Bootstrapping Multi-atlas Hippocampal Segmentation with MAGeT-Brain

Jon Pipitone[1], Matt T. Park[1], Julie Winterburn[1], Tristram A. Lett[1], Jason P. Lerch[2,3],
Jens C. Pruessner[4], Martin Lepage[4], Aristotle Voineskos[1,5], M. Mallar Chakravarty[1,5,6] and
Alzheimer's Disease Neuroimaging Initiative

[1] *Kimel Family Translational Imaging-Genetics Lab, Centre for Addiction and Mental Health, Toronto, ON, Canada*
[2] *Neurosciences and Mental Health, Hospital for Sick Children, Toronto, ON,Canada*
[3] *Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*
[4] *Brain Imaging Group, Douglas Hospital Research Centre, Verdun, QC, Canada*
[5] *Department of Psychiatry, University of Toronto, Toronto, ON, Canada*
[6] *Institute of Biomedical Engineering, University of Toronto, Toronto, ON, Canada*

Abstract

Contact:
Jon Pipitone and M. Mallar Chakravarty
Kimel Family Translation Imaging-Genetics Research Laboratory
Research Imaging Centre
Centre for Addiction and Mental Health
250 College St.
Toronto, Canada M5T 1R8
jon.pipitone@camh.ca; mallar.chakravarty@camh.ca

1   Introduction

The hippocampus is a neuroanatomical structure situated in the medial temporal lobe of the brain, and has long been associated with learning and memory (den Heijer et al., 2012; Scoville and Milner, 2000). In addition to its known functional roles, the hippocampus is of interest to neuroscientists because it is implicated in several forms of brain disfunction such as Alzheimer's disease (Sabuncu et al., 2011) and schizophrenia (Narr et al., 2004; Karnik-Henry et al., 2012). In neuroimaging experiments, magnetic resonance images (MRI) are often used for the identification of the hippocampus. As such, accurate segmentation of the hippocampus and its subregions in MRI is a necessary first step to better understand the unique neuroanatomy of subjects. Typically, the gold standard for neuroanatomical segmentation is manual delineation by an expert human rater. However the rapid increase in the availability of MRI data and the time and expertise required for manual segmentation is prohibitive (Mazziotta et al., 1995, 2001; Mazziotta et al.; Pausova et al., 2007). Further, there is little agreement between researchers regarding how exactly the hippocampus should be identified in MRI images (Geuze et al., 2004) and this has led to efforts to create an unified hippocampal segmentation protocol (Jack et al., 2011).

Automated segmentation techniques that are reliable, objective, and reproducible are a necessary alternative to manual segmentation. In the case of classical model-based segmentation methods (Haller et al., 1997; Csernansky et al., 1998), an MRI atlas that was previously manually labelled by an expert rater is matched to target images using nonlinear registration methods. The resulting nonlinear transformation is applied to the manual labels (i.e. *label propagation*) to move them into the target image space. While this methodology has been used successfully in several contexts (Chakravarty et al., 2008, 2009; Collins et al.,

1995; Haller et al., 1997), it is limited in accuracy by the introduction of errors due inaccuracies in the non-linear transformation itself, partial volume effects in label resampling, and irreconcilable differences between the neuroanatomy represented within the atlas and target images.

One methodology that can be used to mitigate these sources of errors involves the use of multiple manually segmented atlases and probabilistic segmentation techniques, such as those found in the FreeSurfer package (Fischl et al., 2002). FreeSurfer uses a probabilistic atlas of anatomical and tissue classes along with spatial constraints for class labels encoded using a Markov random field model to segment the entire brain.

More recently, many groups have been using multiple atlases to improve overall segmentation accuracy (i.e. multi-atlas segmentation) over model-based approaches (Heckemann et al., 2006a, 2011; Collins and Pruessner, 2010; Lötjönen et al., 2010; Aljabar et al., 2009; Leung et al., 2010; Wolz et al., 2010). Each atlas image is registered to a target image, and label propagation is performed to produce several labellings of the target image (one from each atlas). A *label fusion* technique, such as voxel-wise voting, is used to merge these labels into a definitive segmentation for the target. In addition, weighted voting procedures that use *atlas selection* techniques are often used to exclude atlases from label fusion that are dissimilar to a target image in order to reduce error from unrepresentative anatomy (Aljabar et al., 2009). This involves the selection of a subset of atlases using a similarity metric such a cross-correlation (Aljabar et al., 2009) or normalized mutual information. Such selection has the added benefit of significantly reducing the number of nonlinear registrations. For example Collins and Pruessner (2010) demonstrated that only 14 atlases, selected based on highest similarity between medial temporal lobe neuroanatomy as evaluated by normalized mutual information (Studholme et al., 1999) from a library of 80 atlases, were required to achieve accurate segmentations of the hippocampus. Several methods have been explored for label fusion including the STAPLE algorithm (Warfield et al., 2004) that computes a probabilistic segmentation using an expectation maximization framework from an set of competing segmentations; or others where a subset of segmentations can be estimated using metrics such as the sum of squared differences in the regions of interest to be segmented (Coupé et al., 2012).

However, many of these methods require significant investment of time and resources for the creation of the atlas library; ranging from atlas libraries that require between 30 (Heckemann et al., 2006a) and 80 (Collins and Pruessner, 2010) manually segmented atlases. This strategy has the main drawback of being inflexible as it does not easily accommodate varying the definition of the hippocampal anatomy (such as the commonly used heuristic of subdividing the hippocampus in to head, body, and tail (Poppenk and Moscovitch, 2011; Pruessner et al., 2000)). Furthermore, none of these methods have demonstrated sufficient flexibility to accommodate atlases that are somehow exceptional such as those derived from serial histological data (Chakravarty et al., 2006; Yelnik et al., 2007) or high-resolution MRI data that enables robust identification of hippocampal subfields (Winterburn et al., 2013; Yushkevich et al., 2009; Mueller and Weiner, 2009; Van Leemput et al., 2009; Wisse et al., 2012). Due to the recent availability of the latter, there has been increased interest in the use of probabilistic methods for the identification of the hippocampal subfields on standard T1-weighted images. Our group recently demonstrated that through use of an intermediary automated segmentation stage, robust and accurate segmentation of the striatum, pallidum, and thalamus using a single atlas derived from serial histological data is possible (Mallar Chakravarty et al., 2012). The novelty of this manuscript is the extension our multi-atlas methodology to the hippocampus using more than a single input atlas, while simultaneously limiting the number of possible inputs used during segmentation, and demonstrating that accurate identification of the hippocampal subfields is indeed possible using this methodology.

There are few methods that have attempted to perform multi-atlas segmentation with a limited number of input atlases. The LEAP algorithm is an elegant modification to the basic multi-atlas strategy (Wolz et al., 2010) in which the atlas library is grown, beginning with a set of manually labelled atlases, and successively incorporating unlabelled target images after themselves being labelling 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 ADNI-1 baseline dataset, achieving a mean Dice score of 0.85 with manual segmentations.

To the best of our knowledge there are two other segmentation methods that attempt to define the

hippocampal subfields using standard T1-weighted data. The first is included with the FreeSurfer package (Van Leemput et al., 2009). This work is limited as it omits the tail of the hippocampus and the segmentation protocol has yet to be fully validated. Nonetheless, they demonstrate that the applicability of their work using data from 10 subjects. In the second method, Yushkevich et al. (2009) acquired and labelled hippocampal subfields on high-resolution MRI data from post-mortem medial temporal lobe samples. Using nonlinear registration guided by manually derived hippocampus masks and specific landmarks, they demonstrate accurate parcellation of hippocampal subfields in unlabelled MRI volumes.

Here we address the issue of limiting the number of input atlases by tuning our algorithm, for segmentation of the entire hippocampus, using a multi-fold experiment performed on a subset Alzheimer's Disease Neuroimaging Initiative (ADNI) 1 dataset. Based on the parameters we find in our experiment, we validate our algorithm using all of the data available in the ADNI Complete 1Yr sample and compare our segmentations to the other segmentations that are available through the ADNI informatics portal. To ensure that we have not over-fit our parameters to the aging or neurodegenerative brain, we also apply our segmentations to a dataset of normal controls and individuals suffering from first episode psychosis. Finally, we perform a leave-one-out validation experiment to determine if the subfields can be accurately identified using our multi-atlas framework.

2 Methods

2.1 The MAGeT-Brain Algorithm

In this paper, we will use the term *label* to mean any segmentation (manual or derived) of an MR image. *Label propagation*, or *labelling*, 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 will 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 will use the term *target* to refer to an unlabelled image that is undergoing segmentation.

The simplest form of multi-atlas segmentation, (which we will call *basic multi-atlas segmentation*), involves three steps. First, each labelled input image (i.e. atlas or template) is registered to an unlabelled target image. Second, the labels from each image are propagated to the target image space. Third, the labels are combined into a single labelling by way of a label fusion method (Heckemann et al., 2006a, 2011). This method is described in detail in other publications (Collins and Pruessner, 2010; Heckemann et al., 2011; Aljabar et al., 2009).

MAGeT-Brain bootstraps the creation of a large template library given a limited input atlas library, and then uses the template library in basic multi-atlas segmentation. Images for the template library are selected from a set of input 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 then labelled by each of the atlases. Basic multi-atlas segmentation is then conducted using the template library to segment the entire set of target images (including the targets whose images are 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 atlas $\times$ # of templates).

Figure 1 describes the MAGeT-Brain algorithm in pseudocode. Source code for MAGeT-Brain can be found at `http://github.com/pipitone/MAGeTbrain`.

2.2 Experiments

The following section describes the experiments we conducted to assess the segmentation quality of the MAGeT-Brain algorithm. The first two experiments assess the validity of MAGeT-Brain using a cross-validation design. Experiment 1 investigates the accuracy of MAGeT-Brain whole hippocampus segmentation over a wide range of parameter settings. The results of this experiment enable us to choose the parameter settings offering the best performance for use in subsequent experiments. Experiment 2 tests hippocampal subfield segmentation quality. The last two experiments assess the validity of the MAGeT-Brain

---
Algorithm 1 Pseudocode for the MAGeT-Brain algorithm

```
function BASICMULTIATLASSEGMENTATION(Templates, Subjects)
    for all target do
        for all template do
            propagate all labels for template to target space
            store target labels
        end for
        fuse target labels
    end for
end function

function MAGETBRAIN(Subjects, Atlases, n)
    for i = 1 → n do
        choose a target to be used as a template
        propagate labels from each atlas to template space
        store the template with all of its labels
    end for
    MultiAtlas(Templates, Subjects)
end function
```
---

algorithm when applied to different diseases: first episode schizophrenia (Experiment 3), and Alzheimer's disease (Experiment 4). Additionally, in experiment 4, we compare MAGeT-Brain segmentations with those of well-known automated and manual methods and assess segmentation bias.

2.2.1 Experiment 1: Whole Hippocampus Cross-Validation

Monte Carlo Cross-Validation (MCCV) (Shao, 1993) was performed using a pool of images and manual hippocampal segmentations from ADNI1 dataset. This form of cross-validation allows us to rigorously validate a large number of parameter settings of MAGeT-Brain (atlas and template library sizes, registration algorithm, and label fusion method) and select the best parameters to use in subsequent experiments.

ADNI1 dataset 69 1.5T images were arbitrarily selected from the baseline scans in the *ADNI1:Complete 1Yr 1.5T* standardized dataset. 23 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. Manual segmentations of the left and right whole hippocampi are available (Hsu et al., 2002). These labels have been generated using the SNT tool from Medtronic Surgical Navigation Technologies, Louisville, CO (see Supplementary Materials for detailed discussion of the manual segmentation process used).

