap of the automated and manual segmentations were also compared.

Results—Significant volume reduction in affected hippocampi in mTLE compared to controls was detected by manual hippocampal volume measurement (p < 0.01, effect size 33.2%), Freesurfer (p < 0.01, effect size 20.8%), and FSL-FIRST (p < 0.01, effect size 13.6%) after correction for brain volume. Freesurfer correlated reasonably (r = 0.74, p ≪ 0.01) with this manual segmentation and FSL-FIRST relatively poorly (r = 0.47, p ≪ 0.01). The spatial overlap between manual and automated segmentation was reduced in affected hippocampi, suggesting the accuracy of automated segmentation is reduced in pathologic brains.

Discussion—Expert manual hippocampal volumetry is more sensitive than both automated methods for the detection of hippocampal atrophy associated with mTLE. In our study Freesurfer was the most sensitive to hippocampal atrophy in mTLE and could be used if expert manual segmentation is not available.

Keywords

MRI; Hippocampal volumetry; Hippocampal sclerosis

Hippocampal volumetry is a widely used clinical tool for the detection and lateralization of mesial temporal lobe epilepsy (Jack et al., 1990; Cascino et al., 1991; Cook et al., 1992; Cendes et al., 1993a; Jackson et al., 1993). Traditionally, the volume of the hippocampus is measured by manually segmenting the hippocampus on serial sections of a T1-weighted
magnetic resonance imaging (MRI) scan acquired perpendicular to the long axis of the hippocampus (Jack et al., 1990; Watson et al., 1992). The manual approach requires a trained operator to segment the hippocampus in a reliable and consistent manner. The measured hippocampal volumes tend to have a high interrater variability and, to a lesser extent, have some intersession variability (Niemann et al., 2000).

Recent developments in automated software-based segmentation now allow us to use MRI to obtain estimates of hippocampal volume without requiring manual input (Fischl et al., 2002; Patenaude, 2007). These techniques potentially address interrater and intersession sources of variance in hippocampal volume assessment. A number of automated and semi-automated computational techniques have been used to detect hippocampal volume changes in mTLE (Hogan et al., 2000, 2004; Keller et al., 2002; Bonilha et al., 2004; Chupin et al., 2007; Hammers et al., 2007; McDonald et al., 2008; Pell et al., 2008; Bonilha et al., 2009); however, few of these studies explicitly compared automated volume estimates with the corresponding manual estimate in the same subject (Chupin et al., 2007; Hammers et al., 2007). In the studies in which a direct comparison of manual and automated techniques was made (Chupin et al., 2007; Hammers et al., 2007), an automated segmentation technique was used that was different from the two publicly available methods presented in this article.

The aim of this study is to determine if automated hippocampal segmentation methods are suitable for detection of hippocampal atrophy in a mesial temporal lobe epilepsy cohort with unilateral hippocampal sclerosis (mTLE). We also aim to compare the relative sensitivity of the automated and manual techniques. The outcomes of this study will help determine whether automated techniques should replace the manual technique as the standard hippocampal segmentation methodology for the detection and lateralization of hippocampal atrophy in a clinical setting.

**Methods**

The automated and manual techniques were compared in two ways. The first was in terms of the magnitude of the volume difference between the affected hippocampus in mTLE and control hippocampal volumes. This comparison allowed us to determine which technique was most sensitive for the detection of mTLE and indicated whether automated segmentation could be used instead of manual segmentation for clinical assessment in mTLE. The second analysis investigated similarity between the automated and manual segmentations in terms of the measured volumes and spatial overlap of the automated and manual hippocampal segmentations. These volume and spatial overlap analyses allowed us to investigate the agreement between the different segmentation methodologies, and determine if the accuracy of the automated segmentation techniques were affected by structural abnormalities in the mTLE group.

