Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates

Jon Pipitone, Min Tae M. Park, Julie Winterburn, Tristram A. Lett, Jason P. Lerch, Jens C. Pruessner, Martin Lepage, Aristotle N. Vineskos, M. Mallar Chakravarty

**Abstract**

Introduction: Advances in image segmentation of magnetic resonance images (MRI) have demonstrated that multi-atlas approaches improve segmentation over regular atlas-based approaches. These approaches often rely on a large number of manually segmented atlases (e.g., 30–80) that take significant time and expertise to produce. We present an algorithm, MAGeT-Brain (Multiple Automatically Generated Templates), for the automatic segmentation of the hippocampus that minimises the number of atlases needed whilst still achieving similar agreement to multi-atlas approaches. Thus, our method acts as a reliable multi-atlas approach when using special or hard-to-define atlases that are laborious to construct.

Method: MAGeT-Brain works by propagating atlas segmentations to a template library, formed from a subset of target images, via transformations estimated by nonlinear image registration. The resulting segmentations are then propagated to each target image and fused using a label fusion method.

We conduct two separate Monte Carlo cross-validation experiments comparing MAGeT-Brain and basic multi-atlas whole hippocampal segmentation using differing atlas and template library sizes, and registration and label fusion methods. The first experiment is a 10-fold validation (per parameter setting) over 60 subjects taken from the Alzheimer’s Disease Neuroimaging Database (ADNI), and the second is a five-fold validation over 81 subjects having had a first episode of psychosis. In both cases, automated segmentations are compared with manual segmentations following the Pruessner-protocol. Using the best settings found from these experiments, we segment 246 images of the ADNI1:Complete 1Yr 1.5T dataset and compare these with segmentations with manual segmentations following the Pruessner-protocol. Using the best settings found from these experiments, we segment 246 images of the ADNI1:Complete 1Yr 1.5T dataset and compare these with segmentations

Results: In the ADNI cross-validation, using 9 atlases MAGeT-Brain achieves a mean Dice’s Similarity Coefficient (DSC) score of 0.869 with respect to manual whole hippocampus segmentations, and also exhibits significantly lower variability in DSC scores than multi-atlas segmentation. In the younger, psychosis dataset, MAGeT-Brain achieves a mean DSC score of 0.892 and produces volumes which agree with manual segmentation volumes better than those produced by the FreeSurfer and FSL FIRST methods (mean difference in volume: 80 mm³, 1600 mm³, and 800 mm³, respectively). Similarly, in the ADNI1:Complete 1Yr 1.5T dataset, MAGeT-Brain produces hippocampal segmentations well correlated (r > 0.85) with SNT semi-automated reference volumes.
within disease categories, and shows a conservative bias and a mean difference in volume of 250 mm³ across the entire dataset, compared with FreeSurfer and FSL FIRST which both overestimate volume differences by 2600 mm³ and 2800 mm³ on average, respectively. Finally, MAGeT-Brain segments the CA1, CA4/DG and subiculum subfields on standard 3T T1-weighted resolution images with DSC overlap scores of 0.56, 0.65, and 0.58, respectively, relative to manual segmentations. Conclusion: We demonstrate that MAGeT-Brain produces consistent whole hippocampal segmentations using only 9 atlases, or fewer, with various hippocampal definitions, disease populations, and image acquisition types. Additionally, we show that MAGeT-Brain identifies hippocampal subfields in standard 3T T1-weighted images with overlap scores comparable to competing methods.

© 2014 Elsevier Inc. All rights reserved.
the basic multi-atlas strategy in which an atlas library is grown, beginning with a set of manually labelled atlases, by successively incorporating unlabelled target images once they themselves have been labelled using multi-atlas techniques. The sequence in which target images are labelled is chosen so that the similarity between the atlas images and the target images is minimised at each step, effectively allowing for deformations between very dissimilar images to be broken up into sequences of smaller deformations. Although Wolz et al. (2010) begin with an atlas library of 30 MR images, this method could theoretically work using a much smaller atlas library. In their validation, LEAP was used to segment the whole hippocampus in the ADNI1 baseline dataset, achieving a mean Dice score of 0.85 against semi-automated segmentations.

Also of interest to this manuscript are the methods that attempt to define hippocampal subfields using standard T1- or T2-weighted data, of which there are few. Van Leemput et al. (2009) demonstrate that the applicability of hippocampal subfield segmentation in T1-weighted images by Bayesian techniques using Markov random field shape priors learned from 10 manual segmentations. This work, available as part of the FreeSurfer package, is limited in that the segmentation omits the tail of the hippocampus and the protocol has yet to be fully validated. Yushkevich et al. (2009) manually segment hippocampal subfields on high-resolution datasets (either 0.2 mm isotropic or 0.2 mm × 0.3 mm × 0.2 mm resolution voxels) T2-weighted MR images acquired from five post-mortem medial temporal lobe samples. Then, using nonlinear registration guided by shape-based models of the subfield segmentations and manually derived hippocampus masks of the target images, the authors demonstrate accurate parcellation of hippocampal subfields, with respect to manual segmentations, in clinical T1 T1-weighted MR volumes. Using multi-atlas with bias correction techniques, Yushkevich et al. (2010) demonstrate a semi-automated method of subfield segmentation on in vivo focal T2-weighted MR acquisitions of the temporal lobe. Manual input is only needed to mark divisions between the head, body and tail of the hippocampus on target images.

In this paper we describe a thorough validation of the MAGeT-Brain algorithm for the fully automatic segmentation of the hippocampus and a proof-of-concept validation of its application to the segmentation of hippocampal subfields in standard T1-weighted images. First, we address the very idea of generating a template library from a limited number of input atlases (Chakravarty et al., 2013) for whole hippocampus segmentation by conducting a multi-fold validation experiment over a range of atlas and template library sizes, registration and label fusion methods. This type of validation is done first on a subset of the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset with manual segmentations following the Pruessner-protocol (Pruessner et al., 2000), and then replicated on a first episode psychosis patient dataset to determine the behaviour of MAGeT-Brain when segmenting younger and differently diseased subjects. Next, we compare MAGeT-Brain with other popular segmentation algorithms (FreeSurfer, FSL FIRST, MAPER, and SNT) on all the images available in the ADNI1:Complete 1Yr 1.5T sample. Lastly, using the optimal parameter settings for MAGeT-Brain found from the previous experiments, we investigate hippocampal subfield segmentation by conducting a leave-one-out validation using the Winterburn et al. (2013) manually segmented high-resolution MR atlases.

The MAGeT-Brain algorithm

We use the term label to mean any segmentation (manual or derived) of an MR image. Label propagation is the process by which two images are registered and the resulting transformation is applied to the labels from one image to bring them into alignment with the other image. We use the term atlas to mean a manually segmented image, and the term template to mean an automatically segmented image (i.e. via label propagation). The terms atlas library and template library describe any set of such images. Additionally, we use the term target to refer to an unlabelled image that is undergoing segmentation.

The simplest form of multi-atlas segmentation, which we call basic multi-atlas segmentation, involves three steps. First, each labelled input image is registered to an unlabelled target image. Second, the labels from each image are propagated to the target image. Third, the labels are combined into a single label by label fusion (Heckemann et al., 2006a, 2011). The basic multi-atlas segmentation method is described in detail in other publications (Aljabar et al., 2009; Collins and Pruessner, 2010; Heckemann et al., 2011). When only a single atlas is used, basic multi-atlas segmentation degenerates into model-based segmentation: labels are propagated from the atlas to a target, and no label fusion is needed.

