Bootstrapping Multi-atlas Segmentation Using Multiple Automatically Generated Templates for the Segmentation of the Whole Hippocampus and Subfields

Jon Pipitone[1], Min Tae M. Park[1], Julie Winterburn[1], Tristram A. Lett[1,9], Jason P. Lerch[2,3], Jens C. Pruessner[4], Martin Lepage[4,5], Aristotle N. Voineskos[1,6,9], M. Mallar Chakravarty[1,6,7,8] and the 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 Laboratory, Hospital for Sick Children, Toronto, ON, Canada*
[3] *Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*
[4] *Douglas Mental Health University Institute, Verdun, QC, Canada*
[5] *Department of Psychiatry, McGill University, Montreal, QC, Canada*
[6] *Department of Psychiatry, University of Toronto, Toronto, ON, Canada*
[7] *Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada*
[8] *Rotman Research Institute, Baycrest, Toronto, ON, Canada*
[9] *Institute of Medical Science, University of Toronto, Toronto, ON, Canada*

Abstract

Introduction: Advances in image segmentation of magnetic resonance images (MRI) have demonstrated that multi-atlas approaches improve segmentation accuracy and precision over regular atlas-based approaches. These approaches often rely on a large number of such manually segmented atlases (e.g. 30-80) that take significant time and expertise to produce. We present an algorithm, MAGeT-Brain (**M**ultiple **A**utomatically **G**enerated **T**emplates), for the automatic segmentation of the hippocampus that minimizes the number of atlases needed while still achieving similar accuracy to multi-atlas approaches. Thus, our method acts as an accurate multi-atlas approach when using special, 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 non-linear image registration. The resultant segmentations are then propagated to each target image and fused using a label fusion method. To explore the effect of atlases and templates library size, registation and label fusion method, we conduct a 10-fold Monte Carlo cross-validation of MAGeT-Brain of whole hippocampal segmentation on 60 ADNI subjects, manually segmented using the Pruessner-protocol, over a range of parameter settings.

Using the best settings found, we then conduct a leave-one-out cross-validation (LOOCV) of hippocampal subfield segmentation using five high-resolution manually segmented atlases. Two final experiments assess MAGeT-Brain when applied to first episode psychosis and Alzheimer's disease populations, and MAGeT-Brain segmentations are compared with existing automated methods (FSL FIRST, FreeSurfer, MAPER) and biases are explored.

Results: Using 9 atlases and 19 template images, MAGeT-Brain achieves a mean Dice's Similarity Coefficient (DSC) of 0.869 (to Pruessner-protocol manual segmentations) over 10-folds of Monte Carlo significantly lower variability in DSC than multi-atlas segmentation. In a LOOCV, MAGeT-Brain reproduces the volumes of the cornu ammonis (CA) 1; CA4/dentate gyrus (DG); and strata radiatum (SR), strata lacunosum (SL), and strata moleculare (SM) hippocampal subfields with a percent error in volume that is at or lower than that produced by image resampling. MAGeT-Brain produces hippocampal volumes in a first episode psychosis patient population that are highly correlated with expert manual segmentation volumes (Pearson $r = 0.877, t = 16.244, p < 0.001$). Compared to FSL FIRST and FreeSurfer, MAGeT-Brain shows much smaller fixed volume bias (within $250mm^3$ on average) to semi-automated (SNT) segmentations available from ADNI, as well as a conservative, rather than exaggerated, proportional volume bias.

Conclusion: We demonstrate that MAGeT-Brain produces accurate hippocampal segmentations using only 5 atlases over different hippocampal definitions, disease populations, and acquisition types, as well as showing that accurate identification of the hippocampal subfields is possible.

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 brain structure situated in the medial temporal lobe, and has long been associated with learning and memory (den Heijer et al., 2012; Scoville and Milner, 2000). The hippocampus is of interest to clinical neuroscientists because it is implicated in many forms of brain dysfunction, including Alzheimer's disease (Sabuncu et al., 2011) and schizophrenia (Narr et al., 2004; Karnik-Henry et al., 2012). In neuroimaging studies, structural magnetic resonance images (MRI) are often used for the volumetric assessment of the hippocampus. As such, accurate segmentation of the hippocampus and its subfields in MRI is a necessary first step to better understand the inter-individual variability of subject neuroanatomy.

