Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study

Alexander Hammers, a,b,⁎ Rolf Heckemann, a,c Matthias J. Koepp, a,b John S. Duncan, b Jo V. Hajnal, c Daniel Rueckert, d and Paul Aljabar d

a MRC Clinical Sciences Centre and Division of Neuroscience, Faculty of Medicine, Imperial College London, Hammersmith Hospital, DuCane Road, London, UK
b Department of Clinical and Experimental Epilepsy, Institute of Neurology, UCL, Queen Square, London, UK, and National Society for Epilepsy MRI Unit, Chalfont St Peter, UK
c Imaging Sciences Department, MRC Clinical Sciences Centre, Imperial College London, Hammersmith Hospital, DuCane Road, London, UK
d Department of Computing, Imperial College London, UK

Received 1 December 2006; revised 8 February 2007; accepted 26 February 2007
Available online 24 March 2007

In temporal lobe epilepsy (TLE), hippocampal atrophy (HA) is a marker of poor prognosis regarding seizure remission, but predicts success of anterior temporal lobe resection. Manual quantification of HA on MRI is time-consuming and limited by investigator availability. Normal ranges of hippocampal volumes, both in absolute terms and relative to intracranial volume, and of hippocampal asymmetry were defined using an automatic label propagation and decision fusion technique based on thirty manually derived atlases of healthy controls. Manual test–retest reliability and overlaps of automatically and manually determined hippocampal volumes were quantified with similarity indices (SIs). Correct clinical identification of ipsilateral HA, and contralaterally normal hippocampal volumes, was determined in nine patients with histologically confirmed hippocampal sclerosis in terms of volumes and asymmetry indices (AIs) for standard statistical thresholds and with receiver operating characteristic (ROC) analysis. Manual test–retest reliability was very high, with SIs between 0.87 and 0.90. Manual and automatic hippocampus labels overlapped with a SI of 0.83 on the unaffected but with 0.76 on the atrophic side. Accuracy was higher for less atrophic hippocampi. The automatic method correctly identified 6/9 HAs in terms of absolute volume, 7/9 in terms of relative volume at a standard 2 SD threshold, and 9/9 for AIs. ROC-determined thresholds allowed clinically desirable correct identification of all HAs (100% sensitivity) with 85–100% specificity for volumes, and 100% specificity for AIs. The method has the potential to automatically detect unilateral HA, but further work is needed to determine its performance in detecting clinically important bilateral disease.

© 2007 Elsevier Inc. All rights reserved.

⁎ Corresponding author. Room 243, Cyclotron Building, Department of Clinical Neuroscience, Imperial College London, Hammersmith Campus, London W12 0NN, UK. Fax: +44 20 8383 1783.
E-mail address: alexander.hammers@csc.mrc.ac.uk (A. Hammers).
Available online on ScienceDirect (www.sciencedirect.com).

Introduction

Temporal lobe epilepsy (TLE) is the most common of the focal epilepsies, and hippocampal sclerosis (HS) is the most common underlying cause of medically refractory TLE. Detecting HS is clinically important as these patients have a 60% chance of becoming seizure free with surgery (Engel, 1996; Wiebe et al., 2001).

HS is readily detected using magnetic resonance imaging (MRI) and characterized by hippocampal atrophy (HA), i.e. volume loss, and increased T2 relaxation times (Van Paesschen et al., 1997a). HA at the time of diagnosis predicts poorer medical prognosis and may not be visually evident but only volumetrically detectable when mild (Van Paesschen et al., 1997b; Semah et al., 1998; Salmenpera et al., 2005). HA is a predictor of seizure recurrence if antiepileptic drugs are discontinued in seizure-free adults (Cardoso et al., 2006). On the other hand, HA on MRI predicts HS on histological examination (Lenz et al., 1992; Van Paesschen et al., 1997c) and indicates a better surgical prognosis (e.g. Jack et al., 1992).