The MAGeT-Brain (Multiple Automatically Generated Templates) algorithm creates a large template library given a much smaller sized input atlas library and then uses this template library in basic multi-atlas segmentation. The images used in the template library are selected from the target images, either arbitrarily or so as to reflect the neuroanatomy or demographics of the target set as a whole (for instance, by sampling equally from cases and controls). The template library images are automatically labelled by each of the atlases via label propagation. Basic multi-atlas segmentation is then conducted using the template library to segment the entire set of target images (including the target images used in the construction of the template library). Since each template library image has multiple labels (one from each atlas), the final number of labels to be fused for each target may be quite large (i.e. # of atlases × # of templates).

Fig. 1 illustrates the MAGeT-Brain algorithm schematically. Source code for MAGeT-Brain can be found at http://github.com/pipitone/MAGeTBrain.

Experiments

The following section describes experiments conducted to assess the segmentation quality of the MAGeT-Brain algorithm:

- Experiment 1 investigates MAGeT-Brain whole hippocampus segmentation of ageing and Alzheimer’s diseased subjects over a wide range of parameter settings using a Monte Carlo cross-validation design. The results of this experiment enable us to choose the parameter settings offering the best performance for use in subsequent experiments.
- Experiment 2 is a similar cross-validation to explore MAGeT-Brain segmentations on the brain images of young, first episode psychosis patients. MAGeT-Brain segmentations with two different atlas segmentation protocols are compared to automated segmentations by the FSL FIRST and FreeSurfer algorithms. The results of this experiment combined with the previous experiment establish parameter settings that do not overfit to the neuroanatomical features of a specific patient cohort.
- Experiment 3 bridges MAGeT-Brain with the existing segmentation literature by comparing MAGeT-Brain whole hippocampus segmentations with those of several well-known automated methods (FreeSurfer, FSL FIRST, MAPER, SNT) on the entire ADNI1:Complete 1Yr 1.5T dataset consisting of 246 brain images of subjects diagnosed as cognitively normal, having mild cognitive impairment, or Alzheimer’s disease.
- Experiment 4 assesses hippocampal subfield segmentation quality in a leave-one-out cross-validation on the five high-resolution manually segmented Winterburn MR atlases (Winterburn et al., 2013).

Experiment 1: Whole hippocampus segmentation cross-validation — Alzheimer’s disease

In this experiment we explore the very idea of generating a template library for multi-atlas-based segmentation from a small number of input
Multi-Atlas Segmentation

MAGeT Brain Segmentation

Legend
- Anatomical T1
- Anatomical Label
- Image Registration + Label Propagation
- Label Fusion

Fig. 1. A schematic illustration of basic multi-atlas segmentation and MAGeT-Brain segmentation. In multi-atlas segmentation, manual labels from atlas images are warped (propagated) into subject space by applying the transformations estimated from nonlinear image registration. The resulting candidate labels from all atlas images are then fused to create a final segmentation. In MAGeT-Brain segmentation, a template library is created by sampling (either randomly or representatively) from the subject images. Atlas labels are propagated to all template images and then to each subject image (including those used in the template library). The candidate labels for a subject are then fused into a final segmentation.

atlas. To do so, we conduct repeated cross-validations of MAGeT-Brain whilst varying the composition and sizes of the atlas and template libraries used, as well as varying the registration algorithm and label fusion method. The data used in this experiment are images from the ADNI dataset (Jack et al., 2008) along with whole hippocampus labels manually segmented following the Pruessner-protocol (Pruessner et al., 2000).

Note, in the Supplementary Materials we have replicated this experiment using the SNT semi-automated segmentations included as part of the ADNI dataset.

Experiment 1: Materials and methods

ADNI1: Complete 1Yr 1.5T dataset. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), and other biologic markers can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal (CN) older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Sixty 1.5T images were arbitrarily selected from the baseline scans in the ADNI1: Complete 1Yr 1.5T standardized dataset. Twenty subjects were chosen from each disease category: cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). Demographics for this subset are shown in Table 1. Fully manual segmentations of the left and right whole hippocampi in these images were provided by one author (JCP) according to the segmentation protocol specified in Pruessner et al. (2000).

Clinical, demographic and pre-processed T1-weighted MRI were downloaded by the authors from the ADNI database (adni.loni.usc.edu) between March 2012 and August 2012. The image dataset used was the ADNI1: Complete 1Yr 1.5T standardized dataset available from ADNI2 (Wyman et al., 2012). This image collection contains uniformly pre-processed images which have been designated to be the “best” after quality control. All images were acquired using 1.5T scanners (General Electric Healthcare, Philips Medical Systems or Siemens Medical Solutions) at multiple sites using the protocol described in Jack et al. (2008). Representative 1.5T imaging parameters were TR = 2400 ms, TI = 1000 ms, TE = 3.5 ms, flip angle = 8°, field of view = 240 × 240 mm, a 192 × 192 × 166 matrix (x, y, and z directions) yielding voxel dimensions of 1.25 mm × 1.25 mm × 1.2 mm.

2 http://adni.loni.usc.edu/methods/mri-analysis/adni-standardized-data/.
Experiment details. Monte Carlo Cross-Validation (MCCV), also known as repeated random sub-sampling cross-validation, consists of repeated rounds of validation conducted on a fixed dataset (Shao, 1993). In each round, the dataset is randomly partitioned into a training set and a validation set. The method to be validated is then given the training data, and its output is compared with the validation set. In this experiment, our dataset consists of 60 1.5T images and corresponding Pruessner-protocol manual segmentations. In each validation round, the dataset is partitioned into a training set consisting of images and manual segmentations used as an atlas library, and a validation set consisting of the remaining images to be segmented by both MAGeT-Brain and multi-atlas. The computed segmentations are compared to the manual segmentations (see Evaluation below).

A total of ten validation rounds were performed on each subject in the dataset, over each combination of parameter settings. The parameter settings explored are: atlas library size (1–9), template library size (1–20), registration method (ANTS or ANIMAL, described below), and label fusion method (majority vote, cross-correlation weighted majority vote, and normalised mutual information weighted majority vote, described below). In each validation round, both a MAGeT-Brain and multi-atlas segmentation is produced. A total of 10 × 60 × 9 × 20 × 2 × 3 = 6.48 × 10^5 validation rounds were conducted and resulting segmentations analysed.

Before registration, all images underwent prepossessing with the N3 algorithm (Sled et al., 1998) to minimise intensity nonuniformity. In this experiment we compared two nonlinear image registration methods:

Automatic Normalization and Image Matching and Anatomical Labeling (ANIMAL). The ANIMAL algorithm carries out image registration in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that maximises the correlation coefficient between blurred MRI intensities and gradient magnitude over the whole brain (Collins et al., 1994). In the second phase, nonlinear registration is completed using the ANIMAL algorithm (Collins et al., 1995): an iterative procedure that estimates a 3D deformation field between two MR images. At first, large deformations are estimated using a blurred version of the input data. These large deformations are then input to subsequent steps where the fit is refined by estimating smaller deformations on data blurred with a Gaussian kernel with a smaller full width at half maximum (FWHM). The final transformation is a set of local translations defined on a bed of equally spaced nodes that were estimated through the optimization of the correlation coefficient. For the purposes of this work we used the regularization parameters optimised in Robbins et al. (2004), displayed in Table 2.

Automatic Normalization Tools (ANTS). ANTS is a diffeomorphic registration algorithm which provides great flexibility over the choice of transformation model, objective function, and the consistency of the final transformation (Avants et al., 2008). The transformation is estimated in a hierarchical fashion where the MRI data is subsampled, allowing large deformations to be estimated and successively refined at later hierarchical stages (where the data is subsampled to a finer grid). The deformation field and the objective function are regularized with a Gaussian kernel at each level of the hierarchy. The ANTS algorithm is freely available. We used an implementation of the ANTS algorithm compatible with the MINC data format, mincANTS.