Clinical, demographic and pre-processed T1-weighted MRI were downloaded by the authors from the ADNI database (adni.loni.ucla.edu) between March 2012 and August 2012. The image dataset download was the "ADNI1:Complete 1Yr 1.5T " standardized dataset available from ADNI [1] (Wyman et al., 2012). This image collection contains uniformly preprocessed 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 = 2400ms, TI = 1000ms, TE = 3.5ms, flip angle = 8∘, field of view = 240 x 240mm, a $192 \times 192 \times 166$ matrix (x, y, and z directions) yielding a voxel resolution of $1.25 \times 1.25 \times 1.2mm^3$.

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 69 1.5T images and corresponding manual segmentations. In each validation round, the dataset is partitioned into a training set consisting of images and their manual labels to be used as an atlas library, and a validation set consisting of the remaining images segmented by both MAGeT-Brain and multi-atlas. The resulting segmentations are compared to the manual segmentations for the images.

---
[1] http://adni.loni.ucla.edu/methods/mri-analysis/adni-standardized-data/

Table 1: **ADNI-1 cross-validation subset demographics.** Clinical Dementia Rating-Sum of Boxes (CDR-SB), Alzheimer's Disease Assessment Scale (ADAS), MiniMental State Examination (MMSE).

	CN $N = 23$	LMCI $N = 23$	AD $N = 23$	Combined $N = 69$
Age at baseline Years	72.2 75.5 78.5	71.0 77.1 81.4	71.7 77.8 81.8	71.5 76.6 81.3
Sex : Female	43% (10)	43% (10)	43% (10)	43% (30)
Education	16.0 16.0 18.0	15.0 16.0 18.0	12.0 16.0 16.5	14.0 16.0 18.0
CDR-SB	0.00 0.00 0.00	0.75 1.50 1.50	4.00 4.50 5.00	0.00 1.50 4.00
ADAS 13	4.67 5.67 12.34	14.34 16.00 20.50	23.83 29.00 31.66	10.00 16.00 25.33
MMSE	28.5 29.0 30.0	25.0 27.0 28.0	21.0 23.0 24.0	24.0 27.0 29.0

$a\ b\ c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Numbers after percents are frequencies.

Table 2: **ANIMAL registration parameters.**

Parameters	Stage 1	Stage 2	Stage 3
Model Blur (FWHM)	8	8	4
Input Blur (FWHM)	8	8	4
Iterations	30	30	10
Step	8x8x8	4x4x4	2x2x2
Sub-Lattice	6	6	6
Lattice Diameter	24x24x24	12x12x12	6x6x6

A total of ten validation rounds are performed on each subject in the dataset, over each combination of parameter settings. The parameter settings we explore are: atlas library size (1-9), template library size (1-20), registration method (ANTS or ANIMAL), and label fusion method (majority vote, cross-correlation weighted majority vote, and normalized mutual information weighted majority vote). A total of $10 \times 69 \times 9 \times 20 \times 2 \times 3 = 7\,452\,000$ validation rounds were conducted, resulting in a total of $1\,490\,400$ segmentations analysed.

Before registration, all images underwent preprocessing with the N3 algorithm (Sled et al., 1998) to minimize intensity nonuniformity. In this experiment we use one of two non-linear image registration methods.

Automatic Normalization and Image Matching and Anatomical Labeling (ANIMAL) The ANIMAL algorithm carries out image registration is in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that maximizes the correlation between blurred MR intensities and gradient magnitude over the whole brain (Collins et al.). 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 blurred version of the input data. These larger 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 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 optimized 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 `http://www.picsl.upenn.edu/ANTS/`. We used an implementation of the ANTS algorithm compatible with the MINC data format, mincANTS `https://github.com/vfonov/mincANTS`.

We used the following command line when running ANTS :

```
mincANTS 3 -m PR[target_file.mnc,source_file.mnc,1,4]
 --number-of-affine-iterations 10000x10000x10000x10000x10000
 --affine-gradient-descent-option 0.5x0.95x1.e-4x1.e-4
 --use-Histogram-Matching --MI-option 32x16000
 -r Gauss[3,0] -t SyN[0.5] -i 100x100x100x20
 -o transformation.xfm
```

These settings were adapted from the "reasonable starting point" given in the ANTS manual [2].

Label fusion methods Label fusion is a term given to the process of combining the information from several candidate labellings for an intensity 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 labellings. Ties are broken arbitrarily.

Cross-correlation Weighted Majority Vote An optimal combination of targets from the template library has previously been shown to improve segmentation accuracy (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 normalized 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 (Mallar Chakravarty et al., 2012). The number of top ranked 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[3] is used to calculate the normalised mutual information measure between to images.

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

$$\text{Dice's coefficient (DSC)} = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B are the regions being compared, and the cardinality is the volume measured in voxels.

The manual segmentations (SNT) provided as part of the ADNI dataset are used as the gold standard to compare with (Hsu et al., 2002). The segmentation accuracy reported is averaged over the ten validation rounds for each parameter setting.

In order to investigate the performance of MAGeT-Brain in a real world setting in which only one set of atlas and template images are used, we explore the variability in label agreement at fixed parameter settings when the choice for atlas and template images is varied. This is achieved by first computing the standard deviation and variance of DSC scores in each block of ten validation rounds per subject. The distribution of these statistics 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.

---

[2]https://sourceforge.net/projects/advants/files/Documentation/
[3]https://github.com/vfonov/EZminc

6

Table 3: **ADNI-1 cross-validation subset demographics.** Clinical Dementia Rating-Sum of Boxes (CDR-SB), Alzheimer's Disease Assessment Scale (ADAS), MiniMental State Examination (MMSE).

	N	
Age	16	31.0 53.0 63.8
Sex : Male	16	38% (6)
Education : 0.01	15	7% (1)
12		13% (2)
13		13% (2)
14		20% (3)
16		13% (2)
18		33% (5)
Handedness : R	16	94% (15)

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

2.2.2 Experiment 2: Hippocampal Subfield Cross-Validation

In this experiment, the accuracy of the MAGeT-Brain algorithm on hippocampal subregion segmentation is tested using a leave-one-out cross-validation (LOOCV) design. The optimal parameter settings for MAGeT-Brain found in Experiment 1 are used in this experiment.

Winterburn Atlases dataset   The Winterburn atlases (Winterburn et al., 2013) are digital hippocampal segmentations of five in-vivo 300μ isotropic T1-weighted MR images. The segmentations include subfield segmentations for the cornus ammonis (CA) 1, CA4, dentate gyrus, subiculum, and CA 2 and 3 combined. Subjects in the Winterburn atlases range in age from 29-57 years (mean age of 37), and include two males and three females.

In addition to the high-resolution scans distributed as part of the Winterburn atlases, we also obtained additional 3T T1 BRAVO images (0.9mm-isotropic voxels) of four of the five Winterburn atlas subjects.

Experiment details   Leave-one-out cross-validation (LOOCV) is an approach in which the method to be validated is given all but one item in a dataset as training data, and the output is compared with the left-out item. This is done, in turn, for each item in the dataset.

In this experiment the Winterburn atlases are used as the cross-validation dataset. The five 300μ-isotropic voxel images and labels are used as atlases, and LOOCV is conducted using the five Winterburn atlas subject images subsampled (using trilinear interpolation) to 0.9mm-isotropic voxel resolution (referred to as the *Subsampled* dataset) as input subjects. Subsampling the subject images allows us to assess MAGeT-Brain in a typical segmentation scenario (high-resolution atlases and lower-resolution subjects). The template library is composed of the subject images, plus an additional set of 3T T1 images (0.9mm-isotropic voxels) of healthy subjects acquired separately (Table 3). The optimal size of template library, registration method, and label fusion method found in Experiment 1 are used. Each resampled Winterburn atlas subject image is segmented by MAGeT-Brain with that subject's image excluded from the atlas library.

We reproduce this experiment in a separate LOOCV run in which the resampled Winterburn atlas subject images are substituted for separately acquired T1 BRAVO images of four of the subjects (referred to as the *BRAVO* dataset).

Evaluation method   To assess the MAGeT-Brain LOOCV segmentations we compute the relative percent error in hippocampal volume with the full resolution Winterburn atlas segmentations. In addition, by computing the relative error in volume of the Winterburn atlas labels resampled (with nearest-neighbour interpolation) to 0.9mm-isotropic voxels, we obtain a baseline error to assess against.

2.2.3 Experiment 3: Application to the segmentation first episode schizophrenia patients

To validate that MAGeT-Brain algorithm works effectively in the context of other neuropsychiatric disorders, we use the Winterburn atlases with MAGeT-Brain to predict the hippocampal segmentation of dataset of

Table 4: **Schizophrenia First Episode Patient Demographics.** Clinical Dementia Rating-Sum of Boxes (CDR-SB), Alzheimer's Disease State Examination (MMSE).

	N	FEP
		$N = 81$
Age	80	21 23 26
Gender : M	81	63% (51)
Handedness : ambi	81	6% (5)
left		5% (4)
right		89% (72)
Education	81	11 13 15
SES : lower	81	31% (25)
middle		54% (44)
upper		15% (12)
FSIQ	79	88 102 109

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

Schizophrenia patient MR images. The resulting segmentations are assessed for quality by comparison with expert manual segmentations.

SZ-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 1530 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 program details see Malla 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=22ms, echo time=9.2ms, flip angle=30◦, 180 1mm contiguous sagittal slices). The rectangular field-of-view for the images was 256mm (SI)×204mm (AP). Subject demographics are shown in Table 4.

Each subject hippocampus is traced following a validated segmentation protocol (Pruessner et al., 2000).

Experiment details MAGeT-Brain is configured with an atlas library composed of the Winterburn T1 atlases (see Experiment 2). All images from the SZ-FEP dataset are segmented by MAGeT-Brain . The optimal size of template library, registration method, and label fusion method found in Experiment 1 are used.

Evaluation method The manual and Winterburn hippocampal segmentation protocols differ slightly in the neuroanatomical features that are delineated. This poses a problem for evaluation by measuring overlap. That is, since different protocols will necessarily produce segmentations that do not perfectly overlap, the degree of overlap cannot be solely used to compare segmentation methods using different protocols. In place of an overlap metric, we can assess the degree of correlation in average hippocampal volume of the subjects produced by each method. Specifically, Pearson correlation is used.

2.2.4 Experiment 4: Application to the segmentation of Alzheimer's disease patients

To validate that MAGeT-Brain algorithm works as well as established automated methods, MAGeT-Brain is applied to the ADNI1 dataset and the resulting segmentations are compared to those produced by FreeSurfer, FSL, MAPER, and by expert manual segmentation.

Table 5: **ADNI1 1.5T Complete 1Yr dataset demographics.** Clinical Dementia Rating-Sum of Boxes (CDR-SB), Alzheimer's Disease Assessment Scale (ADAS), MiniMental State Examination (MMSE).

	N	CN $N = 584$	LMCI $N = 931$	AD $N = 404$	Combined $N = 1919$
Age at baseline Years	1919	72.4 75.8 78.5	70.5 75.1 80.4	70.1 75.3 80.2	71.1 75.3 79.8
Sex : Female	1919	48% (278)	35% (327)	49% (198)	42% (803)
Education	1919	14 16 18	14 16 18	12 15 17	13 16 18
CDR-SB	1911	0.0 0.0 0.0	1.0 1.5 2.5	3.5 4.5 6.0	0.0 1.5 3.0
ADAS 13	1895	5.67 8.67 12.33	14.67 19.33 24.33	24.67 30.00 35.33	10.67 18.00 25.33
MMSE	1917	29 29 30	25 27 29	20 23 25	25 27 29

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

ADNI1 dataset revisited All images from the *ADNI1:Complete 1Yr 1.5T* standardized dataset described in Experiment 1 are used. Clinical and demographic data are shown in Table 5.