**Subjects and MRI acquisition**

Ten patients [mean age 40.4 ± 12.3 standard deviation (SD) years, six women] with drug-refractory temporal lobe epilepsy associated with left unilateral hippocampal sclerosis were studied. Diagnosis was determined in pre-surgical investigations carried out by the Comprehensive Epilepsy Program at Austin Health, Australia. Investigations include electroencephalography (EEG), video monitoring, 1.5 T magnetic resonance (MR) scanning and neuropsychological evaluation. The patients were recruited as a consecutive series from the surgical program in our tertiary epilepsy center. Postoperative confirmation of the diagnosis was confirmed in all cases using histopathology. Patients were compared with 10 healthy controls (mean age 40.9 ± 12.5 years, five women). The study was approved by the Austin Health Human Research Ethics Committee. All research MR imaging was performed on a 3 T GE LX Horizon scanner (General Electric, Milwaukee, WI, U.S.A.). The structural

*Epilepsia. Author manuscript; available in PMC 2011 March 10.*
scan was a T_1-prepared 3D spoiled gradient recalled (SPGR) acquisition (echo time, TE = 2.7 ms; repetition time TR = 13.8 ms; flip angle = 20°; inversion time TI = 500 ms; voxel size 0.48 × 0.48 × 2 mm).

**Hippocampal volumetry**

Hippocampi were manually delineated (by author HP) using ImageJ software (version 1.41e, Rasband, 2008; http://rsbweb.nih.gov/ij/), following the protocol detailed in Watson et al. (1992). The manual segmentations were undertaken blind to subject classification (mTLE/control). The volume of the hippocampus was measured by summing the area of the hippocampus in each coronal slice. Hippocampi were automatically labeled using (1) the subcortical segmentation image processing stream provided in the Freesurfer software distribution (version 3.0.5, Fischl et al., 2002; http://surfer.nmr.mgh.harvard.edu/) and (2) the subcortical segmentation routines “FIRST,” provided as part of the FSL software distribution (version 4.1.2, Patenaude, 2007; http://www.fmrib.ox.ac.uk/fsl/).

For the Freesurfer-based segmentation, the MR image was processed using the default analysis settings. The primary error-checking undertaken was confirmation of a correct registration of the image to Talairach space (Talairach & Tournoux, 1988), and visual inspection to confirm correct labeling of the hippocampus. Both the Talairach transform and hippocampus were correctly identified in all subjects.

For the FSL-FIRST–based segmentation, the user is able to set the number of modes of variation for the hippocampal template to be warped to fit the individual hippocampi. The optimal number of modes of variation for the detection of hippocampal volume differences between mTLE and control hippocampi were investigated by systematically varying the number of modes of variation from 10–300. The optimal number of modes was selected based on maximizing the effect size for the mTLE/control difference (expressed as a percentage volume reduction in mTLE subjects compared to controls). Similar to the Free-surfer-based automated segmentation, the initial step for FSL-FIRST was an affine registration of each brain to MNI-152 space (Mazziotta et al., 2001). The correct affine registration was visually confirmed in all cases. Each manual and automatically segmented hippocampus was saved as an inclusive binary mask in the same space as the original image.

**Brain volume measurement**

The measured hippocampal volumes were adjusted for head size by dividing each hippocampal volume estimate by the brain volume. For each subject in this study, the brain volume was measured by applying the software tool “BET” (brain extraction tool, Smith, 2002), provided as part of the FSL software distribution, to each T_1-weighted MR scan, and summing the number of extracted voxels to provide an estimate of the total brain volume. Each hippocampal volume estimate was reported as a percentage fraction of the total brain volume for each subject.

**Data analysis**

Manual hippocampal volumetry is historically the standard technique for assessment of hippocampal volume loss in mTLE (Jack et al., 1990; Cascino et al., 1991; Cook et al., 1992; Cendes et al., 1993a). In the absence of an independent measure of hippocampal volume, it may be considered the de facto “gold standard.” To determine in an unbiased fashion whether this remains the preferred method, the estimated effect size associated with volume differences between the mTLE and control hippocampal volumes were measured using the three methods. That is, the relative sensitivity of each technique for the detection of mTLE was assessed for a population known to have the condition. The correlation
between each automated method and the manual estimate was used to quantify the similarity between the measured volumes using each automated technique.

The spatial overlap of the hippocampal segmentations for each technique was measured using the Dice coefficient (Dice, 1945). If \( N(A) \) and \( N(B) \) represent the number of voxels labeled as hippocampus using technique A and technique B, respectively, and \( N(A + B) \) represents the number of voxels labeled as hippocampus by both techniques, the Dice coefficient in this case is given by

\[
d = \frac{2N(A+B)}{N(A)+N(B)}
\]

Thus \( d \) can range from 0 (when there are no common voxels) to 1 (when the two methods yield identical segmentations). The variability of the Dice coefficient across the subjects in a group indicates how consistent the segmentation methods are for that group. By comparing the Dice coefficient for the two automated methods, we can see if certain hippocampi are difficult for either technique to segment relative to the manual segmentation.