We used the following command line when running ANTS:

```
mincANTS 3 -r Pr[target_file.mnc,source_file.mnc,1,4]
```

Table 1

<table>
<thead>
<tr>
<th></th>
<th>CN (N = 20)</th>
<th>LMC (N = 20)</th>
<th>AD (N = 20)</th>
<th>Combined (N = 60)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at baseline</td>
<td>72.2 75.5 80.3</td>
<td>70.9 75.6 80.4</td>
<td>69.4 74.9 80.1</td>
<td>70.9 75.2 80.2</td>
</tr>
<tr>
<td>Sex: Female</td>
<td>50% (10)</td>
<td>50% (10)</td>
<td>50% (10)</td>
<td>50% (30)</td>
</tr>
<tr>
<td>Education</td>
<td>14.0 16.0 18.0</td>
<td>13.8 16.0 16.5</td>
<td>12.0 15.5 18.0</td>
<td>13.0 16.0 18.0</td>
</tr>
<tr>
<td>CDR-SB</td>
<td>0.00 0.00 0.00</td>
<td>1.00 2.00 2.50</td>
<td>3.50 4.00 5.00</td>
<td>0.00 1.75 2.62</td>
</tr>
<tr>
<td>ADAS 13</td>
<td>6.00 7.67 11.00</td>
<td>14.92 20.50 25.75</td>
<td>24.33 27.00 32.09</td>
<td>9.50 18.84 26.25</td>
</tr>
<tr>
<td>MMSE</td>
<td>28.8 29.5 30.0</td>
<td>26.0 27.5 28.2</td>
<td>22.8 23.0 24.0</td>
<td>24.0 27.0 29.0</td>
</tr>
</tbody>
</table>

The settings were adapted from the “reasonable starting point” given in the ANTS manual.

Label fusion methods. Label fusion is a term given to the process of combining the information from several candidate labels for an image into a single labelling. In this experiment we explore three fusion methods:

Voxel-wise Majority Vote Labels are propagated from all template library images to a target. Each output voxel is given the most frequent label at that voxel location amongst all candidate labels.

Cross-correlation Weighted Majority Vote An optimal combination of targets from the template library has previously been shown to improve segmentation accuracy with respect to manual segmentations (Aljabar et al., 2009; Collins and Pruessner, 2010). In this method, each template library image is ranked in similarity to each unlabelled image by the normalised cross-correlation (CC) of image intensities after linear registration, over a region of interest (ROI) generously encompassing the hippocampus. Only the top ranked template library image labels are used in a voxel-wise majority vote. The ROI is heuristically defined as the extent of all atlas labels after linear registration to the template, dilated by three voxels (Chakravarty et al., 2013). The number of top ranked

Table 2

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model blur (FWHM)</td>
<td>8</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Input blur (FWHM)</td>
<td>8</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Iterations</td>
<td>30</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>Step</td>
<td>8 × 8 × 8</td>
<td>4 × 4 × 4</td>
<td>2 × 2 × 2</td>
</tr>
<tr>
<td>Sub-lattice</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Lattice diameter</td>
<td>24 × 24 × 24</td>
<td>12 × 12 × 12</td>
<td>6 × 6 × 6</td>
</tr>
</tbody>
</table>

template library image labels is a configurable parameter and displayed as the size of the template library in the rest of the paper. The `xcorr_vol` utility from the ANIMAL toolkit is used to calculate the cross-correlation similarity measure.

**Normalised Mutual Information Weighted Majority Vote** This method is similar to cross-correlation-weighted voting except that image similarity is calculated by the normalised mutual information score over the region of interest (Studholme et al., 2001). The `itk_similarity` utility from the EZMinc toolkit is used to calculate the normalised mutual information measure between two images.

**Evaluation method.** The Dice similarity coefficient (DSC), also known as Dice’s Kappa, assesses the agreement between two segmentations. It is one of the most widely used measures of segmentation agreement, and we use it as the basis of comparison in this experiment.

\[
\text{DSC} = \frac{2|A \cap B|}{|A| + |B|}
\]

where \(A\) and \(B\) are the regions being compared, and the cardinality is the volume measured in voxels. The labels produced by MAGeT-Brain and multi-atlas segmentation are compared to the manual labels using the Dice similarity coefficient, and the recorded value for each subject at each parameter setting explored in this experiment is the average over ten validation rounds.

Additionally, the sensitivity of MAGeT-Brain and multi-atlas to atlas and template library composition is evaluated by comparing the variability in Dice scores over all validation rounds at fixed parameter settings. This is achieved by first computing the variance of DSC scores in each block of ten validation rounds per subject. The distribution of this statistic across all subjects is then compared between MAGeT-Brain and multi-atlas using a Student’s t-test. A significant difference between distributions is taken to show either a larger or smaller level of variability between methods.

**Experiment 1: Results**

We find that for MAGeT-Brain segmentations, similarity score increases as atlas and template library size is increased, although with diminishing returns with larger atlas libraries. Peak increase (+0.025) is found when using 9 atlases, 19 templates, ANTS registration, and multi-atlas using a Student’s t-test. A significant difference between distributions is taken to show either a larger or smaller level of variability between methods.

**Results**

First Episode Psychosis (FEP) dataset. All patients were recruited and treated through the Prevention and Early Intervention Program for Psychoses (PEPP-Montreal), a specialized early intervention service at the Douglas Mental Health University Institute in Montreal, Canada. People aged 14 to 35 years from the local catchment area suffering from either affective or non-affective psychosis who had not taken antipsychotic medication for more than one month with an IQ above 70 were consecutively admitted as either in- or out-patients. Of those treated at PEPP, only patients aged 18 to 30 years with no previous history of neurological disease or head trauma causing loss of consciousness were eligible for the neuroimaging study; only those suffering from schizophrenia spectrum disorders were considered for this analysis. For complete programme details see Mallia et al. (2003).

**Experiment 2: Whole hippocampus segmentation cross-validation — first episode of psychosis**

To validate that the MAGeT-Brain works effectively in the context of other neurological disorders, in this experiment we replicate the cross-validation done in Experiment 1 with a dataset of patients having had a single episode of psychosis. We also compare MAGeT-Brain segmentations with those of two well-known automated segmentation methods, FSL FIRST and FreeSurfer.

**Experiment 2: Materials and methods**

First Episode Psychosis (FEP) dataset. All patients were recruited and treated through the Prevention and Early Intervention Program for Psychoses (PEPP-Montreal), a specialized early intervention service at the Douglas Mental Health University Institute in Montreal, Canada. People aged 14 to 35 years from the local catchment area suffering from either affective or non-affective psychosis who had not taken antipsychotic medication for more than one month with an IQ above 70 were consecutively admitted as either in- or out-patients. Of those treated at PEPP, only patients aged 18 to 30 years with no previous history of neurological disease or head trauma causing loss of consciousness were eligible for the neuroimaging study; only those suffering from schizophrenia spectrum disorders were considered for this analysis. For complete programme details see Mallia et al. (2003).

Scanning of 81 subjects was carried out at the Montreal Neurological Institute on a 1.5-T Siemens whole body MRI system. Structural T1 volumes were acquired for each participant using a three-dimensional (3D) gradient echo pulse sequence with sagittal volume excitation (repetition time = 22 ms, echo time = 9.2 ms, flip angle = 30°, 180 mm contiguous sagittal slices). The rectangular field-of-view for the images was 256 mm (SI) × 204 mm (AP). Subject demographics are shown in Table 3.

Expert whole hippocampal manual segmentation of each subject is produced following a validated segmentation protocol (Pruessnser et al., 2000).