Experiment details MAGeT-Brain is configured with an atlas library composed of the Winterburn T1 atlases (see Experiment 2). All images from the ADNI1:Complete 1Yr 1.5T dataset are segmented. The optimal size of template library, registration method, and label fusion method found in Experiment 1 are used. The template library is composed of equal numbers of images from each disease class (AD, MCI, and cognitively normal controls).

Evaluation method As in Experiment 3, the manual (SNT) and Winterburn hippocampal segmentation protocols differ in the neuroanatomical features delineated, and so we must assess MAGeT-Brain by the degree of correlation of average hippocampal volume across all subjects produced by MAGeT-Brain and by manual segmentation. Specifically, Pearson correlation is used. For comparison, we also compute the correlation in hippocampal volume between the existing, established automated segmentation methods: FSL, FreeSurfer, and MAPER. Additionally, we evaluate the volume-related fixed and proportional biases in all segmentation methods using Bland-Altman plots (Bland and Altman, 1986).

3 Results

3.1 Experiment 1 Results: Whole Hippocampus Cross-Validation

In this experiment we conducted 10 rounds of MAGeT-Brain and multi-atlas segmentation of each of 69 subjects at a range of atlas and template library sizes, registration algorithm (ANTS or ANIMAL), and three label fusion techniques. Hippocampal MAGeT-Brain -based segmentations using both ANIMAL and ANTS registration algorithm demonstrate good overlap with SNT derived gold-standards (Figure 1). Qualitatively, both ANIMAL and ANTS -based segmentations demonstrate trend overlap accuracy that increases with the size of atlas library and template library. Improvement in accuracy diminishes noticeably with template libraries larger than roughly ten images.

No marked difference in segmentation accuracy is seen when either ANIMAL or ANTS registration is used with any number of atlases or templates. In every parameter configuration, the use of MAGeT-Brain with ANTS registration shows a pronounced increase in segmentation accuracy over MAGeT-Brain with ANIMAL registration. Surprisingly, the label fusion method used does not significantly improve label accuracy, contrary to the findings of Aljabar et al. (2009) when using weighted voting on much larger atlas/template libraries. In the remainder of this section, only results using the ANTS registration algorithm and majority vote fusion will be shown.

With an increasing number of templates, MAGeT-Brain shows improvement in overlap accuracy over multi-atlas-based segmentation when using the same number of atlases and voting method (Figure 2). The magnitude of improvement over multi-atlas-based segmentation decreases with an increasing number of atlases, with accuracy converging with 7 atlases. Peak improvement in MAGeT-Brain accuracy (0.02 DSC) is found when one atlas is used with a template library of 20 images.

In addition to an improvement in accuracy over multi-atlas-based segmentation, MAGeT-Brain also shows a decrease in the variability of segmentation accuracy (Figure 3). The size of template library necessary to reach a significant ($p < 0.05$) decrease in variance and standard deviation grows with the size of atlas library used. A template library of 19 images is sufficient to show significant decrease in variance and standard deviation for 3-7 atlases.

We have omitted results obtained when using an even number of atlases or templates since with this configuration we found significantly decreased performance. We believe this is as a result of an inherent bias in the majority vote fusion method used (see Discussion).
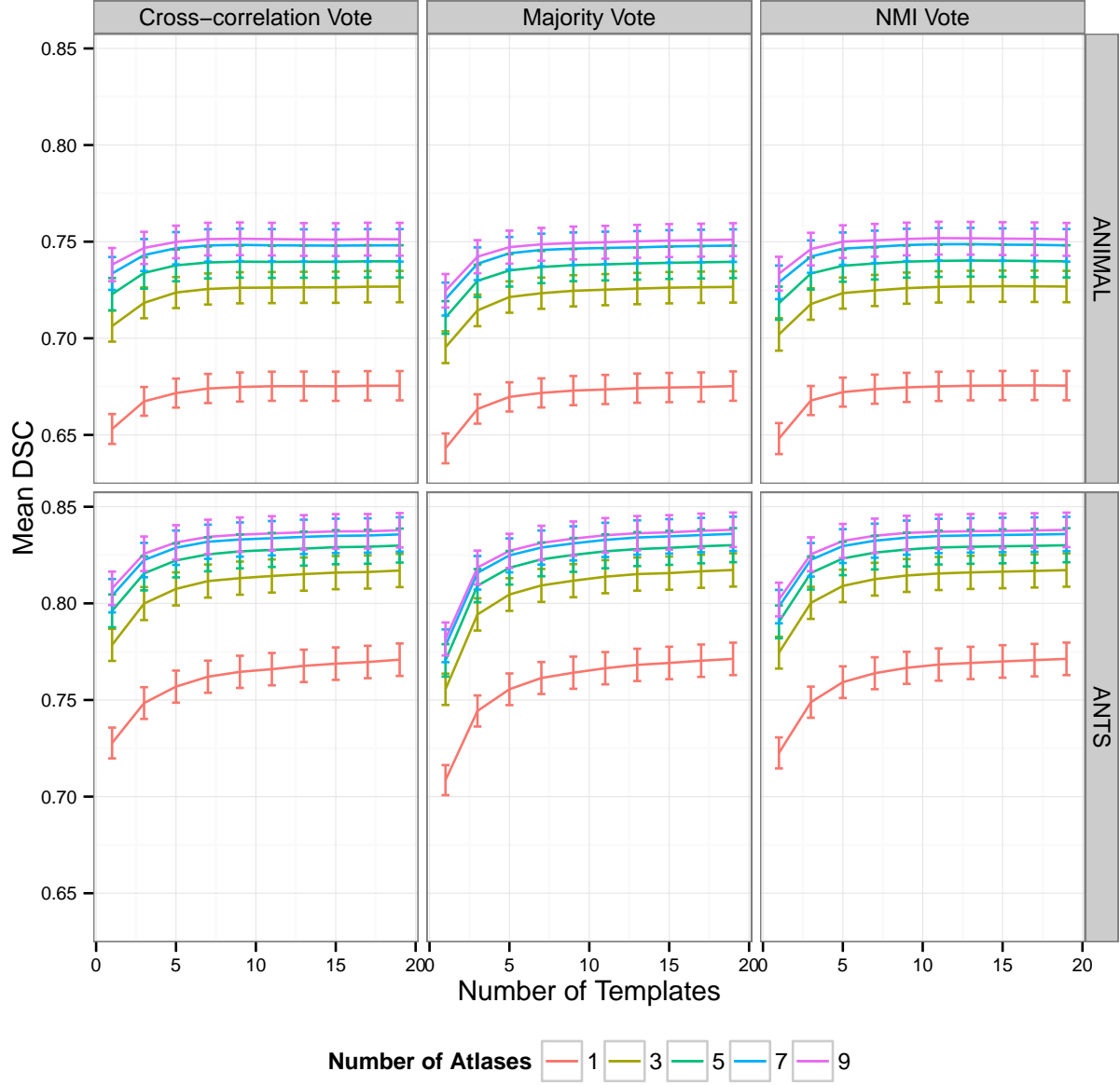
Figure 1: Mean Dice's Similarity Coefficient of MAGeT-Brain segmentations with manual segmentations (SNT) for 69 ADNI1 subjects vs. atlas and template library size, registration algorithm, and label fusion method. Points above zero indicate an MAGeT-Brain parameter settings yielding a higher mean DSC score than multi-atlas segmentation using the same number of atlases.
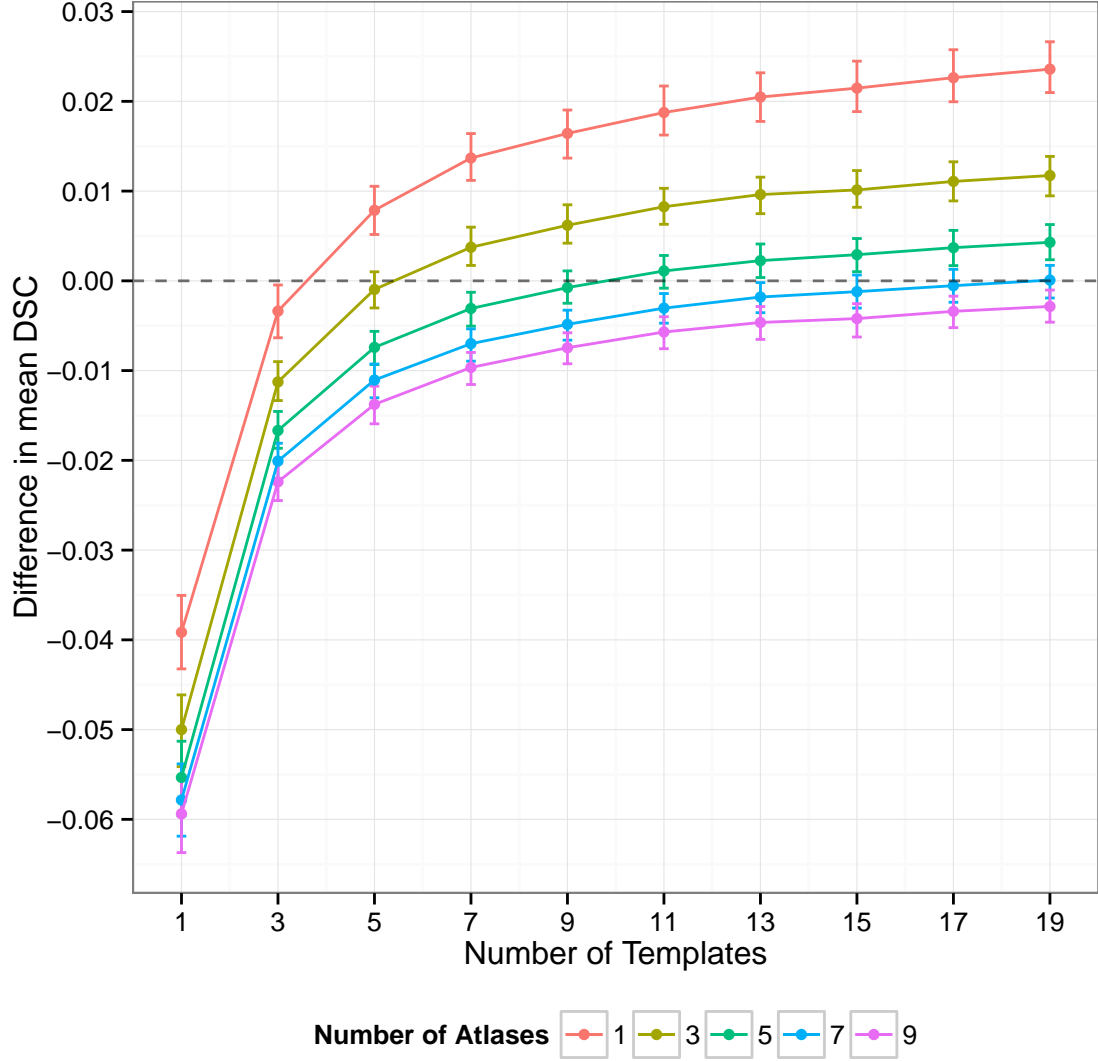
Figure 2: Difference in mean Dice's Similarity Coefficient of MAGeT-Brain and multi-atlas segmentations with manual (SNT) segmentations for a range of atlas and template library sizes.
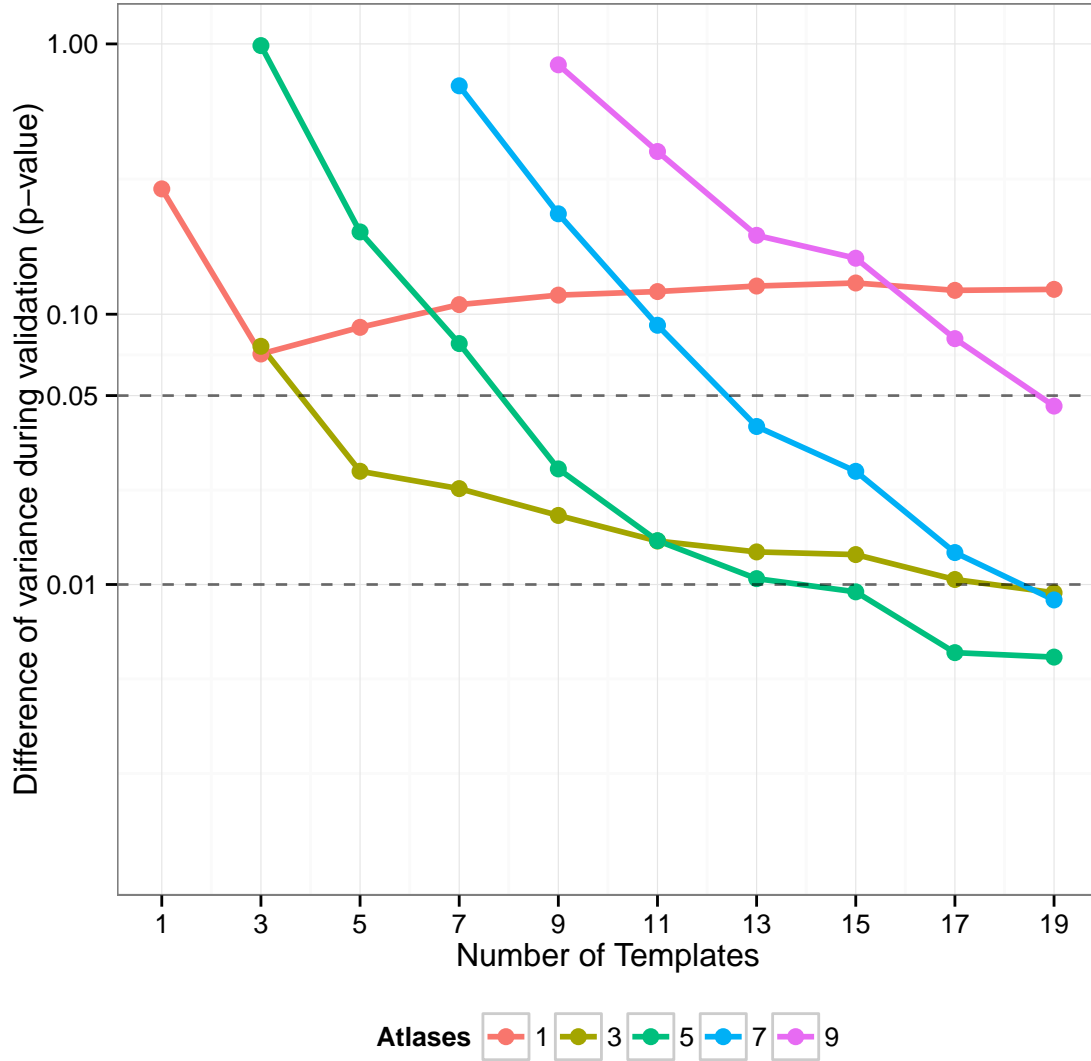
Figure 3: **Difference in variability of MAGeT-Brain vs. multi-atlas segmentation accuracy.** Variance of segmentation accuracy between MAGeT-Brain and multi-atlas segmentation is computed for each subject across all ten rounds of validation. Shown on the y-axis (scaled logarithmically) is the p-value resulting from a t-test comparing the distribution of variances at each parameter setting (atlas/template library size). Only points where MAGeT-Brain mean variability is lower than multi-atlas are shown.