The ability to detect HA and other structural abnormalities depends heavily on both the quality of the MRI data and the training and experience of the interpreting radiologist (Von Oertzen et al., 2002). Marked asymmetric HA with a ratio of smaller to larger hippocampus of less than 70% is readily detected by experts, while lesser degrees of hippocampal atrophy generally require quantification. This is performed manually on serial MRI slices through the hippocampus, with volumes derived using the Cavalieri principle (Free et al., 1995). Disadvantages include the need for trained operators, workstations and software, as well as substantial time commitments (approximately 60 min per dataset). Consequently, quantification is usually restricted to tertiary centres, and many potentially surgically treatable HAs are initially missed (Von Oertzen et al., 2002).
An automated and reliable way of measuring hippocampal volumes in TLE is, therefore, highly desirable, but the issue has received surprisingly little attention. An early investigation of a deformable model/region growing technique requiring only one manual initialization outline on a sagittal slice (Ashton et al., 1997) lateralized asymmetries of a large volume-of-interest (VOI) containing the hippocampus to the side of seizure onset in 8/9 patients with refractory temporal lobe epilepsy. It is unclear whether these patients had HA on MRI, normal control ranges were not derived, nor were physiological asymmetries (Pedraza et al., 2004) taken into account. The manual method used yielded lateralizations opposite to the automatic method in three and opposite to the seizure focus in two of the nine cases. Another study used tissue intensity within a standard mean hippocampal VOI derived from 30 controls which had been automatically transformed by global and regional affine registration (Webb et al., 1999). The authors aimed for 100% specificity; the resulting sensitivity was approximately 60%. No measures of spatial overlap between automatic and manual delineations were reported. Another study applied a semi-automated single-brain atlas warping method, combined with the manual selection of 20 landmarks, to five patients with HA (Hogan et al., 2000). There was difficulty in obtaining a manual gold standard for sclerotic hippocampi, but overlap between manual inter-rater segmentations was comparable to overlap between the semiautomatic and the less variable of two manual segmentations.

In this paper, we apply a recently established method for automatic definition of regional brain volumes on clinical T1-weighted MRI images of the head (Heckemann et al., 2006) using multiple anatomical atlases (Hammers et al., 2003, 2007) combined with automatic tissue classification (Ashburner and Friston, 2005) and apply this to a cohort of patients with HS. We verify the results obtained by reference to manual segmentations of the hippocampi and evaluate the ability of the method to correctly identify pathological volumes as well as the clinical yield of the method.

Materials and methods

Patients and controls

We retrospectively studied nine patients (five women; median age of 38 years, range 22–49, 36±9 years [mean±SD]) with unilateral HS diagnosed using qualitative and quantitative MRI, confirmed histologically after an anterior temporal lobe resection and extensively characterized in other studies (Koepp et al., 1996, 1998; Hammers et al., 2001, 2005). Out of the original cohort of 15 patients, original untreated MRI data for nine patients was available for this proof-of-principle study; some datasets were unavailable for trivial technical reasons.

Controls were 30 subjects (15 women; median age of 31 years, range 20–54, 31±8 years [mean±SD]; difference was not significant: Student’s t-test, p>0.1) scanned with the same protocol on the same machine and used in the creation of atlases of human neuroanatomy (Hammers et al., 2002, 2003, 2007; Wang et al., 2005; Heckemann et al., 2006).

All subjects were Caucasians. All data were acquired between May 1995 and September 1999, with patients’ and controls’ scans intermixed. Prior to scanning, written informed consent and the approval of local Ethics Committees were obtained.

MRI image acquisition and preprocessing

MRI scans were acquired on the 1.5 T GE Signa Echospeed scanner at the National Society for Epilepsy. A coronal T1-weighted 3D volume was obtained using an inversion recovery prepared fast spoiled gradient recall sequence (GE), TE/TR/NEX of 4.2 ms (fat and water in phase)/15.5 ms, time of inversion (TI) of 450 ms, flip angle 20°, yielding 124 slices of 1.5 mm thickness with a field of view of 18×24 cm for a 192×256 matrix, covering the whole brain with voxel sizes of 0.9375×0.9375×1.5 mm. Non-uniformity correction was performed using Sled’s method (N3; Sled et al., 1998). The images were re-oriented with the horizontal line defined by the anterior and posterior commissures (AC–PC orientation) and the sagittal planes parallel to the midline (Mitchell et al., 2003). Images were resliced to create isotropic voxels of 0.9375×0.9375×0.9375 mm³ using windowed sinc interpolation to preserve the native resolution. We used the fully automatic tissue class segmentation algorithm contained within the Statistical Parametric Mapping package (SPM5, Wellcome Trust Centre for Neuroimaging at UCL, London) to segment MRI datasets into probabilistic images of grey matter, white matter and CSF (Ashburner and Friston, 2005). This information was not used during the manual delineation, but used post hoc to extract the grey matter component of the manually or automatically defined hippocampi by thresholding the grey matter maps above 50% probability and multiplying this thresholded map with the hippocampal maps. The rationale for this thresholding was the discovery, during pilot studies, that for hippocampal sclerosis patients, the automatic algorithm occasionally captured the boundary between CSF and white matter in the temporal horn, rather than the boundary between hippocampus and CSF. Hippocampal volumes may vary with brain or intracranial size, and both brain and hippocampal size may be affected by acquired disease. We corrected hippocampal volumes for total intracranial volumes, using a fully automatic algorithm validated specifically for this purpose (Exbrain; Lemieux et al., 1999, 2003).