**Winterburn atlases.** The Winterburn atlases (Winterburn et al., 2013) are digital hippocampal segmentations of five in-vivo 0.3 mm isotropic T1-weighted MR images. The segmentations include subfield segmentations for the cornu ammonis (CA) 1; CA2 and CA3; CA4 and dentate gyrus; subiculum; and strata radiatum (SR), strata lacunosum (SL), and strata moleculare (SM). Subjects in the Winterburn atlases range in age from 29 to 57 years (mean age of 37), and include two males and three females.

**Experiment details.** The same overall design as Experiment 1 is followed in this experiment: a Monte Carlo cross-validation (MCCV) is conducted using the pool of 81 first episode psychosis subject brain images and corresponding Pruessnser-protocol manual segmentations. Five rounds of validation are conducted for each subject, and each atlas and template library size combination (1–9 atlases, 1–19 templates). In each round,
images and their manual labels are randomly selected from the pool, and
the remaining images are segmented using MAGeT-Brain with a random
subset of the unlabelled images also serving as template images. Majority
vote fusion, and the ANTS registration algorithm are used, as these have
shown to behave favourably in previous experiments.

In addition to the MCCV, we segment the entire first episode psychosis
data set using MAGeT-Brain using two different atlases sets, as well as
with two popular automated segmentation packages, FSL FIRST and
FreeSurfer. Specifically, MAGeT-Brain is run once with the five
Winterburn atlas images and labels as atlases and a randomly selected
subset of 19 target images as templates. MAGeT-Brain is run a second
time using the same template images, but we used five additional first ep-
isode psychosis subjects and corresponding manual segmentations (not
included above) as atlases. FSL FIRST and FreeSurfer are run with the de-
fault settings: FSL FIRST run_first_all script was used according to the
FIRST user guide,7 and FreeSurfer was run with the command recon-all
-all.


---

Fig. 2. Whole hippocampus segmentation cross-validation on ADNI subjects with Pruessner-protocol manual segmentations. (2a) Average DSC score of MAGeT-Brain with manual segmentations for 60 ADNI subjects taken over 10 folds of cross-validation at each parameter setting. Error bars indicate standard error. (2b) Increase in DSC of MAGeT-Brain over multi-atlas segmentations. (2c) shows the significance of t-tests comparing the variability in DSC scores of MAGeT-Brain and multi-atlas across validation folds. Only points where MAGeT-Brain mean variability is lower than multi-atlas are shown. Dashed lines indicate p-values of 0.05 and 0.01.
proportional bias: FreeSurfer and FSL FIRST markedly underestimate smaller hippocampi and over-estimate large hippocampi (the limits of agreement are between $-2482\, \text{mm}^3$ and $-784\, \text{mm}^3$, and between $-1653\, \text{mm}^3$ and $-79\, \text{mm}^3$, respectively), whereas both MAGeT-Brain methods show a much less exaggerated, but conservative bias (limits of agreement between $-67\, \text{mm}^3$ and $-766\, \text{mm}^3$ when using FEP atlases, and between $-333\, \text{mm}^3$ and $-504\, \text{mm}^3$ when using Winterburn atlases). On average, FreeSurfer and FSL FIRST overestimate hippocampal volume by about $1600\, \text{mm}^3$ and $800\, \text{mm}^3$, respectively. In contrast, on average MAGeT-Brain underestimates volumes by about $300\, \text{mm}^3$ when using FEP atlases and by about $80\, \text{mm}^3$ when using Winterburn atlases (compared to the Pruessner-protocol manual segmentations).

**Experiment 3: Whole hippocampus segmentation comparison — ADNI1 complete 1Yr**

To validate MAGeT-Brain segmentation quality with respect to other established automated hippocampal segmentation methods, we apply MAGeT-Brain to a large dataset from the ADNI project. The resulting segmentations are compared to those produced by FreeSurfer, FSL FIRST, MAPER, as well as semi-automated whole hippocampal segmentations (SNT) provided by ADNI.

**Experiment 3: Materials and methods**

ADNI1:Complete 1Yr 1.5T dataset. The ADNI1:Complete 1Yr 1.5T standardized dataset contains 1919 images in total. SNT, MAPER, and FreeSurfer hippocampal volumes for a subset of images were provided by ADNI, along with quality control data for each FreeSurfer segmentation (guidelines described in (Hartig et al., 2010)). See Experiment 1 for study details, inclusion criteria and imaging characteristics. Demographics are shown in Table 4.

For a subset of the ADNI images, semi-automated segmentations of the left and right whole hippocampi generated using the SNT tool from Medtronic Surgical Navigation Technologies, Louisville, CO (see Supplementary Materials for detailed discussion of the segmentation process) are made available (Hsu et al., 2002). These labels are used as the reference labels in several other studies of (semi-) automated segmentation methods (see Discussion). In addition, ADNI also distributes hippocampal segmentations and volumes determined using MAPER (Heckemann et al., 2011), a multi-atlas segmentation tool, and the FreeSurfer tool (including quality control data, with guidelines described in Hartig et al. (2010)).
Experiment details. MAGeT-Brain was configured with an atlas library composed of the five Winterburn atlas images (Experiment 2) and segmentations. A template library of 19 images were randomly selected from the target dataset of ADNI subjects, and ANTS registration and majority vote label fusion were used as these were found to perform favourably in earlier experiments.

**Fig. 3.** First Episode Patient dataset validation. All manual segmentation of the 81 subjects is done with the Pruessner-protocol. MAGeT-Brain uses ANTS registration and majority vote label fusion. (3a) shows mean DSC score of MAGeT-Brain segmentations, as atlas and template library size is varied over a 5-fold validation. Error bars indicate standard error. (3b) shows segmentation volumes from FSL FIRST, FreeSurfer, MAGeT-Brain using the five Winterburn atlases (MAGeT-Winterburn), and MAGeT-Brain using five manually segmented FEP subjects as atlases (MAGeT-FEP). Linear fit lines are shown, with the shaded region showing standard error. (3c) shows the agreement between computed and manually volumes. The overall mean difference in volume, and limits of agreement (±1.96SD) are shown by dashed horizontal lines. Linear fit lines are shown for each diagnosis group. Note, points below the mean difference indicate overestimation of the volume with respect to the manual volume, and vice versa.

FSL FIRST segmentation was performed using the `run_first_all` script according to the FIRST user guide. All images in the ADNI1:Complete 1Yr 1.5T dataset were segmented by both methods.

---

One author (MP) performed visual quality inspection for MAGeT-Brain and FSL FIRST segmentations using similar quality control guidelines to those used by FreeSurfer. If either hippocampus was under or over segmented by 10 mm or greater in three or more slices then the segmentation did not pass. Only images meeting the conditions of having segmentations from all methods (SNT, MAPER, FreeSurfer, FSL FIRST, and MAGeT-Brain) and also passing quality control inspection were included in the analysis (Table 5).

Evaluation method. As in previous experiments, the Winterburn hippocampal segmentation protocol differs in the delineated neuroanatomical features (Winterburn et al. (2013), and Table 9, Discussion) and so we assess MAGeT-Brain by the degree of (Pearson) correlation of average hippocampal volume across subjects. We also computed the correlation in hippocampal volume between existing, established automated segmentation methods — FSL FIRST, FreeSurfer, and MAPER, and SNT semi-automated segmentations. Additionally, we evaluate the volume-
related fixed and proportional biases in all segmentation methods using Bland–Altman plots (Bland and Altman, 1986).

Experiment 3: Results

We found a close relationship in total bilateral hippocampal volume between all methods and the SNT semi-automated label volumes (Fig. 4a). Volumes are well correlated ($r > 0.78$) for all methods, and across disease categories. Within disease categories (Fig. 4b), MAGeT-Brain is consistently well correlated to SNT volumes ($r > 0.85$), but appears to slightly over-estimate the volume of the AD hippocampus compared to the SNT segmentations.