3.2 Experiment 2 Results: Winterburn Atlases Cross-Validation

This experiment explores MAGeT-Brain segmentation of hippocampal subfields. To achieve this, a leave-one-out validation is conducted in which lower-resolution images ($0.9mm^3$ voxels) of each Winterburn atlas subject are segmented using the remaining Winterburn atlases.

In general, across hippocampal subregions the percent error in volume of MAGeT-Brain segmentations compares favourably to the error resulting from image resampling (Figure 4). In particular, the CA1, CA4, and Dentate subregions all show near or smaller percent errors. The Sibiculum and CA2/CA3 subregions show distinctly larger error than is found through resampling.

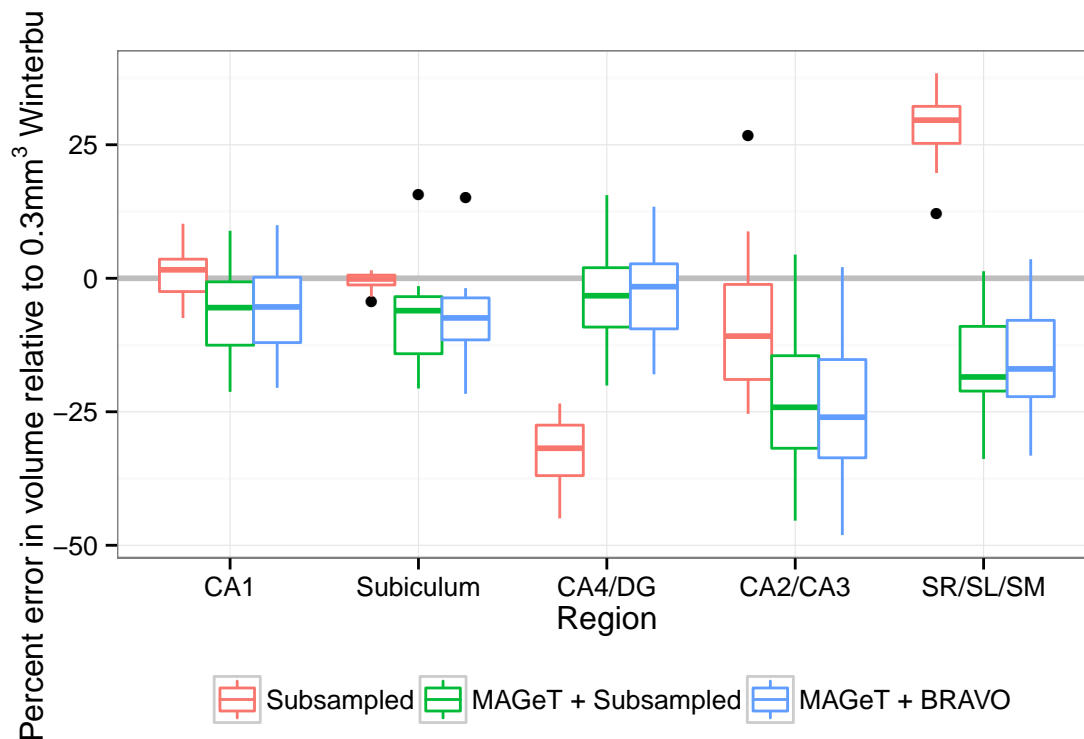Figure 5 shows a qualitative comparison of MAGeT-Brain subfield segmentation.

Figure 4: **Percent error in segmentation volume by hippocampus subregion.** Percent error is measured against the volumes of of the unmodified Winterburn atlas segmentations. **Subsampled** are volumes of the manual segmentations of the Winterburn atlases after resampling to $0.9mm^3$. **MAGeT + WA Subsampled** volumes are MAGeT-Brain segmentations of the Winterburn atlas images after resampling to $0.9mm^3$ voxels. **MAGeT + WA BRAVO** volumes are MAGeT-Brain segmentations of T1 BRAVO images ($0.9mm^3$ voxels) acquired separately of four of the five Winterburn atlas subjects.

Figure 5. Show MAGeT segmentations on BRAVO images. Also, show coronal/transverse slices and MNI coordinates
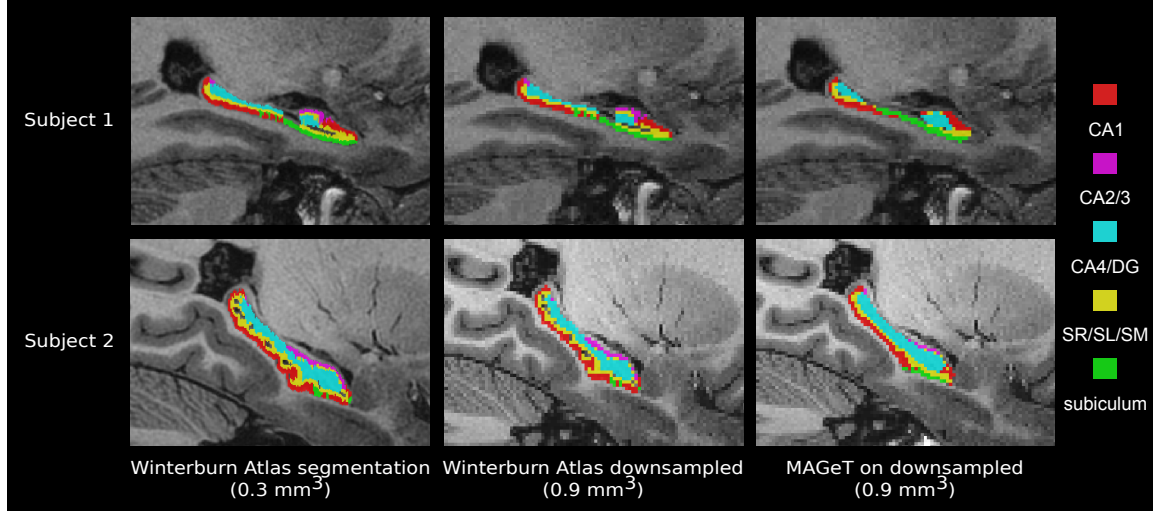
Figure 5: Sagittal slices from two subjects showing a comparison of the original Winterburn atlas subfield segmentations (at 0.3mm-isotropic voxel resolution), the subsampled Winterburn segmentations (at 0.9mm-isotropic voxel resolution), and the MAGeT-Brain labels on the subsampled atlas image.

3.3 Experiment 3 Results: Application to the segmentation of first episode schizophrenia patients

In this experiment MAGeT-Brain is applied to a dataset of images of first episode schizophrenia patients, using the Winterburn atlases and a template library of 21 subject images selected at random. Expert manual whole hippocampal segmentations are used as gold standards.