Manual definition of hippocampi

All manual delineations of hippocampi were performed according to the same protocol previously described and validated in terms of intrarater and interrater intraclass correlation coefficients for volumes (Niemann et al., 2000; Hammers et al., 2003). In brief, a detailed anatomical algorithm for manual delineation was derived from textbooks, atlases and own histological slices (Niemann et al., 2000), and volumetry performed in the region-of-interest module of the Analyze software (Robb, 2001). Traing was performed on all relevant coronal slices with a mouse-controlled cursor, with boundaries displayed in real time on the MRI slices and also displayed in the remaining orthogonal slices. Training required about 60–90 min per dataset (for two hippocampi). Hippocampi in the atlas datasets were delineated by two raters from earlier studies [R.A. (Hammers et al., 2003) and C.H.C. (Hammers et al., 2007)] and all reviewed by A.H. All hippocampi in the patient datasets (n=9×2 sides, five of which were used twice for the test–retest study) were outlined by A.H., as were the five randomly chosen control test–retest datasets (n=5×2 sides×2 time points). A.H. has co-developed the method (Niemann et al., 2000) and has over 10 years’ experience in hippocampal volumetry. R.A. and C.H.C. had been extensively trained by A.H. Both have an estimated >1000 h of experience with manual anatomical segmentation.
In order to establish manual test–retest variation as a gold standard upper bound as a benchmark for automatic labelling, we randomly selected five controls and five patients and re-measured hippocampal volumes after a minimum of 2 weeks. Test–retest variation was evaluated as volume differences (retest minus test volumes, divided by the average of test and retest volumes, expressed as a percentage). Measures of overlap are more important to judge the veracity of a segmentation than volumes. We chose the similarity index (SI), defined as the number of voxels labelled as hippocampus in both the test and retest measurement, divided by the average of voxels labelled as hippocampus in the test or retest measurement (Zijdenbos et al., 1994). The SI is also known as the Dice similarity coefficient (Dice, 1945) and compares to another measure of overlap in frequent use, the overlap ratio (OR) or Jaccard coefficient of community (coefficient de communauté; Jaccard, 1907), as \( OR = SI/(2 – SI) \). The SI and OR range from zero for no overlap to unity for complete overlap.

**Automatic definition of hippocampi**

We used an atlas propagation and label fusion technique described previously in healthy volunteers (Heckemann et al., 2006) to automatically derive hippocampal volumes. In brief, each MRI of each of the 30 healthy controls was automatically warped to each of the other 29 controls and to each of the nine patients. We used an algorithm exploiting normalized mutual information as a voxel-based similarity measure and a staged process whereby global parameters are described as an affine transformation while local MRI intensity correspondence is described by a free-form deformation based on B-splines (Rueckert et al., 1999). This algorithm results in a high-dimensional warping of the source to the target brain.

The MRI warping parameters were then used to transform the associated manually derived hippocampal atlases (Hammers et al., 2003) onto the new target MRIs, yielding 30 independent estimates of hippocampal VOIs per patient hippocampus and 29 independent estimates of hippocampal VOIs per control hippocampus in target space. These were then combined by using the mode of labels (hippocampus versus non-hippocampus), with a random decision in the case of nonunique modes, and finally thresholded with the 50% grey matter map (Ashburner and Friston, 2005). All these steps are fully automatic but require approximately 11 h of CPU time per atlas warping.

**Statistical analysis**

We first evaluated test–retest reliability as described above.

We then gathered descriptive statistics of hippocampal volumes and compared percentage volume differences and measures of overlap of automatically versus the first manually obtained volume estimates with the same SI formula as described above for test–retest data.

We then calculated the yield or detection rate for patients (i.e. how many patients are correctly classified as having HA ipsilateral to the epileptogenic side as defined electroclinically and with manual hippocampal segmentation, and contralaterally normal hippocampal volumes) and the false positive rate for controls (i.e. how many controls were incorrectly classified as having HA), for a threshold of 2 and 2.5 SD from the control mean volume.

Besides the volume differences, we also determined asymmetry indices (AI), another frequently clinically used quantification method, from the hippocampal volumes, which we defined as right minus left hippocampal volume, divided by their average.

Finally we performed a Receiver Operator Characteristic (ROC) curve analysis to determine optimum cut-off values for our sample.

Statistical analysis was performed using SPSS version 12 for Microsoft Windows.

**Results**

**Manual test–retest reliability**

The average percentage volume difference from first to second manual measurement of hippocampal volume in the five randomly selected controls was 2.1 ± 5.0% (range, –2.8% to 8.8%) for the right and 1.6 ± 5.1% (range, –4.2% to 7.4%) for the left hippocampus (Table 1).

Average SIs were 0.90 ± 0.01 (0.88–0.92) for the right and 0.89 ± 0.02 (0.86–0.92) for the left hippocampus in controls.