Bland–Altman plots illustrate the level of agreement of each method with SNT segmentation hippocampal volumes (Fig. 4c). All methods show an obvious proportional bias: FreeSurfer and FSL FIRST markedly under-estimate smaller hippocampi and over-estimate large hippocampi, whereas MAPER and MAGeT-Brain show a reverse, conservative bias (Fig. 4c). Additionally, all methods show a fixed volume bias, with FreeSurfer and FSL FIRST most dramatically over-estimating hippocampal volume by 2600 mm$^3$ and 2800 mm$^3$ on average, respectively, and MAPER and MAGeT-Brain within 250 mm$^3$ on average.

Fig. 5 shows a qualitative comparison of MAGeT-Brain and SNT hippocampal segmentations for 10 randomly selected subjects in each disease category, and illustrates some of the common errors found during manual segmentations for 10 randomly selected subjects in each disease category. Simulated manual 0.33 mm-isotropic voxel segmentations by one voxel in the form of standard 3T T1-weighted images. Image subsampling is performed using trilinear subsampling techniques. In each round of LOOCV, a single atlas image is selected and treated as a target image to be segmented by MAGeT-Brain. So as to have an odd-sized atlas library, atlas image is segmented once using each possible triple of atlas images, and corresponding manual segmentations, from the remaining four unselected atlases. Thus, for each of the five atlases, a total of $3 \times 4 = 20$ segmentations are evaluated, resulting in a combined total of 5 × 4 = 20 segmentations evaluated overall. We chose an atlas library with an odd number of images so as to ensure unbiased label fusion when using majority voting (see Discussion).

The template library used has a total of 19 images composed of all five resampled atlas images plus the additional 14 images from the healthy control dataset. The ANTS registration algorithm was used for image registration, and majority voting was used for label fusion, as these methods proved most favourable in the previous whole hippocampal validation experiments.

Evaluation method. Evaluating the agreement of automated hippocampal subfield segmentations with manual segmentations for T1 images at 0.9 mm-isotropic voxels is inherently ill-defined since there are no manual protocols for segmentation at this resolution. Instead, we must evaluate how well the lower-resolution MAGeT-Brain hippocampal subfield segmentations correspond in form to the segmentation protocol used in the high-resolution images. By directly resampling the Winterburn atlas segmentations to 0.9 mm$^3$ voxels (using standard nearest-neighbour image resampling techniques) we obtain a subsampled version of the labels which preserve the original segmentation protocol within the limits of error from rounding and interpolation.

Therefore, using the resampled Winterburn segmentations as definitive for the 0.9 mm$^3$ resolution we evaluate agreement of MAGeT-Brain segmentations using DSC overlap scores and evaluate consistency across the range of hippocampal sizes using Bland–Altman plots of subfield volumes.

Additionally, by shifting the original manual 0.33 mm-isotropic voxel segmentations by one voxel in the $x$, $y$, and $z$ directions and then resampling it to 0.9 mm-isotropic voxels we obtain a simulated manual segmentation having a small amount of error. We can compare the DSC overlap score of the shifted labels (relative to the directly resampled labels) with the DSC score of the MAGeT-Brain generated labels in order to evaluate their relevance.

Experiment 4: Results

Fig. 6a shows the overlap similarity scores between the MAGeT-Brain segmentations and the resampled Winterburn atlases for each

### Table 6

Demographics for the hippocampal subfield cross-validation healthy control subject sample used in the template library (excluding the Winterburn atlas subjects). Education is shown in years. Values are presented as lower quartile, median, and upper quartile for continuous variables, or as a percentage (frequency) for discrete variables. N is the number of non-missing values.

<table>
<thead>
<tr>
<th>Age</th>
<th>N</th>
<th>Control</th>
<th>N = 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex: male</td>
<td>14</td>
<td>43%</td>
<td>(6)</td>
</tr>
<tr>
<td>Education: 12</td>
<td>13</td>
<td>15%</td>
<td>(2)</td>
</tr>
<tr>
<td>13</td>
<td>8%</td>
<td>(1)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>23%</td>
<td>(3)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15%</td>
<td>(2)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>38%</td>
<td>(5)</td>
<td></td>
</tr>
<tr>
<td>Handedness: R</td>
<td>14</td>
<td>93%</td>
<td>(13)</td>
</tr>
</tbody>
</table>

### Table 7

Overlap similarity results for the each of the subfields of the hippocampus. Simulated overlap similarity results are also given for manual labels that were translated by one voxel (i.e.: 0.3 mm) in all directions and then resampled. Values are given as mean Dice’s Similarity Coefficient (DSC) ± standard deviation.

<table>
<thead>
<tr>
<th>Subfield</th>
<th>MAGeT</th>
<th>0.9 mm translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA1</td>
<td>0.56 ± 0.05</td>
<td>0.27 ± 0.03</td>
</tr>
<tr>
<td>CA2/CA3</td>
<td>0.41 ± 0.10</td>
<td>0.12 ± 0.05</td>
</tr>
<tr>
<td>CA4/SG</td>
<td>0.65 ± 0.05</td>
<td>0.42 ± 0.05</td>
</tr>
<tr>
<td>SR/SLSM</td>
<td>0.43 ± 0.05</td>
<td>0.19 ± 0.04</td>
</tr>
<tr>
<td>Subiculum</td>
<td>0.58 ± 0.06</td>
<td>0.14 ± 0.04</td>
</tr>
</tbody>
</table>

### Table 8

MAGeT-Brain segmentations and the resampled Winterburn atlases for each

### Experiment 4: Materials and methods

Healthy control dataset. T1 MR images of 14 subjects were acquired as a part of an ongoing study at the Centre for Addiction and Mental Health (Table 6). Subjects were known to be free of neuropsychiatric disorders and gave informed consent. These images were acquired on a 3T GE Discovery MR 750 system (General Electric, Milwaukee, WI) using an 8-channel head coil with the enhanced fast gradient recalled echo 3-dimensional acquisition protocol, FGRE-BRAVO, with the following parameters: $TE/TR/TI = 3.0$ ms/6.7 ms/650 ms, flip angle = 8°, FOV = 15.3 cm, slice thickness = 0.9 mm, 170 in-plane steps for an approximate 0.9 mm-isotropic voxel resolution.

Experiment details. Leave-one-out cross-validation (LOOCV) is a validation approach in which an algorithm is given all but one item in a dataset as training data (in our case, atlas images and labels) and then the algorithm is applied to the left-out item. This is done, in turn, for each item in the dataset and the output across all items is evaluated together.

In this experiment, the Winterburn atlases (Experiment 2) are resampled to 0.9 mm-isotropic voxel resolution to simulate standard 3T T1-weighted resolution images. Image subsampling is performed using trilinear subsampling techniques. In each round of LOOCV, a single atlas image is selected and treated as a target image to be segmented by MAGeT-Brain. So as to have an odd-sized atlas library, atlas image is segmented once using each possible triple of atlas images, and corresponding manual segmentations, from the remaining four unselected atlases. Thus, for each of the five atlases, a total of $3 \times 4 = 20$ segmentations are evaluated, resulting in a combined total of 5 × 4 = 20 segmentations evaluated overall. We chose an atlas library with an odd number of images so as to ensure unbiased label fusion when using majority voting (see Discussion).

The template library used has a total of 19 images composed of all five resampled atlas images plus the additional 14 images from the healthy control dataset. The ANTS registration algorithm was used for image registration, and majority voting was used for label fusion, as these methods proved most favourable in the previous whole hippocampal validation experiments.