**Classification of atrophy in individual subjects**

In our research center, we classify an epilepsy subject as having significantly reduced hippocampal volume if their hippocampal volume is <1.65 SDs below the mean ipsilateral hippocampal volume of a representative control group. A hippocampal volume below this cut-off is outside the one-tailed 95% confidence interval (CI) (90% confidence interval in the case of a two-tailed distribution) of a standard normal distribution with the same mean and SD as our control group. The use of a one-tailed interval is justified because numerous previous studies have demonstrated volume reduction in mTLE as opposed to volume difference (see previously cited references).

We defined a cutoff point for each of the three methods using the preceding definition and classified each mTLE subject as a false negative if it had a left hippocampal volume above the defined threshold, that is, within the control range. Using this definition of significant atrophy will result in a false-positive rate of 5%, that is, 5% of controls would be found to have significant hippocampal atrophy according to our threshold.

**Results**

The manual and automated segmentations of the hippocampi in an example mTLE subject are shown in Fig. 1. It can be seen that both automated estimates tend to label more tissue as hippocampus than the manual technique. The optimal number of modes for detecting volume differences between mTLE subjects and controls using FSL-FIRST was found by plotting the effect size, expressed as the percentage volume reduction in mTLE subjects compared to controls, versus number of modes (see Supporting Information). The maximum difference in volume occurs when the number of modes is set at 50. Throughout the rest of this article any reference to FSL-FIRST hippocampal volume estimates refers to estimates generated using a number of modes of 50.

The distribution of measured hippocampal volumes for each segmentation technique demonstrates that both automated techniques give higher estimates of hippocampal volume than the manual segmentation technique (Fig. 2). The expected volume decrease in the affected (left) mTLE hippocampi can be observed in the volume distribution for the manual assessment (Fig. 2, left graph), but the volume difference between affected mTLE
hippocampi and the contralateral and control hippocampi is reduced with the automated methods (Fig. 2, center and right graph).

Analysis of volume differences between controls and mTLE subjects in the left hippocampi for each technique confirms that manual segmentation has a greater effect size for mTLE hippocampal volume loss than either automated method (Table 1). The greater effect size is indicated by the relative volume reduction (compared to controls) in left mTLE hippocampi of 33.2% for the manual method, compared to 20.8% for the Freesurfer technique and 13.6% for the FSL-FIRST technique. Therefore, of the automated segmentation methods, the Freesurfer-based technique is more sensitive to hippocampal atrophy in mTLE than the FSL-FIRST method.

The similarity between the volume estimates for the manual segmentation and the automated techniques can be investigated by regressing each automated hippocampal volume estimate against the manual estimate. For both techniques there is a significant linear relationship between the automated and manual estimates (Fig. 3. Freesurfer: $r = 0.74$, $p \ll 0.01$, FSL-FIRST: $r = 0.47$, $p \ll 0.01$). The increased spread of the data around the regression line for the FSL-FIRST technique (Fig. 3, right) indicates that there is less agreement between this technique and the manual estimate compared to the Freesurfer analysis. Analysis of the spatial agreement between the automated and manual methods, as quantified using the Dice coefficient, shows that segmentation accuracy is reduced in the affected hippocampi in the mTLE group (Table 2, $p < 0.01$ for each control/mTLE comparison).

The use of manual hippocampal segmentation to classify significant atrophy in the mTLE group on a case-by-case basis results in fewer false negatives than either automated segmentation method (Table 3). Freesurfer-based segmentation also shows greater discrimination than FSL-FIRST. This result follows as a consequence of the greater separation between the distributions of measured volumes for the manual method (Fig. 2).