In the five randomly selected patients, these statistics were minimally lower with 0.89 ± 0.02 (0.86–0.92) for the right and 0.87 ± 0.02 (0.85–0.90) for the left hippocampus (or, in expression in relation to pathology, 0.88 ± 0.02 (0.85–0.90) for the ipsilateral and 0.89 ± 0.03 (0.86–0.92) for the contralateral hippocampus).

The average percentage volume difference from first to second manual measurement in patients was –7.2 ± 7.9% (range, –20.5% to –0.5%) for the right hippocampus and –7.4 ± 9.9% (range, –16.9% to 4.2%) for the left hippocampus.

**Table 1**

<table>
<thead>
<tr>
<th>Number</th>
<th>Right volume 1</th>
<th>Right volume 2</th>
<th>% Difference</th>
<th>R SI</th>
<th>Left volume 1</th>
<th>Left volume 2</th>
<th>% Difference</th>
<th>L SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1792</td>
<td>1783</td>
<td>–0.5</td>
<td>0.89</td>
<td>1445</td>
<td>1538</td>
<td>6.2</td>
<td>0.88</td>
</tr>
<tr>
<td>C16</td>
<td>1925</td>
<td>2103</td>
<td>8.8</td>
<td>0.90</td>
<td>1834</td>
<td>1853</td>
<td>1.0</td>
<td>0.90</td>
</tr>
<tr>
<td>C17</td>
<td>1832</td>
<td>1914</td>
<td>4.4</td>
<td>0.90</td>
<td>1767</td>
<td>1902</td>
<td>7.4</td>
<td>0.88</td>
</tr>
<tr>
<td>C18</td>
<td>2052</td>
<td>1996</td>
<td>–2.8</td>
<td>0.92</td>
<td>2068</td>
<td>1983</td>
<td>–4.2</td>
<td>0.92</td>
</tr>
<tr>
<td>C26</td>
<td>2064</td>
<td>2070</td>
<td>0.3</td>
<td>0.88</td>
<td>1866</td>
<td>1824</td>
<td>–2.3</td>
<td>0.90</td>
</tr>
<tr>
<td>P2 (rHS)</td>
<td>1174</td>
<td>1168</td>
<td>–0.5</td>
<td>0.89</td>
<td>1678</td>
<td>1744</td>
<td>3.9</td>
<td>0.86</td>
</tr>
<tr>
<td>P4 (iHS)</td>
<td>2249</td>
<td>2146</td>
<td>–4.7</td>
<td>0.92</td>
<td>1128</td>
<td>1105</td>
<td>–2.1</td>
<td>0.88</td>
</tr>
<tr>
<td>P6 (rHS)</td>
<td>1437</td>
<td>1400</td>
<td>–2.6</td>
<td>0.90</td>
<td>2202</td>
<td>2133</td>
<td>–3.2</td>
<td>0.90</td>
</tr>
<tr>
<td>P8 (rHS)</td>
<td>1566</td>
<td>1274</td>
<td>–20.6</td>
<td>0.86</td>
<td>1993</td>
<td>1682</td>
<td>–16.9</td>
<td>0.86</td>
</tr>
<tr>
<td>P9 (iHS)</td>
<td>2484</td>
<td>2305</td>
<td>7.5</td>
<td>0.91</td>
<td>1154</td>
<td>957</td>
<td>–18.7</td>
<td>0.85</td>
</tr>
</tbody>
</table>

C/P, control/patient (number of the subject chosen randomly for the five test–retest studies); rHS/iHS: right/left hippocampal sclerosis; % differences: percentage differences, \((\text{Volume 2} – \text{Volume 1})/(\text{Volume 2} + \text{Volume 1})/2\)* 100; SI = similarity index (number of voxels labelled as hippocampus in both the test and retest measurement, divided by the average of voxels labelled as hippocampus in the test or retest measurement).
to 3.8%) for the left hippocampus or, expressed in relation to pathology, −8.9±9.9% (range, −20.5% to −0.5%) for the ipsilateral and −5.7±7.5% (range, −16.9% to 3.8%) for the contralateral hippocampus. The test–retest volumes were not significantly different on the right or left (paired t-test, P > 0.3). When all patients’ and all controls’ test–retest differences were considered together, the average percentage volume test–retest difference was significantly higher in patients (−7.3±8.4%) than in controls (1.8±4.6%) (Student’s t-test, p < 0.01), suggesting that pathological hippocampi may be more difficult to quantify.