MAGeT-Brain produces hippocampus segmentation volumes that are highly correlated with manual segmentation volumes (Pearson $r = 0.877$, $t = 16.244$, $p < 0.001$; Figure 6).
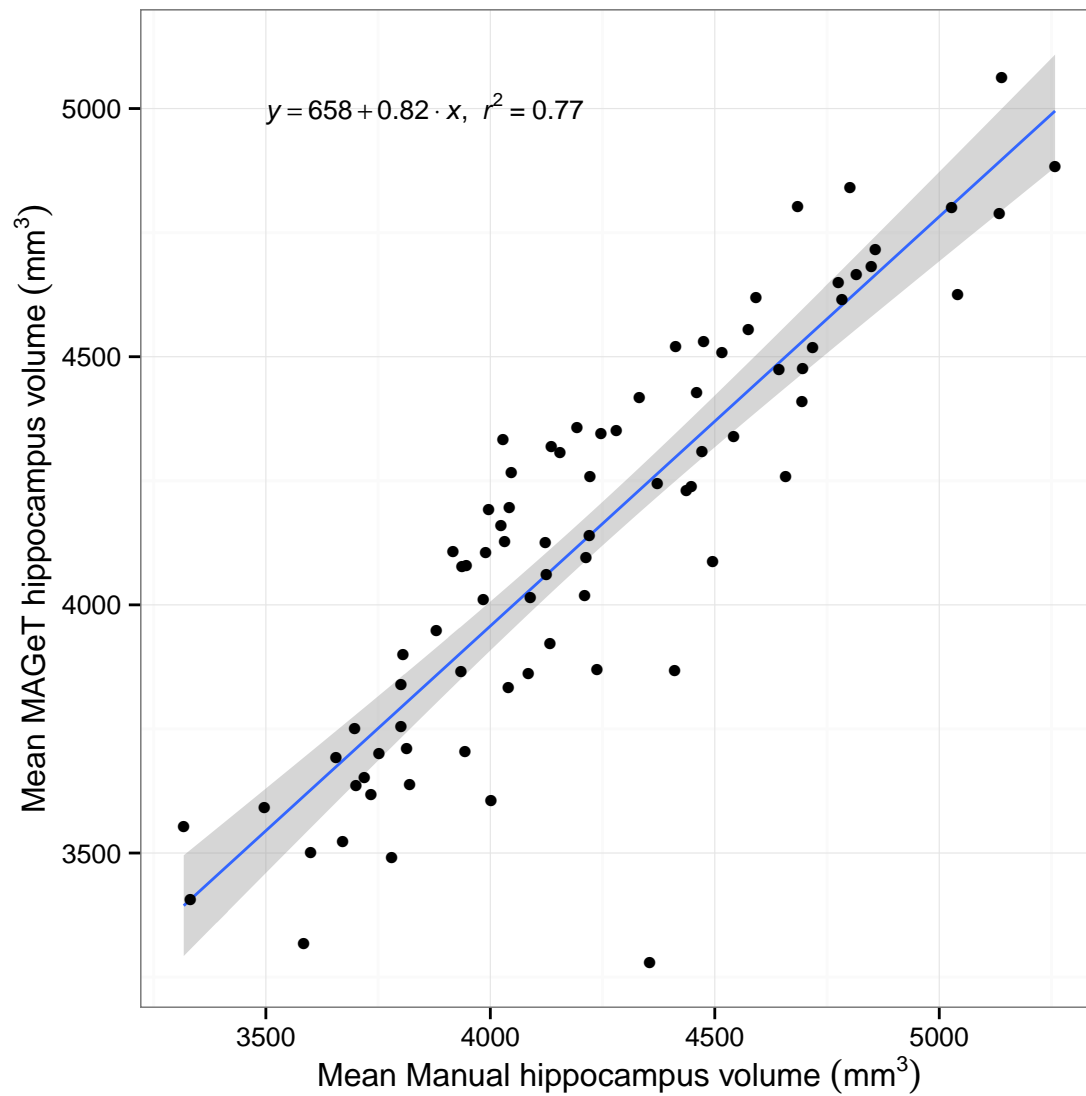
Figure 6: Subject mean hippocampal volumes by MAGeT-Brain vs. manually segmented volumes from the First Episode Patients with Schizophrenia dataset. A linear fit and standard error are shown.

Table 6: **Number of segmented images and quality control failures of ADNI1:Complete 1Yr 1.5T dataset by method.**

X	SNT	MAGeT	MAPER	FSL	FS
Images	1445	1909	636	1876	1530
Failures	–	34	–	27	304

3.4 Experiment 4 Results: Application to the segmentation of Alzheimer's disease patients

Based on the results from the ADNI1 Cross-Validation experiment, in this experiment MAGeT-Brain was configured with a template library of 21 randomly chosen subject images (7 from each disease class) and used majority vote label fusion. The entire ADNI1:Complete 1Yr 1.5T dataset was segmented by MAGeT-Brain and we now compare the resulting volumes with those obtained by manual segmentation (SNT), and other automated segmentation techniques (MAPER, FreeSurfer, and FSL). Table 6 shows the total count of segmentations available, including a count of those which have failed a quality control inspection. Quality control for FreeSurfer segmentations was provided along with the segmentations from ADNI Hartig et al. (2010). One of the authors (MP) performed visual quality inspection for MAGeT and FSL segmentations using similar quality control guidelines (if either hippocampus was under or over segmented by greater than 10mm in three or more slices then the segmentation did not pass).

Only those images which had quality segmentations from each method are included in the following analysis (a total of 361 images).

We find a close relationship in total bilateral hippocampal volume between all methods and manually segmented volumes (Figure 8). Volumes are correlated with Pearson $r > 0.78$ for all methods across disease categories. Within disease categories (Figure 7), MAGeT-Brain is consistently well correlated to manual volumes (Pearson $r > 0.85$), but appears to slightly over-estimate the volume of the AD hippocampus.

To investigate the level of agreement with manually segmented hippocampal volumes, we constructed Bland-Altman plots for each method (Figure 9). As Bland and Altman (1986) noted, high correlation amongst measures of the same quantity does not necessarily imply agreement (as correlation can be driven by a large range in true values, for instance). What is most striking in Figure 9 is that all methods show an obvious proportional bias: FreeSurfer and FSL markedly under-estimate smaller hippocampi and over-estimate large hippocampi, whereas MAPER and MAGeT-Brain more conservatively show the reverse bias. Additionally, all methods show a fixed bias, with FreeSurfer and FSL most dramatically over-estimating hippocampal volume by $2600mm^3$ and $2800mm^3$ on average, respectively, and MAPER and MAGeT-Brain within $250mm^3$ on average.

Figure 10 shows a qualitative comparison of MAGeT-Brain and manual (SNT) hippocampal segmentations for 10 randomly selected subjects in each disease category, and illustrates some of the common errors found during visual inspection. Mostly frequently, we find MAGeT-Brain improperly includes the vestigial hippocampal sulcus and, although not anatomically incorrect, MAGeT-Brain under-estimates the hippocampal body in comparison to the manual (SNT) segmentation.

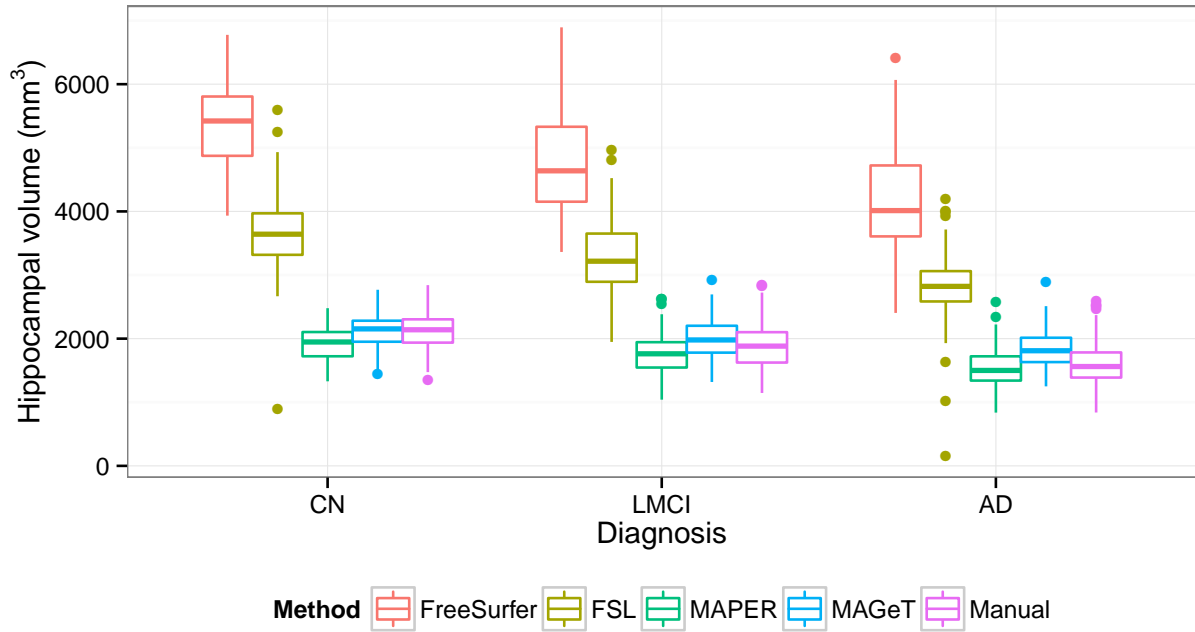Figure 10. Show coronal/transverse slices and MNI coordinates

Figure 7: Subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by FreeSurfer, FSL, MAPER, MAGeT-Brain , and manual raters (SNT) vs. disease category.
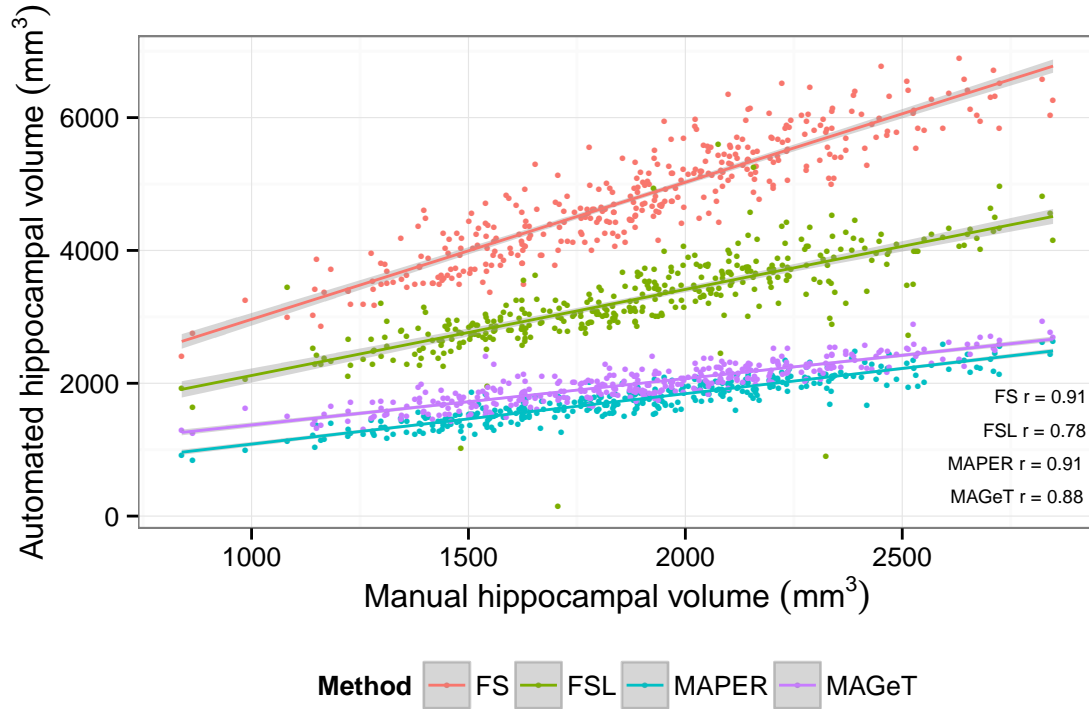