**Discussion**

In this study we have tested the ability of conventional manual segmentation and two automatic hippocampal segmentation methods to determine disease-specific atrophy of the hippocampus using whole-brain T₁-weighted MRI scans. Structural changes associated with mTLE have been well characterized in numerous previous studies (Jack et al., 1990; Cook et al., 1992; Cendes et al., 1993a; Jackson et al., 1993; Van Paesschen et al., 1995; Marsh et al., 1997; Kuzniecky, 1998; Mackay et al., 2000; Keller et al., 2002; Bernasconi et al., 2003; Bonilha et al., 2004; Townsend et al., 2004; Hammers et al., 2007; Pell et al., 2008), and so this subject group is a good model system for testing the relative sensitivity of the manual and automated methods. It is also important to test automated segmentation methods in pathologic brains in which structural abnormalities may not be confined to the structure of interest (Briellmann et al., 1998, 2004).

The data presented in this article indicate that manual hippocampal segmentation is more sensitive than both automated segmentation techniques. Of the automated approaches, the Freesurfer-based hippocampal segmentation method was more sensitive to hippocampal atrophy in mTLE than the FSL-FIRST method. A recent study (Morey et al., 2009) investigated the similarity between hippocampal segmentation using Freesurfer and FSL-FIRST and manual segmentation in a control group and confirmed our findings that Freesurfer-based hippocampal segmentation exhibits higher spatial overlap and correlation with manual volume estimates than FSL-FIRST.

The impact of the relative difference in sensitivities may be illustrated by using the distribution of control hippocampal volumes measured using the three techniques to classify
significant hippocampal volume loss in mTLE subjects (Table 3). The use of a less sensitive automated technique such as FSL-FIRST decreases our ability to detect significant hippocampal volume loss on a patient-by-patient basis. This means that if only the volume was relied on, clinically significant hippocampal atrophy would be missed in only one case with manual segmentation, three cases using Freesurfer, and seven cases using FSL-FIRST (Table 3).

Although not explicitly addressed in this study, it is worth considering whether volumetry provides any improvement over visual inspection of MR scans to detect hippocampal atrophy in patients with mTLE. Severe mTLE with hippocampal sclerosis can usually be detected by a trained neurologist or neuroradiologist with visual inspection (Jackson et al., 1993), but there are three circumstances in which volumetry can improve upon (Cendes et al., 1993b; Reutens et al., 1993) or aid visual inspection. (1) In the case of bilateral hippocampal atrophy—with visual inspection the reader tends to use the contralateral hippocampus as a reference for estimating the volume of a hippocampus. In the case of bilateral hippocampal atrophy this reference becomes invalid. (2) When the hippocampal volume changes are subtle. (3) Volumetry can be used as a training aid to establish visual thresholds for calling hippocampal atrophy in centers that are not expert in epilepsy diagnosis.

An alternative approach to manual and fully automated methods for hippocampal segmentation is semi-automated hippocampal segmentation. Typically these involve the manual identification of landmarks to aid in the identification of the hippocampus, and then utilize software-based analysis of intensity variability or higher-dimensional shape parameters to segment the hippocampi. A previous study has shown that these techniques can reduce the effects of inter- and intraoperator variability associated with manual hippocampal segmentation in mTLE (Hogan et al., 2000).

Comparison of the overlap of the automated techniques with the manual segmentation shows that there is greater disagreement between manual and both automated techniques in the mTLE group, indicating that the automated methods do not perform as well in the presence of pathology in the patient group. This is consistent with a previous report utilizing a different automated segmentation methodology (Hammers et al., 2007). The relatively low dispersion of the overlap of the automated segmentations and the manual segmentation (3–6% for controls, 6–11% for mTLE) at least indicate that the automated techniques are correctly identifying the structure of interest in each case.

It is likely that the accuracy of the automated and manual segmentation methods, and the agreement between them, is dependent on the spatial resolution of the MRI scans. In this study we observed that the agreement between the automated and manual techniques was lower at the posterior end of the hippocampus, where the hippocampal tail curves in a superior–inferior direction. The anisotropic voxel size (2 mm through-plane resolution) means partial volume effects will be worse in this part of the hippocampus, making it difficult to reliably distinguish between the hippocampus and surrounding tissue. Similarly at the anterior end of the hippocampus a higher through-plane resolution makes it easier to distinguish between the hippocampal head and the amygdala. In terms of the clinical utility of automated hippocampal segmentation methods versus manual segmentation for diagnosis and lateralization of mTLE, the results of this study suggest that manual hippocampal segmentation should still be used as the standard technique for clinical assessment of hippocampal volume.
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by Program Grant 400121 from the National Health and Medical Research Council (NH&MRC), Australia. Heath Pardoe has salary part funded by Grant R37NS31146 (June 1, 2007–May 31, 2011) from NIH-NINDS, USA. We confirm that we have read the Journal’s position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References