Automatic volume estimates and bias in relation to manual estimates

In controls, mean automatically determined hippocampal GM volumes were 2318±279 mm³ on the right and 2072±235 mm³ on the left. This right-over-left volume difference of 11% was highly significant (paired t-test, p < 0.001) and very close to that found when outlining hippocampi manually with the same protocol used for the atlases (9% in the current cohort) and with the same protocol in a different cohort of 20 males (12%; Niemann et al., 2000). Therefore, patient hippocampi had to be evaluated separately by side and by side of sclerosis. In patients, the automatically estimated unaffected (contralateral) hippocampal volumes differed by 0.9±10% in volume from their control counterparts (Fig. 1). The automatic method detected substantially smaller hippocampi on the affected side compared with their healthy counterparts; reductions were by an average of 30% on the right and 41% on the left.

Comparing automatic estimates with the manual gold standard, the true degree of hippocampal atrophy on the ipsilateral side was underestimated by an average of 17±11% (Fig. 1).

Similarly, while the SIs measuring overlap between automatically derived and manually derived hippocampi were close to control values on the unaffected (contralateral) side (0.83±0.03 (0.80–0.87) in patients versus the previously established control values (Heckemann et al., 2006) of 0.83±0.04 (0.71–0.89) (right) and 0.81±0.04 (0.70–0.87) (left), they were lower on the side ipsilateral to the sclerosis (0.76±0.04, 0.71–0.83). These SI values in patients were also positively correlated with manually derived hippocampal volumes (Pearson’s r=0.70, p < 0.001), indicating that hippocampi closer to control volumes were easier to quantify automatically.

An example of automatic and manual labels is shown in Fig. 2.

Clinical detection rates: volumes and fractional volumes

Individual values and detection are listed in Table 2.

The cut-off values for normality derived from the automatically determined absolute volumes for controls (compare Fig. 1) were 1760 mm³ for the right and 1602 mm³ for the left hippocampus at the 2 SD threshold; at the 2.5 SD thresholds these were 1620 mm³ for the right and 1485 mm³ for the left hippocampus.

This led to correct detection of 4/9 hippocampal atrophies (two each on right and left side) at the 2.5 SD threshold, with 6/9 detected at the 2 SD threshold (two additional left HA detected; for details see Table 2). The undetected HAs were the less severe ones judged from manual segmentation (equally uncorrected for ICV). All nine contralateral hippocampal volumes were correctly identified as normal, i.e. there were no false positives in the patients. The false positive rate in controls, i.e. the proportion of controls incorrectly identified as having HA, was 0/60 values (i.e. for 30 right and 30 left hippocampi) at 2.5 SD for an expected false positive rate of 0.7/60 and 1/60 at 2 SD for an expected false positive rate of 2.7/60.

There was a significant correlation between intracranial volumes and the volumes of both the right hippocampus (Pearson’s r=0.70, p < 0.001) and the left hippocampus (Pearson’s r=0.75, p < 0.001). We therefore expected an increase in sensitivity following narrowing of the normal range by relating absolute hippocampal volumes to intracranial volume. Indeed the detection rate in patients rose to 5/9 for the 2.5 SD threshold and to 7/9 for the 2 SD threshold. All nine contralateral hippocampal volumes were again correctly identified as normal, i.e. there were no false positives in the patients. The corresponding false positive rates in controls remained 0/60 for the 2.5 SD threshold and 1/60 for the 2 SD threshold.

Clinical detection rates: asymmetry indices

Relating hippocampal size to the contralateral hippocampus offers another means of standardising a measurement, and AIs are routinely used for clinical purposes. A very important limitation is their inability to detect bilateral disease; missing such contralateral pathology may have deleterious consequences. Pearson’s correlation coefficient between automatically determined right and left hippocampal volumes in controls was r=0.87 (p < 0.001), i.e. even higher than the correlation between hippocampal volumes and intracranial volumes, and we therefore expected a further increase in sensitivity.
The normal range for the AI was 0.00 to 0.22 at 2 SD and −0.03 to 0.25 at 2.5 SD, with the distance from zero reflecting the physiological rightward asymmetry.

All 9/9 subjects were automatically detected as having hippocampal volume asymmetry, with the ipsilateral side correctly identified as being smaller, at both the 2 SD and the 2.5 SD thresholds. Mean ±SD absolute $Z$-scores from control mean were 5.8 ±2.0 (range, 3.2–9.1, Table 2). There was one false positive AI among the 30 controls at both thresholds.

**ROC analysis**

Finally, we performed a ROC analysis to determine optimum sensitivities and specificities. Absolute and fractional volumes as well as AIs had areas under the curve close to unity (area under the curve 0.971–1; $p < 0.002$ in all cases). For a screening test like MRI in suspected HA, a 100% sensitivity is desirable, and we chose our cut-off to reflect this.

For automatically determined hippocampal volumes, the cut-off for normality at 100% sensitivity was 1914 mm$^3$ (1.4 SD) for the right and 1563 mm$^3$ (2.2 SD) for the left hippocampus. At that cut-off, specificity was 97% for right and 100% for left hippocampal volumes.