Figure 8: Subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by each of the four automated methods investigated (FreeSurfer (FS), FSL, MAPER, MAGeT-Brain) vs. manual rating (SNT). Linear fit lines and pearson correlations are shown for each method.
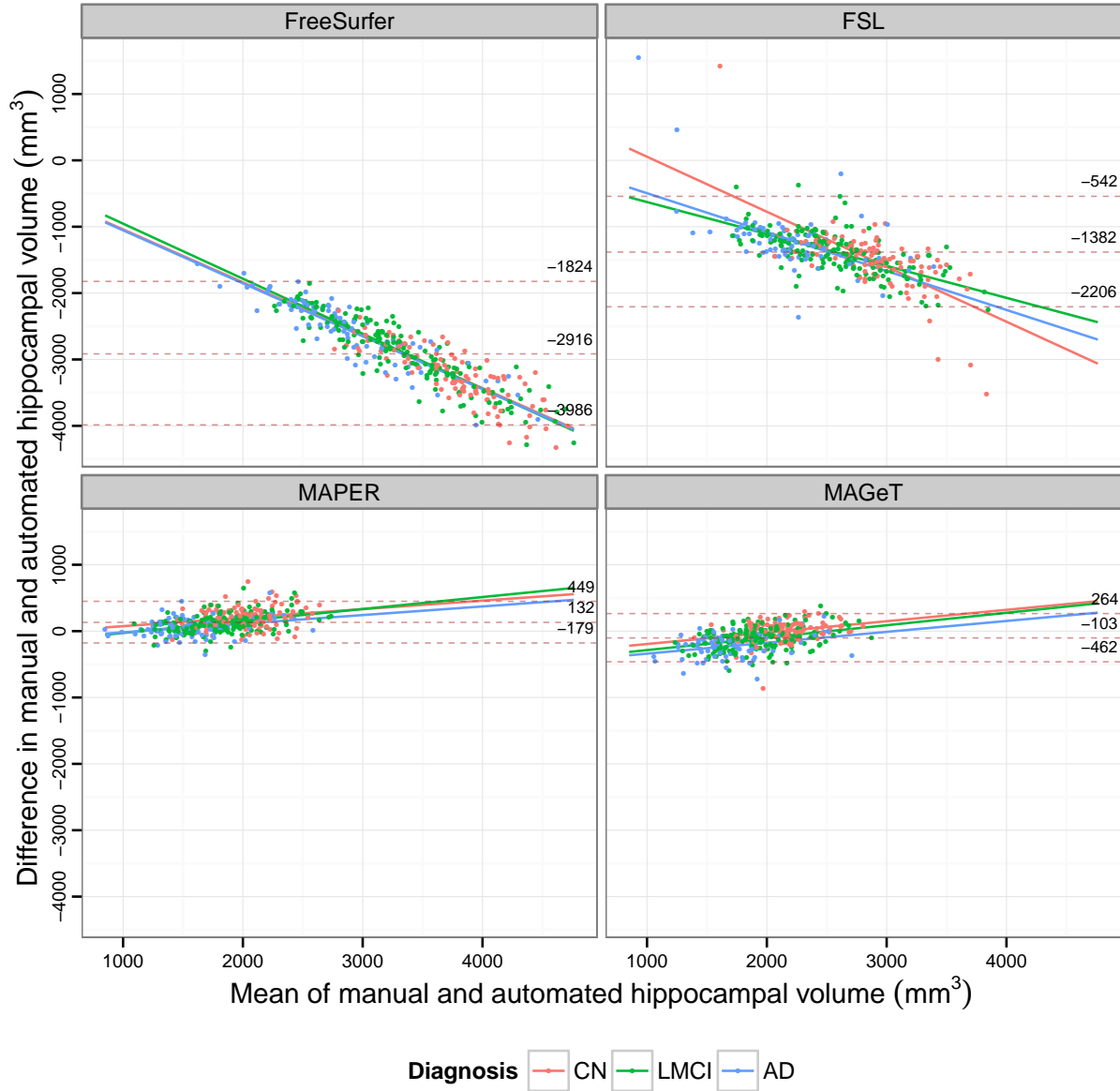
Figure 9: Bland-Altman plots comparing subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by manual raters (SNT) and each of the four automated methods investigated (FreeSurfer, FSL, MAPER, MAGeT-Brain). The overall mean difference in volume, and limits of agreement ($\pm 1.96 SD$) 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 rating, and vice versa.
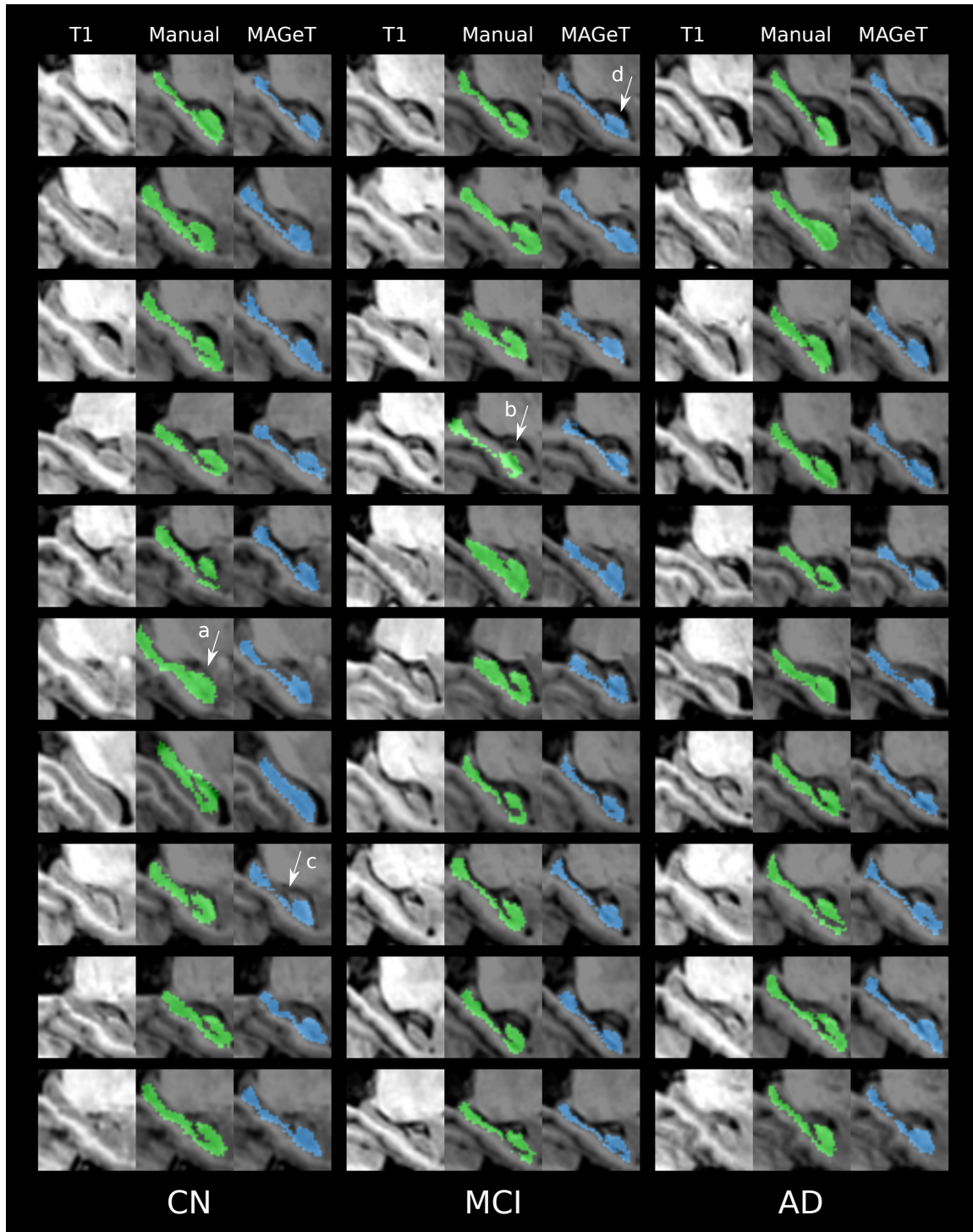
Figure 10: Manual and MAGeT segmentation results for 30 ADNI1 subjects (10 subjects randomly selected from each disease category in the subject pool used in Experiment 1). Sagittal slices are shown for subject unlabelled T1-weighted anatomical image, SNT manual label (green), and MAGeT-Brain label (blue). Noted are examples of common segmentation idiosyncrasies: *(a)* over-estimation of hippocampal head and *(b)* translated manual segmentation by SNT; *(c)* under-estimation of hippocampal body and *(d)* improper inclusion of the vestigial hippocampal sulcus by MAGeT-Brain .

Table 7: **Automated segmentation accuracy (overlap with manual labels) of the ADNI dataset.**
For each method, the number of manually labelled atlases used for training, the best Dice's overlap measure,
the disease classes measured, and the validation procedure are shown. Unless specified, validation datasets
are composed equally of subjects diagnosed with Alzheimer's Disease (AD), Mild Cognitive Impairment
(MCI), and Cognitively Normal (CN). LOOCV = Leave-one-out cross-validation. Some studies of auto-
mated segmentation of ADNI images are excluded because they do not not provide overlap measures for the
hippocampus Heckemann et al. (2011); Chupin et al. (2009).

Method	Atlases	DSC	Reference	Validation
MAGeT-Brain	9	0.838		10 rounds of Monte Carlo CV on a pool of 69 subjects
LEAP	30	0.848	Wolz et al. (2010)	Segmentation of 60 subjects
ACM (AdaBoost-based)	21	0.862	Morra et al. (2008)	LOOCV on atlases
Patch-based label fusion	16	0.883 (CN) 0.838 (AD)	Coupe et al. (2011)	LOOCV on atlases
Multi-atlas	30	0.885	Lötjönen et al. (2010)	Segmentation of 60 subjects
Multi-atlas + weighted fusion	20	0.898 (CN) 0.798 (left HC, MCI)	Wang et al. (2011)	10 rounds of Monte Carlo CV on 20 subjects, pool of 139 (CN/MCI)
Multi-atlas (MAPS)	55	0.890	Leung et al. (2010)	Segmentation of 30 subjects (10 AD, MCI, and CN)

4 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.
We have demonstrated that our methodology robustly provides accurate and consistent segmentations in
populations with different ageing and neuropsychiatric characteristics. Further, we have demonstrated that
algorithmic performance is not dependent on a single definition of the hippocampus but is effective with
differing hippocampal definitions (Winterburn et al., 2013; Pruessner et al., 2000; Hsu et al., 2002). Finally,
we also demonstrate that MAGeT-Brain provides accurate automatic identification of the hippocampal
subfields despite contrast and resolution limitations in standard T1-weighted image volumes.

Throughout the cross validation in Experiment 1 (10-fold Monte Carlo cross validation in the ADNI1:Complete
1Yr 1.5T dataset subsample) we find that two parameter choices improve segmentation accuracy: increasing
the number of atlases, and the number of templates. However, after setting the parameters to 5 atlases and
15 templates there are diminishing returns with respect to this improvement. We were somewhat surprised
to find that cross-correlation and normalized mutual information based weighted voting did little to improve
segmentation accuracy. This suggests that using an intermediate template library generated by accurate
nonlinear registration methods such as ANTS (Avants et al., 2008) sufficiently filters sources of error such
as nonlinear registration error and partial volume effects that arise through the use of nearest neighbour
resampling. From our previous work on the MAGeT-Brain algorithm we have shown that these increases in
error are not simply a smoothing or averaging effect (Mallar Chakravarty et al., 2012).

Although, the goal of this manuscript was to not 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 weights each segmentation based
upon its estimated performance level with respect to the other available candidate segmentations. Further,
the sensitivity and specificity parameters could also be tuned to potentially improve segmentation accuracy
and reliability. It is rather likely that using more sophisticated voting methods would have a positive effect
on the overall segmentation performance.

To this end, more work is required in order 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, we determined that
this pattern was also true in the analysis of 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, 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 (eg. Leung et al. (2010)).

Another concern is the moderate-to-small improvement observed in MAGeT-Brain in comparison to multi-atlas segmentation when using the same number of atlases. The actual benefit in using MAGeT-Brain is consistency of the labelling regardless of atlas or template choice. This is an important consideration that few have touched on previously. The 10-fold Monte Carlo leave-one-out cross-validation that we present in Experiment 1 is amongst one of the most stringent performed in the multi-atlas/segmentation literature. To the best of our knowledge, other groups using ADNI data for validation do at most a single round of leave-one-out-validation (Table 7), aside from (Wang et al., 2011) who perform a similar validation as us. The thoroughness of our validation suggests that our results are reflective of a true average over the choice of parameter settings and are independent of atlas or template choice (provided the input atlases are properly segmented).