Evaluation method. Evaluating the agreement of automated hippocampal subfield segmentations with manual segmentations for T1 images at 0.9 mm-isotropic voxels is inherently ill-defined since there are no manual protocols for segmentation at this resolution. Instead, we must evaluate how well the lower-resolution MAGeT-Brain hippocampal subfield segmentations correspond in form to the segmentation protocol used in the high-resolution images. By directly resampling the Winterburn atlas segmentations to 0.9 mm$^3$ voxels (using standard nearest-neighbour image resampling techniques) we obtain a subsampled version of the labels which preserve the original segmentation protocol within the limits of error from rounding and interpolation.

Therefore, using the resampled Winterburn segmentations as definitive for the 0.9 mm$^3$ resolution we evaluate agreement of MAGeT-Brain segmentations using DSC overlap scores and evaluate consistency across the range of hippocampal sizes using Bland–Altman plots of subfield volumes.

Additionally, by shifting the original manual 0.33 mm-isotropic voxel segmentations by one voxel in the $x$, $y$, and $z$ directions and then resampling it to 0.9 mm-isotropic voxels we obtain a simulated manual segmentation having a small amount of error. We can compare the DSC overlap score of the shifted labels (relative to the directly resampled labels) with the DSC score of the MAGeT-Brain generated labels in order to evaluate their relevance.

Experiment 4: Results

Fig. 6a shows the overlap similarity scores between the MAGeT-Brain segmentations and the resampled Winterburn atlases for each
hippocampal subfield across all subjects and folds of the validation. Mean and standard deviation DSC scores of the subfields are shown in Table 7, along with DSC scores for the resampled atlas segmentations when perturbed slightly and compared to the originals. We find that the CA4/DG subfield shows the highest mean DSC score of 0.647 ± 0.051, followed by the Subiculum and CA1 subfields having scores of 0.563 ± 0.046 and 0.58 ± 0.057, respectively. Both the CA4/DG and molecular regions score below 0.5. These scores may seem low but not when taken in context and compared to existing semi-automated methods (see Discussion). The whole hippocampus is segmented with a mean DSC score of 0.816 ± 0.023.

Fig. 6b contains Bland–Altman plots comparing MAGeT-Brain volumes with manual volumes across all validation folds. MAGeT-Brain displays a conservative proportional bias — small hippocampi are overestimated in volume, and larger hippocampi are underestimated (a mean maximum difference of approximately 200 mm³ across all subfields). MAGeT-Brain display a slight conservative fixed bias, tending to underestimate all subfields except CA4/DG (mean underestimation: CA1 = 76 mm³, CA2/3 = 56 mm³, CA4/DG = −16 mm³, Subiculum = 48 mm³, SR/SL/SM = 96 mm³).

Fig. 7 shows slices subfield segmentations for a single subject for qualitative inspection.

### Discussion

In this manuscript we have presented the implementation and validation of the MAGeT-Brain framework — a methodology that requires very few input atlases in order to provide accurate and reliable segmentations with respect to manual segmentations. Both Experiment 1 (Section 3.1) and Experiment 2 (Section 3.2) compare MAGeT-Brain to basic-multi-atlas segmentation by characterising the change in segmentation quality with varying parameter settings ( atlases and template library sizes, registration method, and label fusion method) and differing age and neuropsychiatric populations. Together, these experiments allow us to choose optimal MAGeT-Brain parameter settings for use in subsequent experiments. Experiment 3 (Section 3.3) demonstrates that across 246 images from the ADNI1:Complete 1Yr 1.5T dataset, MAGeT-Brain performs as well as, or better, than other established and popular methods, and has a much more conservative proportional bias in segmentation volume. Finally, Experiment 4 (Section 3.4) is a proof-of-concept validation demonstrating the reliability of MAGeT-Brain in producing subfield segmentations which match the segmentation protocol of the input atlases despite contrast and resolution limitations in standard T1-weighted image volumes. All of these experiments together demonstrate that MAGeT-Brain’s algorithmic performance is not dependent on a single definition of the hippocampus but is effective with differing hippocampal definitions (Hsu et al., 2002; Pruessner et al., 2000; Winterburn et al., 2013), across image types, and subject populations.

The core claim the MAGeT-Brain method is based on – that a useful template library can be generated from a small set of labelled atlas images – is validated in the cross-validation conducted in Experiment 1 (and the replication in Experiment 2 and Experiment 5, Supplementary Materials). We find that both increasing the number of atlases and the number of templates used improves MAGeT-Brain segmentation over and above basic-multi-atlas segmentations using the same number of atlas images. That is, by taking the extra step of generating a template library from the input atlases, we can produce accurate and reliable segmentations without resorting to the use of additional atlases.

### Table 8

Summary of automated segmentation methods of the hippocampus. This table summarizes published Dice’s overlap measure between automated and manual segmentations of the hippocampus. Unless otherwise specified, validation datasets are composed equally of cases and control subjects, and use manual segmentation labels as ground truth in computing DSC scores. AD = Alzheimer’s Disease; MCI = Mild Cognitive Impairment; CN = Cognitively Normal (CN); FEP = First Episode of Psychosis; LOOCV = Leave-one-out cross-validation; MCCV = Monte Carlo cross-validation; SNT = Surgical Medtronic Navigation Technologies semi-automated labels. Some studies of automated segmentation of ADNI images are excluded because they do not provide overlap measures for the hippocampus (Chupin et al., 2009; Heckemann et al., 2011).

<table>
<thead>
<tr>
<th>Method</th>
<th>Atlases</th>
<th>DSC mean</th>
<th>Reference</th>
<th>Validation</th>
<th>Dataset (truth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAGeT-Brain</td>
<td>9</td>
<td>0.841</td>
<td></td>
<td>10-Fold MCCV on 69 subjects</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>Patch-based label fusion</td>
<td>16</td>
<td>0.861*</td>
<td>Coupe et al. (2011)</td>
<td>LOOCV</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>Joint Label Fusion</td>
<td>20</td>
<td>0.848</td>
<td>Wang et al. (2011)</td>
<td>10-Fold MCCV on 20 of 139 subjects</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>ACM (AdaBoost-based)</td>
<td>21</td>
<td>0.862</td>
<td>Morra et al. (2008)</td>
<td>LOOCV</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>LEAP</td>
<td>30</td>
<td>0.846</td>
<td>Wolz et al. (2010)</td>
<td>LOOCV</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>Multi-atlas</td>
<td>30</td>
<td>0.885</td>
<td>Lüttjohann et al. 2010</td>
<td>Segmentation of 60 subjects</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>Multi-atlas (MAPS)</td>
<td>55</td>
<td>0.890</td>
<td>Leung et al. (2010)</td>
<td>Segmentation of 30 subjects</td>
<td>ADNI (SNT)</td>
</tr>
<tr>
<td>MAGeT-Brain</td>
<td>9</td>
<td>0.869</td>
<td></td>
<td>10-Fold MCCV on 60 subjects</td>
<td>ADNI (Pruessner)</td>
</tr>
<tr>
<td>MAGeT-Brain</td>
<td>9</td>
<td>0.892</td>
<td></td>
<td>5-Fold MCCV on 81 subjects</td>
<td>FEP subjects</td>
</tr>
<tr>
<td>Neural nets</td>
<td>10</td>
<td>0.740</td>
<td>Powell et al. (2008)</td>
<td>Segmentation of 5 subjects</td>
<td>Controls</td>
</tr>
<tr>
<td>Probabilistic atlas</td>
<td>11</td>
<td>0.852</td>
<td>van der Lijn et al. (2008)</td>
<td>11 atlases used in 100 rounds of LOOCV on 20 elderly subjects</td>
<td>Controls</td>
</tr>
<tr>
<td>Probabilistic Atlas</td>
<td>16</td>
<td>0.860</td>
<td>Chupin et al. (2009)</td>
<td>LOOCV</td>
<td>AD subjects</td>
</tr>
<tr>
<td>Anatomically-guided EM</td>
<td>17</td>
<td>0.812</td>
<td>Pohl et al. (2007)</td>
<td>LOOCV on 17 controls, segmentation of 33 mixed subjects</td>
<td>Mixed diagnosis</td>
</tr>
<tr>
<td>Multi-atlas</td>
<td>30</td>
<td>0.820</td>
<td>Heckemann et al. (2006a)</td>
<td>LOOCV</td>
<td>Controls</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.880</td>
<td>Gousias et al. (2008)</td>
<td>30 adult atlas used, segmentation of 33 2 yr old subjects</td>
<td>Controls</td>
</tr>
<tr>
<td>Multi-atlas</td>
<td>80</td>
<td>0.890</td>
<td>Collins and Pruessner (2010)</td>
<td>LOOCV</td>
<td>Controls</td>
</tr>
<tr>
<td>Multi-atlas</td>
<td>55</td>
<td>0.860</td>
<td>Barnes et al. (2008)</td>
<td>LOOCV</td>
<td>Controls and AD</td>
</tr>
<tr>
<td>Multi-atlas</td>
<td>275</td>
<td>0.835</td>
<td>Aijabar et al. (2009)</td>
<td>LOOCV</td>
<td>Controls</td>
</tr>
</tbody>
</table>