Fractional volume cut-offs for 100% sensitivity were 1.47 (0.8 SD) for the right and 1.16 (2.4 SD) for the left hippocampus, yielding specificities of 85% on the right and 100% on the left.

For AIs, cut-offs for 100% sensitivity were −0.03 (2.5 SD) indicating right HA and 0.33 (3.9 SD) indicating left HA. Those cut-offs yielded 100% specificity.

**Discussion**

The main finding of this study is the possibility of detecting unilateral hippocampal atrophy in temporal lobe epilepsy on standard clinical T1-weighted MRI images in an automated fashion.
Table 2
Individual values and detection as pathological/normal with standard thresholds for significance

<table>
<thead>
<tr>
<th>Patient number</th>
<th>Ipsilateral</th>
<th>Contralateral</th>
<th>Asymmetry</th>
</tr>
</thead>
</table>

| R Control | 2525±367 | 2318±279 | 1.58±0.14 |
| L Control | 2072±235 | 2318±279 | 1.58±0.14 |
| 1/rHS/m | 1615 | −2.4 | 1902 | −1.5 | 1.24 | −2.5 | 16 | 2268 | −0.2 | 2533 | 2.0 | 1.65 | 2.1 | 11 | −0.28 | 6.6 |
| 2/rHS/f | 1174 | −3.7 | 1560 | −2.7 | 1.31 | −2.0 | 28 | 1678 | −2.0 | 1807 | −1.1 | 1.52 | 0.9 | 7 | −0.15 | 4.3 |
| 3/lHS/m | 1061 | −3.9 | 1283 | −3.4 | 0.83 | −5.3 | 19 | 2783 | 0.7 | 2530 | 0.8 | 1.63 | 0.4 | −10 | 0.65 | 9.1 |
| 4/lHS/f | 1128 | −3.6 | 1533 | −2.3 | 1.16 | −2.4 | 31 | 2249 | −0.7 | 2336 | 0.1 | 1.76 | 1.3 | 4 | 0.41 | 5.1 |
| 5/lHS/m | 1263 | −3.2 | 1530 | −2.3 | 1.04 | −3.5 | 19 | 2074 | −4.2 | 2290 | −0.1 | 1.55 | −0.2 | 10 | 0.40 | 4.8 |
| 6/lHS/f | 1437 | −2.9 | 1779 | −1.9 | 1.23 | −2.6 | 21 | 2202 | −0.4 | 2150 | 0.3 | 1.48 | 0.6 | −2 | −0.19 | 5.0 |
| Mean±SD | 1321±208 | −3.2±0.5 | 1564±253 | −2.5±0.8 | 1.16±0.21 | −2.6±1.2 | 17±11 | 1996±19 | −0.6±0.8 | 2181±267 | −0.2±1.2 | 1.60±0.08 | 1.0±0.7 | −1±10 | N/A | 5.8±2.0*

Difference manual/automatic: Volumes (automatic−manual)/(automatic+manual)/2, i.e. positive values indicate larger volumes obtained automatically. The same applies to volumes corrected for intracranial volume (ICV-corr). Values preceded by “±” are SDs. Z score: Number of SD from control mean volume. P: patient number; GM: grey matter; r/lHS: right/left hippocampal sclerosis.

1 Abnormal at 2 SD.
2 Abnormal at 2.5 SD from control mean.
3 Mean of absolute values.
with at least 85% specificity for 100% sensitivity. In addition, we show high reliability of our manual measurements of normal and atrophic hippocampal volumes.

**Methodological considerations**

The prior knowledge atlases for our general purpose method have been derived from control data only (Heckemann et al., 2006). There was a positive correlation between SIs for overlap of automatically and manually derived hippocampal labels on the one hand and hippocampal volumes on the other hand. This indicates that larger hippocampi (those closer to the volumes used as the normative atlas database) may be segmented more reliably, i.e. closer to the manual gold standard. The implications are twofold. Firstly, automatically derived hippocampal volumes and AIs may be most reliable when HA is subtle; at the same time this is the situation in which visual inspection will perform worst (Van Paesschen et al., 1995) and clinical need is greatest. Secondly, in order to improve accuracy for more severely atrophic hippocampi, the next logical step will be to incorporate prior information from manually outlined atrophic hippocampi into the *a priori* atlas knowledge base. We hypothesize that this will increase accuracy for pathological hippocampi, and aim to apply the method to a larger clinical cohort including bilateral HA cases.