> could we be seeing a ceiling effect... reaching the limit of what is achievable with a multi-atlas-based segmentation method

In comparison to other methodologies in the field, it is important to note that MAGeT-Brain performs quite well. Table 6 gives a survey of some of the most recent multi-atlas implementations with the reported kappa values used to validate the algorithm on the ADNI dataset (many of which use the same SNT labels for populating their template library and as gold standards for evaluation). While it is difficult to compare segmentation results across studies, gold-standards, evaluation metrics, and algorithms it is worth noting that the methods summarized in Table 6 require more atlases (between 16-55) than our MAGeT-Brain implementation with the Winterburn atlases (Winterburn et al., 2013). Amongst these methods only method 4 yields mean Kappas that are higher than the ones that are reported here.

There are some important differences between our method and these specific methods. Others have reported the difficulty with mis-registrations in candidate segmentation (i.e. segmentations generated that are then input in the voxel-voting procedure (Collins and Pruessner, 2010)). The work of Leung et al. (2010) tackles this problem by using an intensity threshold that is estimated heuristically at the time of segmentation (this work also reports some of the highest kappas for the segmentation of ADNI data). While this works for the ADNI dataset (which is partially homogenized with respect to image acquisition and pre-processing), it is unclear if this type of heuristic is applicable to other datasets. In all cases, these methods require more atlases than our implementation with the Winterburn atlases. Lötjönen et al. (2010) achieve highly accurate segmentation but correct their segmentations using classifications derived using an expectation maximization framework. In their initial work, Chupin et al. (2009) develop their probabilistic methodology using a cohort of 8 healthy controls and 15 epilepsy patients. In subsequent methods, they retuned their methodology using the ADNI sample with a hierarchical experimentation protocol that involved. These methods suggest that some post-processing of the final segmentations would improve accuracy of the segmentation. While that may be true, there is little consensus regarding how to achieve this.

To the best of our knowledge, no other groups have validated their work using multiple atlas segmentation protocols, different acquisitions, and disease populations in order to demonstrate the robustness of their technique. This is one of the clear strengths of this work. Furthermore, unlike some of the algorithms mentioned, our implementation does not require retuning 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 accuracy 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).

> time complexity seems suspicious

In comparison to the methods that we compared and were available through the ADNI database (FreeSurfer, MAPER) and the method that we initialized ourselves (FSL-FIRST) we find extremely variable performance of all methods. With the exception of FSL all methods correlate well with the SNT volumes provided in the ADNI database. However, FreeSurfer and FIRST provide radically different definitions of the size of the hippocampus in comparison to the other methods. Further, when estimating bias of these methods relative to SNT hippocampal volumes we see that large hippocampi are over estimated while small hippocampi are under estimated. By comparison, our method and MAPER are far more conservative suggesting that these methods may be better suited for estimating true-positives. While this comparison with these methods

using only volume, more work is needed to better understand the differences between our method and other methods.

Finally, we have also demonstrated that our algorithmic framework is appropriate for the segmentation of hippocampal subfields in standard T1-weighted data. This has started to become a burgeoning topic in the segmentation literature, although very few methods are available for the segmentation of 3T data (Yushkevich et al., 2009; Van Leemput et al., 2009). However, the FreeSurfer implementation of subfield segmentation. While recent work demonstrates that subfield segmentations can be used for classification of AD, MCI, and NC, there has been no explicit validation of the methodology based on accuracy or precision. Although the initial hippocampal subfield segmentations from the Yushkevich group have been demonstrated to work on the ADNI population, there has also been no validation of their work. In addition, their work requires some manual initialization to properly function. Our work demonstrates that we can reliably identify the CA1, subiculum, and CA4 dentate with only modest amounts of error. The fact that CA2/CA3 and molecular layers cannot be reasonably identified should not be surprising. These are extremely thin and spatially convoluted regions that originally required high-resolution MRI for identification. It is likely that they would be that extents of these regions are well below the resolution offered by standard T1-weighted images. In fact many manual segmentation methodologies do not attempt to parse these regions either (Wisse et al., 2012; Mueller and Weiner, 2009)

4.1 Conclusion

In conclusion, we have presented a flexible multi-atlas framework. It has considerable advantages over other methods as only a small set of atlases is required to initialize the algorithm. We demonstrate that our method works robustly over hippocampal definitions, different populations, and different acquisition types. Finally, we also demonstrate that using this method that accurate identification of the hippocampal subfields in also possible.

5 Supplementary Materials

5.1 ADNI Manual Labels

Semi-automated hippocampal volumetry was carried out using a commercially available high dimensional brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO), that has previously been validated and compared to manual tracing of the hippocampus (Hsu et al., 2002). Measurement of hippocampal volume is achieved first by placing manually 22 control points as local landmarks for the hippocampus on the individual brain MRI data: one landmark at the hippocampal head, one at the tail, and four per image (i.e., at the superior, inferior, medial and lateral boundaries) on five equally spaced images perpendicular to the long axis of the hippocampus. Second, fluid image transformation is used to match the individual brains to a template brain (Christensen et al., 1997). The pixels corresponding to the hippocampus are then labeled and counted to obtain volumes. This method of hippocampal voluming has a documented reliability of an intraclass coefficient better than .94 (Hsu et al., 2002).

References

P. Aljabar, R. a. Heckemann, a. Hammers, J. V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726–38, July 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.02.018. URL http://www.ncbi.nlm.nih.gov/pubmed/19245840.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, Feb. 2008. ISSN 1361-8423. doi: 10.1016/j.media.2007.06.004. URL http://dx.doi.org/10.1016/j.media.2007.06.004.

J. M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, pages 307–310, 1986. URL `http://www.sciencedirect.com/science/article/pii/S0140673686908378`.

M. M. Chakravarty, A. F. Sadikot, S. Mongia, G. Bertrand, and D. L. Collins. Towards a multi-modal atlas for neurosurgical planning. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9(Pt 2): 389–96, Jan. 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/17354796`.

M. M. Chakravarty, A. F. Sadikot, J. Germann, G. Bertrand, and D. L. Collins. Towards a validation of atlas warping techniques. *Medical image analysis*, 12(6):713–26, Dec. 2008. ISSN 1361-8423. doi: 10.1016/j.media.2008.04.003. URL `http://dx.doi.org/10.1016/j.media.2008.04.003`.

M. M. Chakravarty, A. F. Sadikot, J. Germann, P. Hellier, G. Bertrand, and D. L. Collins. Comparison of piece-wise linear, linear, and nonlinear atlas-to-patient warping techniques: analysis of the labeling of subcortical nuclei for functional neurosurgical applications. *Human brain mapping*, 30(11):3574–95, Nov. 2009. ISSN 1097-0193. doi: 10.1002/hbm.20780. URL `http://www.ncbi.nlm.nih.gov/pubmed/19387981`.

G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformation of brain anatomy. *IEEE transactions on medical imaging*, 16(6):864–77, Dec. 1997. ISSN 0278-0062. doi: 10.1109/42.650882. URL `http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=650882`.

M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and O. Colliot. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–87, June 2009. ISSN 1098-1063. doi: 10.1002/hipo.20626. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2837195\&tool=pmcentrez\&rendertype=abstract`.

D. L. Collins and J. C. Pruessner. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52 (4):1355–66, Oct. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.04.193. URL `http://www.ncbi.nlm.nih.gov/pubmed/20441794`.

D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192–205. ISSN 0363-8715. URL `http://www.ncbi.nlm.nih.gov/pubmed/8126267`.

D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, Oct. 1995. ISSN 10659471. doi: 10.1002/hbm.460030304. URL `http://doi.wiley.com/10.1002/hbm.460030304`.

P. Coupe, V. Fonov, S. Eskildsen, J. Manjón, D. Arnold, and L. Collins. Influence of the training library composition on a patch-based label fusion method: Application to hippocampus segmentation on the ADNI dataset. *Alzheimer's & Dementia*, 7(4):S316, July 2011. ISSN 15525260. doi: 10.1016/j.jalz.2011.05.918. URL `http://linkinghub.elsevier.com/retrieve/pii/S1552526011010612`.

P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, and D. L. Collins. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. *NeuroImage*, 59(4):3736–47, Feb. 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.10.080. URL `http://www.ncbi.nlm.nih.gov/pubmed/22094645`.

J. G. Csernansky, S. Joshi, L. Wang, J. W. Haller, M. Gado, J. P. Miller, U. Grenander, and M. I. Miller. Hippocampal morphometry in schizophrenia by high dimensional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 95(19):11406–11411, 1998. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21655\&tool=pmcentrez\&rendertype=abstract`.

T. den Heijer, F. V. der Lijn, M. W. Vernooij, M. de Groot, P. J. Koudstaal, a. V. der Lugt, G. P. Krestin, a. Hofman, W. J. Niessen, and M. M. B. Breteler. Structural and diffusion MRI measures of the hippocampus and memory performance. *NeuroImage*, 63(4):1782–9, Dec. 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.08.067. URL `http://www.ncbi.nlm.nih.gov/pubmed/22960084`.

B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–55, Jan. 2002. ISSN 0896-6273. URL `http://www.ncbi.nlm.nih.gov/pubmed/11832223`.

E. Geuze, E. Vermetten, and J. D. Bremner. MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders. *Molecular Psychiatry*, 10(2):160, Sept. 2004. doi: 10.1038/sj.mp.4001579. URL `http://www.nature.com.myaccess.library.utoronto.ca/mp/journal/v10/n2/full/4001579a.htmlhttp://www.nature.com.myaccess.library.utoronto.ca/mp/journal/v10/n2/pdf/4001579a.pdf`.

J. W. Haller, A. Banerjee, G. E. Christensen, M. Gado, S. Joshi, M. I. Miller, Y. Sheline, M. W. Vannier, and J. G. Csernansky. Three-dimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas. *Radiology*, 202(2):504–510, 1997. URL `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve\&db=PubMed\&dopt=Citation\&list\_uids=9015081`.

M. Hartig, D. Truran-sacrey, S. Raptentsetsang, N. Schuff, and M. Weiner. USCF FreeSurfer Overview and QC Ratings. 2010. URL `http://adni.loni.ucla.edu/wp-content/uploads/2010/12/ADNI\_UCSF\_Freesurfer-Overview-and-QC.pdf`.

R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 46(3):726–38, July 2006a. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.02.018. URL `http://www.ncbi.nlm.nih.gov/pubmed/19245840`.

R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–26, Oct. 2006b. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.05.061. URL `http://dx.doi.org/10.1016/j.neuroimage.2006.05.061http://www.ncbi.nlm.nih.gov/pubmed/16860573`.

R. A. Heckemann, S. Keihaninejad, P. Aljabar, K. R. Gray, C. Nielsen, D. Rueckert, J. V. Hajnal, and A. Hammers. Automatic morphometry in Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 56(4):2024–37, July 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.03.014. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3153069\&tool=pmcentrez\&rendertype=abstract`.

Y.-Y. Hsu, N. Schuff, A.-T. Du, K. Mark, X. Zhu, D. Hardin, and M. W. Weiner. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. *Journal of magnetic resonance imaging : JMRI*, 16(3):305–10, Sept. 2002. ISSN 1053-1807. doi: 10.1002/jmri.10163. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1851676\&tool=pmcentrez\&rendertype=abstract`.

C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. G. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of magnetic resonance imaging : JMRI*, 27(4):685–91, Apr. 2008. ISSN 1053-1807. doi: 10.1002/jmri.21049. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2544629\&tool=pmcentrez\&rendertype=abstract`.