* (AD: 0.838, MCI: n/a, CN: 0.883). b (AD: n/a, MCI: 0.798, CN: 0.898).
Fig. 5. SNT and MAGeT-Brain segmentations for 30 ADNI subjects — 10 subjects randomly selected from each disease category in the subject pool used in Experiment 1 (Section 1). Sagittal slices are shown for each unlabelled T1-weighted anatomical image. SNT labels appear in green, and MAGeT-Brain labels appear in blue. Noted are examples of common segmentation idiosyncrasies: (a) over-estimation of hippocampal head and (b) translated segmentation (seen in SNT segmentations only); (c) under-estimation of hippocampal body and (d) improper inclusion of the vestigial hippocampal sulcus by MAGeT-Brain.
larger atlas libraries we have found that generating a template library reduces the variability in segmentation agreement (i.e. MAGeT-Brain more consistently produces segmentations in greater agreement with manual segmentations than does basic-multi-atlas method, over repeated randomized trials). These effects do not appear dependant on the hippocampal segmentation protocol used.

Interestingly, previous work on multi-atlas segmentation methods (Aljabar et al., 2009; Collins and Pruessner, 2010) has found that cross-correlation and normalised mutual information-based weighted label fusion improves segmentation reliability over simple majority vote label fusion, and yet we did not see a significant indication of this effect in the MAGeT-Brain segmentations. Selectively filtering out atlases with lower image similarity is believed to reduce sources of error from estimating deformations via nonlinear registration, partial volume effects from nearest neighbour image resampling, and neuroanatomical mismatch between atlases and subjects. That MAGeT-Brain does not see the same boost in performance from weighted voting may suggest that the neuroanatomical variability of a template library

![DSC score by subfield](image)

![Bland-Altman plots of computed vs. manual subfield volumes](image)

Fig. 6. Hippocampal subfield cross-validation. (6a) Similarity of MAGeT-Brain segmentation of subfields and the resampled Winterburn atlas segmentations at 0.9 mm$^3$ voxel resolution, over all validation folds. Overlap score for each hemisphere is measured separately. (6b) shows the agreement, by subfield, of computed and manual volumes across all validation folds. The overall mean difference in volume, and limits of agreement (± 1.96SD) are shown by dashed horizontal lines. Linear fit lines are shown. Note, points below the mean difference indication overestimation of the volume with respect to the resampled volume, and vice versa.
constructed from study subjects more closely matches any particular subject and thereby leaving less error to filter. From our previous work on the MAGeT-Brain algorithm we have shown that the reduction in error is not simply a smoothing or averaging effect (Chakravarty et al., 2013).

Although, the goal of this manuscript was not to exhaustively test or validate multiple different voting strategies in the context of our segmentation algorithm, it is important to note that other strategies for voting are available. For example, other groups have used the STAPLE algorithm (Warfield et al., 2004) (or variants of the STAPLE algorithm (Robitaille and Duchesne, 2012)) which weighs each segmentation based upon its estimated performance level with respect to the other available candidate segmentations. Further, the sensitivity and specificity parameters can also be tuned to potentially improve segmentation reliability. It is likely that using more sophisticated voting methods would have a positive effect on the overall segmentation performance, as demonstrated by the STAPLE algorithm. However, it is also important to note that even in the absence of a more sophisticated label fusion algorithm, MAGeT-Brain performs reasonably well in comparison to other groups that have tested new segmentation algorithm with Alzheimer’s disease, mild cognitive impairment, and cognitively normal data from the ADNI database (Table 8). In addition, our validation in Experiment 2 (with the first episode psychosis subjects) yields DSC’s that are amongst the highest reported. Thus, more work is required to determine the extent to which label fusion will improve the reliability of our algorithm.

More work is required to determine the source of the slight decrease in segmentation performance when the number of templates are set to an even number. Our initial concern was that this dip in performance was a by-product of the MAGeT-Brain algorithm itself. However, this pattern is also found in the results of the multi-atlas segmentations we used in our experiments. We believe that our majority voting methodology is biased towards labels with the lowest numeric values when breaking ties (by way of the implementation of the mode function used to determine majority), thus causing the slight bias observed when using an even number of templates. This is another area where the voting scheme could be used to improve performance. However, it is worth noting that this limitation was previously identified by Heckemann et al. (2006b) and, subsequently, other groups have not even considered the potential pitfalls of an even number of candidate labels (e.g. Leung et al. (2010)).

Despite MAGeT-Brain achieving segmentation results which are competitive with the rest of the field (Table 8), a concern may be raised over the modest improvement in segmentation agreement observed when the number of templates are set to an even number. Our initial concern was that this dip in performance was a by-product of the MAGeT-Brain algorithm itself. However, this pattern is also found in the results of the multi-atlas segmentations we used in our experiments. We believe that our majority voting methodology is biased towards labels with the lowest numeric values when breaking ties (by way of the implementation of the mode function used to determine majority), thus causing the slight bias observed when using an even number of templates. This is another area where the voting scheme could be used to improve performance. However, it is worth noting that this limitation was previously identified by Heckemann et al. (2006b) and, subsequently, other groups have not even considered the potential pitfalls of an even number of candidate labels (e.g. Leung et al. (2010)).

Despite MAGeT-Brain achieving segmentation results which are competitive with the rest of the field (Table 8), a concern may be raised over the modest improvement in segmentation agreement observed using MAGeT-Brain over multi-atlas, with the same number of atlases (Experiment 1). As we have shown in that same experiment, the benefit in using MAGeT-Brain is both an increase in the overlap agreement and also in the improved consistency of the labelling regardless of atlas or template choice. Reducing the variability in segmentation agreement is an important consideration that few have touched on previously. In addition, the Monte Carlo cross-validations that we present in Experiment 2 (with the first episode psychosis subjects) yields DSC’s that are amongst the highest reported. Thus, more work is required to determine the extent to which label fusion will improve the reliability of our algorithm.