Figure 1.
A comparison of automated and manual hippocampal segmentation methods. The figure shows posterior (top row), intermediate (middle row), and anterior (bottom row) sections through the long plane of the hippocampus in an example mTLE subject. The red overlay (2nd column) shows the manual segmentation, the blue overlay (3rd column) shows the Freesurfer-based segmentation, and the green overlay (4th column) shows the FSL-FIRST-based segmentation.
Figure 2.
Hippocampal volume estimates in mTLE compared to controls and the contralateral hippocampus. Manual segmentation (left graph) generates lower volumes compared to Freesurfer (middle graph) and FSL-FIRST (right graph), and there is greater separation between the affected hippocampus (mtle lhipp) in mTLE and unaffected hippocampi. The range of measured volumes for each segmentation method and subject group is tabulated in Supporting Information.
Figure 3.
The relationship between the manual hippocampal volume and the Freesurfer-based (left, $r = 0.74$, $p \ll 0.01$) and FSL-FIRST (right, $r = 0.47$, $p \ll 0.01$) hippocampal volume estimates in left and right hippocampi in controls and mTLE subjects. Blue squares represent ipsilateral mTLE, red triangles represent contralateral mTLE, green circles represent control right, and orange diamonds represent control left hippocampi. The solid line shows a line of best fit for all the hippocampal volumes; the dashed line shows a 1:1 relationship, indicating perfect agreement between the automated and manual techniques.
Table 1
A comparison of left hippocampal volume changes in subjects with mTLE and age-matched controls

<table>
<thead>
<tr>
<th>Method</th>
<th>p-Value</th>
<th>Effect size (% volume reduction)</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual segmentation</td>
<td>$1.18 \times 10^{-5}$</td>
<td>33.2</td>
<td>3.07</td>
</tr>
<tr>
<td>Freesurfer</td>
<td>$1.26 \times 10^{-4}$</td>
<td>20.8</td>
<td>2.29</td>
</tr>
<tr>
<td>FSL-FIRST</td>
<td>$5.81 \times 10^{-3}$</td>
<td>13.6</td>
<td>1.49</td>
</tr>
</tbody>
</table>

The p-value is associated with a two-way Student’s t-test of differences in volume between controls and mTLE subjects in the affected hippocampus. The quoted effect size is the percentage volume reduction in subjects with mTLE. Cohen’s d indicates the strength of the observed effect (0.2 is a small effect, 0.5 is a medium effect, and >0.8 is a large effect).
Table 2
Overlap (Dice coefficient) between different hippocampal segmentation methods in controls and left mTLE hippocampi

<table>
<thead>
<tr>
<th>Method Combination</th>
<th>Controls (mean ± SD)</th>
<th>mTLE (mean ± SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual/Freesurfer</td>
<td>0.73 ± 0.028</td>
<td>0.66 ± 0.042</td>
</tr>
<tr>
<td>Manual/FSL-FIRST</td>
<td>0.71 ± 0.046</td>
<td>0.62 ± 0.057</td>
</tr>
<tr>
<td>Freesurfer/FSL-FIRST</td>
<td>0.72 ± 0.037</td>
<td>0.67 ± 0.042</td>
</tr>
</tbody>
</table>

The overlap is reduced in affected mTLE hippocampi, indicating less agreement between manual and automated methods (p < 0.05 for each control/mTLE comparison).
### Table 3

Classification of significant atrophy in mTLE subjects using manual and automated hippocampal segmentation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Threshold (% brain vol)</th>
<th>Number of false negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>Freesurfer</td>
<td>0.26</td>
<td>3</td>
</tr>
<tr>
<td>FSL-FIRST</td>
<td>0.27</td>
<td>7</td>
</tr>
</tbody>
</table>

The threshold for significant atrophy is defined as the mean control volume $-1.65$ standard deviations. The number of false negatives is the number of mTLE subjects with a hippocampal volume greater than the defined threshold.