C. R. Jack, F. Barkhof, M. A. Bernstein, M. Cantillon, P. E. Cole, C. Decarli, B. Dubois, S. Duchesne, N. C. Fox, G. B. Frisoni, H. Hampel, D. L. G. Hill, K. Johnson, J.-F. Mangin, P. Scheltens, A. J. Schwarz, R. Sperling, J. Suhy, P. M. Thompson, M. Weiner, and N. L. Foster. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 7(4):474–485.e4, July 2011. ISSN 1552-5279. doi: 10.1016/j.jalz.2011.04.007. URL `http://www.ncbi.nlm.nih.gov/pubmed/21784356`.

M. S. Karnik-Henry, L. Wang, D. M. Barch, M. P. Harms, C. Campanella, and J. G. Csernansky. Medial temporal lobe structure and cognition in individuals with schizophrenia and in their non-psychotic siblings. *Schizophrenia research*, 138(2-3):128–35, July 2012. ISSN 1573-2509. doi: 10.1016/j.schres.2012.03.015. URL `http://www.ncbi.nlm.nih.gov/pubmed/22542243`.

K. K. Leung, J. Barnes, G. R. Ridgway, J. W. Bartlett, M. J. Clarkson, K. Macdonald, N. Schuff, N. C. Fox, and S. Ourselin. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, 51(4):1345–59, July 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.03.018. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873209\&tool=pmcentrez\&rendertype=abstract`.

J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352–65, Mar. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.10.026. URL `http://dx.doi.org/10.1016/j.neuroimage.2009.10.026`.

A. Malla, R. Norman, T. McLean, D. Scholten, and L. Townsend. A Canadian programme for early intervention in non-affective psychotic disorders. *The Australian and New Zealand journal of psychiatry*, 37 (4):407–13, Aug. 2003. ISSN 0004-8674. URL `http://www.ncbi.nlm.nih.gov/pubmed/12873324`.

M. Mallar Chakravarty, P. Steadman, M. C. van Eede, R. D. Calcott, V. Gu, P. Shaw, A. Raznahan, D. Louis Collins, and J. P. Lerch. Performing label-fusion-based segmentation using multiple automatically generated templates. *Human brain mapping*, 00(September 2011):1–20, May 2012. ISSN 1097-0193. doi: 10.1002/hbm.22092. URL `http://www.ncbi.nlm.nih.gov/pubmed/22611030`.

J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma, H. Hulshoff Pol, T. Cannon, R. Kawashima, and B. Mazoyer. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association : JAMIA*, 8(5):401–30. ISSN 1067-5027. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=131040\&tool=pmcentrez\&rendertype=abstract`.

J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1412):1293–322, Aug. 2001. ISSN 0962-8436. doi: 10.1098/rstb.2001.0915. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1088516\&tool=pmcentrez\&rendertype=abstract`.

J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). *NeuroImage*, 2(2):89–101, June 1995. ISSN 1053-8119. URL `http://www.ncbi.nlm.nih.gov/pubmed/9343592`.

J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, A. W. Toga, C. R. Jack, M. W. Weiner, and P. M. Thompson. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment,

and elderly controls. *NeuroImage*, 43(1):59–68, Oct. 2008. ISSN 1095-9572. doi: 10.1016/j.neuroimage. 2008.07.003. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2624575\&tool= pmcentrez\&rendertype=abstract`.

S. G. Mueller and M. W. Weiner. Selective effect of age, Apo e4, and Alzheimer's disease on hippocampal subfields. *Hippocampus*, 19(6):558–64, June 2009. ISSN 1098-1063. doi: 10.1002/ hipo.20614. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2802542\&tool= pmcentrez\&rendertype=abstract`.

K. L. Narr, P. M. Thompson, P. Szeszko, D. Robinson, S. Jang, R. P. Woods, S. Kim, K. M. Hayashi, D. Asunction, A. W. Toga, and R. M. Bilder. Regional specificity of hippocampal volume reductions in first-episode schizophrenia. *NeuroImage*, 21(4):1563–75, Apr. 2004. ISSN 1053-8119. doi: 10.1016/j. neuroimage.2003.11.011. URL `http://www.ncbi.nlm.nih.gov/pubmed/15050580`.

Z. Pausova, T. Paus, M. Abrahamowicz, J. Almerigi, N. Arbour, M. Bernard, D. Gaudet, P. Hanzalek, P. Hamet, A. C. Evans, M. Kramer, L. Laberge, S. M. Leal, G. Leonard, J. Lerner, R. M. Lerner, J. Mathieu, M. Perron, B. Pike, A. Pitiot, L. Richer, J. R. Séguin, C. Syme, R. Toro, R. E. Tremblay, S. Veillette, and K. Watkins. Genes, maternal smoking, and the offspring brain and body during adolescence: design of the Saguenay Youth Study. *Human brain mapping*, 28(6):502–18, June 2007. ISSN 1065-9471. doi: 10.1002/hbm.20402. URL `http://www.ncbi.nlm.nih.gov/pubmed/17469173`.

J. Poppenk and M. Moscovitch. A Hippocampal Marker of Recollection Memory Ability among Healthy Young Adults: Contributions of Posterior and Anterior Segments. *Neuron*, 72 (6):931–937, Dec. 2011. ISSN 0896-6273. doi: 10.1016/j.neuron.2011.10.014. URL `http: //www.sciencedirect.com/science/article/pii/S089662731100924Xhttp://pdn.sciencedirect. com.myaccess.library.utoronto.ca/science?_ob=MiamiImageURL\&_cid=272195\&_user= 994540\&_pii=S089662731100924X\&_check=y\&_origin=article\&_zone=toolbar\&\ _coverDate=22-Dec-2011\&view=c\&originContentFamily=serial\&wchp=dGLbVlV-zSkzk\ &md5=e75d94a1de9d5c31e146f910b38468da/1-s2.0-S089662731100924X-main.pdfhttp://www. sciencedirect.com.myaccess.library.utoronto.ca/science/article/pii/S089662731100924X`.

J. C. Pruessner, L. M. Li, W. Serles, M. Pruessner, D. L. Collins, N. Kabani, S. Lupien, and A. C. Evans. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. *Cerebral cortex (New York, N.Y. : 1991)*, 10 (4):433–42, Apr. 2000. ISSN 1047-3211. URL `http://www.ncbi.nlm.nih.gov/pubmed/10769253`.

S. Robbins, A. C. Evans, D. L. Collins, and S. Whitesides. Tuning and comparing spatial normalization methods. *Medical image analysis*, 8(3):311–23, Sept. 2004. ISSN 1361-8415. doi: 10.1016/j.media.2004. 06.009. URL `http://www.ncbi.nlm.nih.gov/pubmed/15450225`.

N. Robitaille and S. Duchesne. Label fusion strategy selection. *International journal of biomedical imaging*, 2012:431095, Jan. 2012. ISSN 1687-4196. doi: 10.1155/2012/431095. URL `http://www.pubmedcentral. nih.gov/articlerender.fcgi?artid=3296312\&tool=pmcentrez\&rendertype=abstract`.

M. R. Sabuncu, R. S. Desikan, J. Sepulcre, B. T. T. Yeo, H. Liu, N. J. Schmansky, M. Reuter, M. W. Weiner, R. L. Buckner, R. a. Sperling, and B. Fischl. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of neurology*, 68(8):1040–8, Aug. 2011. ISSN 1538-3687. doi: 10.1001/archneurol.2011.167. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid= 3248949\&tool=pmcentrez\&rendertype=abstract`.

W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. 1957. *The Journal of neuropsychiatry and clinical neurosciences*, 12(1):103–113, 2000. URL `http://www.ncbi.nlm.nih. gov/pubmed/10678523`.

J. Shao. Linear Model Selection by Cross-validation. *Journal of the American Statistical Association*, 88 (422):486–494, June 1993. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476299. URL `http://www. tandfonline.com/doi/abs/10.1080/01621459.1993.10476299`.

J. G. Sled, a. P. Zijdenbos, and a. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging*, 17(1):87–97, Feb. 1998. ISSN 0278-0062. doi: 10.1109/42.668698. URL `http://www.ncbi.nlm.nih.gov/pubmed/9617910`.

C. Studholme, D. Hill, and D. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86, Jan. 1999. ISSN 00313203. doi: 10.1016/S0031-3203(98)00091-0. URL `http://dx.doi.org/10.1016/S0031-3203(98)00091-0`.

C. Studholme, E. Novotny, I. G. Zubal, and J. S. Duncan. Estimating tissue deformation between functional images induced by intracranial electrode implantation using anatomical MRI. *NeuroImage*, 13(4):561–76, Apr. 2001. ISSN 1053-8119. doi: 10.1006/nimg.2000.0692. URL `http://www.ncbi.nlm.nih.gov/pubmed/11305886`.

K. Van Leemput, A. Bakkour, T. Benner, G. Wiggins, L. L. Wald, J. Augustinack, B. C. Dickerson, P. Golland, and B. Fischl. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. *Hippocampus*, 19(6):549–57, June 2009. ISSN 1098-1063. doi: 10.1002/hipo.20615. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2739884\&tool=pmcentrez\&rendertype=abstract`.

H. Wang, J. W. Suh, J. Pluta, M. Altinay, and P. Yushkevich. Optimal weights for multi-atlas label fusion. *Information processing in medical imaging : proceedings of the ... conference*, 22:73–84, Jan. 2011. ISSN 1011-2499. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3226736\&tool=pmcentrez\&rendertype=abstract`.

S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–21, July 2004. ISSN 0278-0062. doi: 10.1109/TMI.2004.828354. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1283110\&tool=pmcentrez\&rendertype=abstract`.

J. L. Winterburn, J. C. Pruessner, S. Chavez, M. Schira, N. J. Lobaugh, A. N. Voineskos, and M. M. Chakravarty. A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. *NeuroImage*, 74:254–65, Feb. 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.02.003. URL `http://www.ncbi.nlm.nih.gov/pubmed/23415948`.

L. E. M. Wisse, L. Gerritsen, J. J. M. Zwanenburg, H. J. Kuijf, P. R. Luijten, G. J. Biessels, and M. I. Geerlings. Subfields of the hippocampal formation at 7 T MRI: in vivo volumetric assessment. *NeuroImage*, 61(4):1043–9, July 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.03.023. URL `http://www.ncbi.nlm.nih.gov/pubmed/22440643`.

R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert. LEAP: learning embeddings for atlas propagation. *NeuroImage*, 49(2):1316–25, Jan. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2009.09.069. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3068618\&tool=pmcentrez\&rendertype=abstract`.

B. T. Wyman, D. J. Harvey, K. Crawford, M. A. Bernstein, O. Carmichael, P. E. Cole, P. K. Crane, C. Decarli, N. C. Fox, J. L. Gunter, D. Hill, R. J. Killiany, C. Pachai, A. J. Schwarz, N. Schuff, M. L. Senjem, J. Suhy, P. M. Thompson, M. Weiner, and C. R. Jack. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, Oct. 2012. ISSN 1552-5279. doi: 10.1016/j.jalz.2012.06.004. URL `http://www.ncbi.nlm.nih.gov/pubmed/23110865`.

J. Yelnik, E. Bardinet, D. Dormont, G. Malandain, S. Ourselin, D. Tandé, C. Karachi, N. Ayache, P. Cornu, and Y. Agid. A three-dimensional, histological and deformable atlas of the human basal ganglia. I. Atlas construction based on immunohistochemical and MRI data. *NeuroImage*, 34(2):618–38, Jan. 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.09.026. URL `http://www.ncbi.nlm.nih.gov/pubmed/17110133`.

P. A. Yushkevich, B. B. Avants, J. Pluta, S. Das, D. Minkoff, D. Mechanic-Hamilton, S. Glynn, S. Pickup, W. Liu, J. C. Gee, M. Grossman, and J. A. Detre. A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. *NeuroImage*, 44(2):385–98, Jan. 2009. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.08.042. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2650508\&tool=pmcentrez\&rendertype=abstract`.