<table>
<thead>
<tr>
<th>Subfield</th>
<th>MAGeT-Brain</th>
<th>Van Leemput et al. (2009)</th>
<th>Yushkevich et al. (2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA1</td>
<td>0.563</td>
<td>0.62</td>
<td>0.875</td>
</tr>
<tr>
<td>CA2/3</td>
<td>0.412</td>
<td>0.74</td>
<td>CA2 = 0.538; CA3 = 0.618</td>
</tr>
<tr>
<td>CA4/DG</td>
<td>0.647</td>
<td>0.68</td>
<td>DG = 0.873</td>
</tr>
<tr>
<td>Presubiculum</td>
<td>0.58</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>Subiculum</td>
<td>0.58</td>
<td>0.74</td>
<td>0.770</td>
</tr>
<tr>
<td>Hippocampal fissure</td>
<td>–</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>SR/SL/SM</td>
<td>0.428</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Fimbria</td>
<td>–</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Tail</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 7. Detailed subfield segmentation results for a single subject. In the upper left corner is the original high-resolution Winterburn atlas manual subfield segmentation; in the upper right corner is the Winterburn atlas segmentation subsampled from 0.3 mm- to 0.9 mm-isotropic voxels; in the lower left corner is the MAGeT-Brain segmentation of the subsampled Winterburn atlas image from a single fold of the cross-validation. In each segmentation, slices from the left hemisphere are shown in Talairach-like ICBM152 space: the first row shows axial slices from inferior to superior; the second row shows sagittal slices from lateral to medial; the third row shows coronal slices from anterior to posterior.
implementation does not require re-tuning for new populations or datasets as it inherently models the variability of the dataset through the template library. However it should be noted that the increased agreement that follows increasing the number of atlases and templates comes at an increased computational cost (O(log(n))), as previously mentioned in other work (Heckemann et al., 2006a).

Amongst the automated segmentation methods we compared in this paper (FreeSurfer, MAPER, FSL FIRST), we find extremely variable performance of all methods. With the exception of FSL FIRST all methods correlate well with the semi-automated SNT volumes provided in the ADNI database. However, the FreeSurfer and FSL FIRST hippocampal segmentations are on average about twice the volume of those from all other methods. Furthermore, when estimating the bias of FreeSurfer and FSL FIRST relative to the SNT hippocampal volumes we see that large hippocampi are over estimated whilst small hippocampi are under estimated. By comparison, MAGeT-Brain and MAPER are far more conservative in volume estimation, suggesting that these methods may be better suited for estimating true-positives, especially in neurodegenerative disease subjects featuring smaller overall hippocampi. However, in this analysis we have only compared methods by total hippocampal volume, and so more work is needed to understand the full extent to which these methods differ.

Finally, we have provided evidence that using the Winterburn high-resolution hippocampal subfield atlases (Winterburn et al., 2013) our algorithmic framework is appropriate for the segmentation of hippocampal subfields in standard T1-weighted data. Subfield segmentation is a burgeoning topic in the literature although very few automated methods are available for the segmentation of 3D data (Van Leemput et al., 2009; Yushkevich et al., 2009, 2010). Table 10 compares segmentation agreement from some of these methods and MAGeT-Brain. The overlap DSC scores for MAGeT-Brain subfields are notably lower but a direct comparison of overlap values must be done cautiously. In the present work, our overlap scores are computed on 0.9 mm-isotropic voxel resolution images, whereas Yushkevich et al. (2010) uses focal 0.4 × 0.5 × 2.0 mm voxel resolution images, and Van Leemput et al. (2009) use supersampled 0.3809 mm-isotropic voxel resolution images. The larger voxel images we use necessarily entail a greater change in DSC for each incorrectly labelled voxel. In addition, our automated segmentations are compared to manual segmentations resampled from 0.3 mm-isotropic voxel labels; the resampling process inevitably introduces noise which may lower overlap scores. Lastly, as our method is aimed specifically at situations when manually produced atlases are scarce, in our cross validation we are forced to use three rather than all five of the Winterburn atlases (which, based on our findings with whole hippocampal segmentation, would have resulted in improved overlap similarity). Although having more atlases would be ideal in this context, these atlases are very time consuming to generate (Winterburn et al., 2013). Nevertheless, the advantage of evaluating MAGeT-Brain on standard 3T T1-weighted resolution MR images with a publicly available atlas library is that our results reflect typical usage scenarios of researchers and clinicians.

Experiments 1, 2, and 5 have demonstrated that our algorithm flexibly accommodates different whole hippocampus manual segmentation methodologies. We have not explicitly evaluated a subfield definition other than the Winterburn protocol, and therefore it is possible that using an alternate subfield definition could improve the reliability of our automated subfield definitions. For example, established definitions such as those from Mueller et al. (2007) could be a prime candidate for further exploration. In addition, the conservative nature of the Mueller definition (labelling of the 5 slices in the hippocampus body only) would likely further aid in reliability measurement. However, there are two main logistical problems that we would have to overcome prior to implementation. The first is that these definitions were developed for data that is highly anisotropic (0.4 × 0.5 × 2 mm), and it is unclear how our algorithms would deal with such atlases used as input. The second is that, since these atlases are not publicly available, we would have to re-implement the protocol using our atlases. At the present time it is unclear how we would adapt these protocols to data that we used, where subfield segmentations are defined on 0.3 mm³ voxels. However, the impact of subfield definitions in the context of our work is an important one and should be considered in subsequent studies.

One further complication common to all subfield segmentation evaluation is that, by its nature, the Dice’s Similarity Coefficient score penalizes structures with high surface area-to-volume ratios. Therefore subfield DSC scores will generally be lower than whole hippocampal segmentations. We attempted to put this effect into perspective by comparing MAGeT-Brain subfield segmentation agreement with the agreement of voxel-shifted manual segmentations (Table 7). The results of this exercise show conclusively, despite the very limited number of atlases we had to work with, that MAGeT-Brain subfield segmentations are well within the bounds of error of a 0.3 mm³ voxel shift.

Our overlap DSC values demonstrate that we can reliably reproduce segmentations for the CA1, subiculum, and CA4 dentate subfields (DSC > 0.5). That the CA2/CA3 and molecular layers are less well reproduced (DSC > 0.5) should not be surprising as these are extremely thin and spatially convoluted regions that originally required high-resolution MRI for identification and so it is likely that the extents of these regions are well below the resolution and contrast offered by standard T1-weighted images.

This points to a larger issue of how to truly validate subfield segmentations, both in high resolution images and in standard T1-weighted images. There are several manual subfield segmentation methodologies, and they do not agree on which regions can be differentiated, even on high-resolution scans. See Table 9 for a comparison of MRI-based manual subfield segmentation methodologies. A further complication is that different researchers have differing operational definitions for the subfields and how they ought to be parcellated. The disagreement in the community has led to an international working group devoted to normalising the ontology and segmentation rules for the hippocampal subfields9. In addition, there have been recent advances from the Yushkevich group to revise their MRI subfield segmentation protocol based on anatomy discerned from serial histological acquisitions (Adler et al., 2014). The definitional and operational disagreements suggest that direct comparison across automated methods using “ground truth”-based overlap similarity metrics, such as Dice’s Similarity Coefficient, are not possible without carefully taking into account the differences in underlying segmentation protocols and image characteristics.

In conclusion, we have demonstrated the viability of leveraging a small number of input atlases to generate a large template library and thereby improve segmentation reliability when using multi-atlas methods. We demonstrated that this method works robustly over hippocampal definitions, different disease populations, and different acquisition types. Finally, we also demonstrate that reliable reproduction of hippocampal subfield segmentations in standard 3T T1-weighted images is possible.

**Acknowledgments**

We wish acknowledge support from the CAMH Foundation, thanks to Michael and Sonja Koerner, the Kimel Family, and the Paul E. Garfinkel New Investigator Catalyst Award. MMC is funded by the W. Garfield Weston Foundation and ANV is funded by the Canadian Institutes of Health Research, Ontario Mental Health Foundation, NARSAD, and the National Institute of Mental Health (R01MH099167).

Computations were performed on the gpc supercomputer at the SciNet HPC Consortium (Loken et al., 2010). SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute

---

9 [http://www.hippocampalsubfields.com/).
Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.04.054.

References