Prior methods for volumetry of healthy hippocampi have usually relied on manual delineation. For semiautomatic methods, aside from differences in accuracy there are major differences in the amount of manual preprocessing and required operator time. Methods achieving good overlap tend to have high demands on an expert operator, with up to about 400 landmarks having to be placed per brain (Shen et al., 2002). A dependency on accurate placement of landmarks is usually found (Hogan et al., 2000; Bueno et al., 2001; Shen et al., 2002) with relatively large differences between different runs of the procedure. A notable exception is a recent method based on homotopically deforming regions and dual competitive growth, only requiring the definition of a bounding box and the positioning of one seed for the hippocampus and one for the amygdala (Chupin et al., 2007). Compared with a manual gold standard of similar reliability to ours, the method yielded SIs of 0.84 for hippocampi of both healthy controls as well as subjects with Alzheimer’s disease, but a limitation is that it does not yield other volumes than those of hippocampus and amygdala. This and our approach could in principle be combined.

Methods which are automatic (excluding trivial manipulations like reorientation along the axes of a co-ordinate system) often rely on prior knowledge in the form of atlases. Comparatively poor measures of overlap are achieved when single-subject-derived atlases are used (e.g. SI of 0.68±0.10 for the overlap of manual and automatic hippocampal regions in 19 elderly controls (Barnes et al., 2006), see experimental confirmations in (Heckemann et al., 2006; Wang et al., 2005) or when overlap is hampered by overly inclusive standard hippocampal atlases [SI 0.65–0.72 for up to three sets of 10 healthy controls each, ANIMAL (Collins et al., 1995) and a new related method (Duchesne et al., 2002)].

Better results in healthy controls were achieved with a statistical shape model based on 30 males used to drive and restrict its elastic warping to the target in 21 male subjects for the left hippocampus with SIs of ~0.75±0.11; however, there were outliers (SI range, ~0.39–0.86) (Kelemen et al., 1999). An atlas encoding intensity information as well as localization and neighbourhood information, linearly transformed to each of seven healthy volunteers in turn in a leave-one-out procedure, yielded SIs of ~0.80 for both inter-rater manual and automatic versus manual measurements of a large (~5000 mm³) hippocampal VOI (Fischl et al., 2002).

Our method shows very good manual test–retest results which are within the best conceivable range of accuracy derived from phantom validation studies (Jack et al., 1990) reflecting the detailed nature of the protocol. Excellent measures of manual test–retest overlap were maintained in atrophic hippocampi. This is important as manual segmentation can only be considered a gold standard if it provides reliable results. It can then be used as a benchmark for assessing automatic methods. The high quality of our benchmark may have led to an overestimation of the performance of our automatic labeling method: It is conceivable that other labeling strategies are comparable or better, but use inferior prior knowledge which limits their performance.

Note that different protocols will lead to different absolute volumes obtained in any one study (Pedraza et al., 2004). Protocols must therefore be published (Bergin et al., 1994), and comparison with a normal range obtained in the same fashion, as in our study, is mandatory.

The only prerequisite for the current method is a broadly comparable spatial orientation of the images, which can be ensured during acquisition. In this instance, we adjusted the patient images to have isotropic voxels like the control datasets and corrected for nonuniformity, but these steps are not essential. Besides this minimal preprocessing, all steps are user-independent.

A significant drawback of our current method is the demand for 30±10.7±5.2 (6.0–37.7) hours of CPU time per dataset to be studied, which in our case was met by using a cluster of Linux PCs (Heckemann et al., 2006). Computing demands have led to pessimistic assessments of a method’s suitability in the past. However, here we have shown a proof of principle for automatically extracting clinically meaningful information from MRI images. It is reasonable to predict that the continuing progress in processing capacity (http://en.wikipedia.org/wiki/Moore’s_Law) will overcome this drawback in due course. Large speed increases could also be achieved by parallelizing the spatial transformation software, by focusing on one area of the brain, or by down-sampling image resolution. Another strategy which we are currently investigating is the prediction of segmentation accuracy for single atlas–target pairs. If factors predicting accuracy can reliably be identified, they may allow to choose a smaller number of atlas classifiers for a given subject.

**Clinical considerations**

Some authors, concerned about false positives in controls, aimed at 100% specificity (Webb et al., 1999). We share concerns that this means avoiding a clinical situation that does not exist in practice (Duchesne et al., 2006). An automated volume detection tool should not be used indiscriminately in asymptomatic subjects, but when the diagnosis of TLE has already been established electroclinically. In this situation, 100% specificity is important to indicate the possibility of considering a presurgical evaluation with good chance of success, and we adjusted our ROC cut-offs accordingly.

Hippocampal volumes were highly correlated with intracranial volumes. In line with previous studies (reviewed by Geuze et al., 2006), see experimental confirmations in (Heckemann et al., 2006), computing demands have led to pessimistic assessments of a method’s suitability in the past. However, here we have shown a proof of principle for automatically extracting clinically meaningful information from MRI images. It is reasonable to predict that the continuing progress in processing capacity (http://en.wikipedia.org/wiki/Moore’s_Law) will overcome this drawback in due course. Large speed increases could also be achieved by parallelizing the spatial transformation software, by focusing on one area of the brain, or by down-sampling image resolution. Another strategy which we are currently investigating is the prediction of segmentation accuracy for single atlas–target pairs. If factors predicting accuracy can reliably be identified, they may allow to choose a smaller number of atlas classifiers for a given subject.
2005), our data therefore confirms a rationale for using volumes corrected for ICV. In view of the possibly widespread involvement of brain structures particularly in the temporal lobe and particularly ipsilaterally (e.g. Bernasconi et al., 2003), correcting for ICV seems more prudent than using cerebral volumes, even though a previous study found a slightly higher correlation with cerebral volume (Free et al., 1995). It should be noted, however, that even with 30 controls to define the normal range, simple thresholding methods (e.g. 2 SD below control mean) did not achieve the desired 100% sensitivity; only 7/9 patients were correctly identified, and only the ROC-derived optimal thresholds led to correct identification and lateralization of all 9 HAs. If standard thresholds are in routine use, this limitation should be kept in mind.

The correlation between right and left hippocampi was even more marked than between ICV and hippocampal volume, and accordingly AIs achieved the best diagnostic yields, reflecting clinical experience. Note that in our cohort, selected to show unilateral HA based on manual volumetry, all contralateral hippocampi were correctly identified as normal, which will have aided the high diagnostic yield of AIs with very high Z scores (Table 2). In clinical practice, bilateral HA may lead to normal AIs, but this normality is deceiving and failure to recognise bilateral pathology prior to surgery can have serious consequences in terms of seizures continuing (Garcia et al., 1994) or amnesia developing (Baxendale, 1998). Patients with normal AIs due to bilaterally normal hippocampal volumes (Jack et al., 1992) have a lower chance of surgical success, and patients with normal AIs due to bilaterally abnormal hippocampal volumes are difficult surgical candidates (Jack et al., 1992, 1995; King et al., 1995). The AI can therefore still be useful as a screening tool, with abnormal AIs pointing towards potential good surgical candidates and normal AIs – whether due to bilaterally normal or abnormal hippocampi – indicating more difficult cases.

An important point is the need to consider the physiological hippocampal asymmetry favouring the right side (Pedraza et al., 2004). This is likely to be dependent on the inclusion or exclusion of mesial temporal substructures and thereby protocol-dependent. With our protocol we find right hippocampi about 9% (current cohort) to 12% (Niemann et al., 2000) larger than those on the left. Disregarding this asymmetry favours detection of left-sided atrophy but hampers detection of right-sided atrophy, as for example in (Ashton et al., 1997).

Our results confirm the rationale for using manual measures as a gold standard, with very high overlap (SI ~0.9) achieved for manual-to-manual intrarater comparisons, allowing meaningful comparisons. In order to evaluate the performance of automatic labeling methods, however, the reliability of manual measurements must be demonstrated in pathology as well. Here, this was the case with manual-to-manual SIs of the atrophied hippocampi statistically indistinguishable from those in healthy controls, but such reliability is more difficult to achieve in disease (Hogan et al., 2000).

**Neurobiological considerations**

The current method focuses on deriving hippocampal volumes and AIs alone as a proof of principle. It has long been known that the hippocampus is not the only anatomical entity exhibiting alterations in TLE (Shorvon, 2006). Our method can be extended to other structures that may be affected by atrophy in TLE, for example amygdala, parahippocampal gyrus and its subdivisions (Bernasconi et al., 2003), temporal pole (Moran et al., 2001), to temporal horn enlargement, or white matter atrophy. Such associated features are present in a large proportion of patients (Meiners et al., 1994) and have recently been successfully used to classify TLE patients overall (Duchesne et al., 2006). Our approach, while offering high accuracy in terms of automatic volumetric assessment, can be extended to incorporate further volumetric criteria, particularly for regions already present in the normative atlases (Hammers et al., 2003). In addition, a reliably defined hippocampal VOI could be used to examine tissue class contributions or measures of texture, and to measure T2 relaxation times on coregistered T2 images (Duncan et al., 1996) or T2 maps (Bernasconi et al., 2000), and finally, automatically defined larger temporal areas could be subjected to general appearance-based methods (Duchesne et al., 2006).

**Acknowledgments**

This work was performed during grant funding from the Medical Research Council to Dr Hammers (G108/585). The funding source had no role in this study. We thank our healthy and patient volunteers, our colleagues at the National Hospital for Neurology and Neurosurgery, Queen Square, for referring patients and for performing the surgical resections, and our colleagues at the National Society for Epilepsy for performing and transferring the MRI scans.

**References**