The gold standard for neuroanatomical segmentation is manual delineation by an expert human rater. However, with the availability of increasingly large MRI datasets, the time and expertise required for manual segmentation becomes prohibitive (Mazziotta et al., 1995, 2001; Mazziotta et al.; Pausova et al., 2007). This effort is complicated by the fact that there is significant variation between segmentation protocols with respect to specific anatomical boundaries of the hippocampus (Geuze et al., 2004) and this has led to efforts to create an unified hippocampal segmentation protocol (Jack et al., 2011; Boccardi et al., 2013b,a). In

addition, there is controversy over the appropriate manual segmentation protocol to use in a particular imaging study (Nestor et al., 2012). Thus, a segmentation algorithm that can easily adapt to different manual segmentation definitions would be of significant benefit to the neuroimaging community.

Automated segmentation techniques that are reliable, objective, and reproducible can be considered complementary 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 warp 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 due to error in the estimated nonlinear 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 used 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 the 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 as 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. Additionally, several methods have been explored for label fusion, including the STAPLE algorithm (Simultaneous Truth And Performance Level Estimation; 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 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 into 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 (Chakravarty et al., 2013). The novelty of this manuscript is the extension of our multi-atlas methodology to the hippocampus using more than a single input atlas, while simultaneously limiting the number of inputs used during segmentation, and demonstrating that accurate identification of the hippocampal subfields is indeed possible using this methodology.

Of particular relevance to the present work is the LEAP algorithm (Learning Embeddings for Atlas Propagation; Wolz et al. (2010)) because of its focus on performing multi-atlas segmentation with a limited number of input atlases. The LEAP algorithm is a clever modification to 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 ADNI-1 baseline dataset, achieving a mean Dice score of 0.85 against Gold Standard segmentations.

Also of interest to this manuscript are methods that attempt to define hippocampal subfields using standard T1-weighted data. To the best of our knowledge, there are only two automated segmentation algorithms that attempt this problem. 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, it demonstrates that the applicability of hippocampal subfield segmentation using data from 10 subjects. In the second method, Yushkevich et al. (2009) hippocampal subfields were labelled on highly-anistropic ($0.380mm$ in-plane resolution, $0.8mm$ coronal slice thickness) T2-weighted MR images acquired from post-mortem medial temporal lobe samples. Using nonlinear registration guided by manually derived hippocampus masks and specific landmarks, the authors demonstrate accurate parcellation of hippocampal subfields in unlabelled MRI volumes.

In this paper we have describe a thorough validation of the MAGeT-Brain algorithm for segmentation of the hippocampus and its subfields. First, we address the feasibility of whole hippocampus segmentation with a limited number of input atlases (Chakravarty et al., 2013) by performing a multi-fold validation experiment using a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset manually segmented using the Pruessner-protocol, over a range of atlas and template library sizes, registration and label fusion methods. We then perform a leave-one-out validation to determine if hippocampal subfields can be accurately identified, using the best parameters discovered in the first experiment. To ensure that we have not overfit our parameters to the aging or neurodegenerative brain, we also apply MAGeT-Brain to a dataset of individuals suffering from first episode psychosis. Finally, we validate our algorithm using all of the data available in the ADNI1:Complete 1Yr 1.5T sample and compare our segmentations to other popular segmentation algorithms.

2   Methods

2.1   The MAGeT-Brain Algorithm

In this paper, 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 (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 label by label fusion (Heckemann et al., 2006a, 2011). The basic multi-atlas segmentation method is described in detail in other publications (Collins and Pruessner, 2010; Heckemann et al., 2011; Aljabar et al., 2009). 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.

MAGeT-Brain (**M**ultiple **A**utomatically **G**enerated **T**emplates) bootstraps the creation of a large template library given a limited input atlas library, and then uses the template library in basic multi-atlassegmentation. 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 automatically labelled by each of the atlases via label propagation. Basic multi-atlassegmentation 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 atlas $\times$ # of templates).

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

2.2   Experiments

The following section describes for experiments conducted to assess the segmentation quality of the MAGeT-Brain algorithm. The first two experiments assess the validity of MAGeT-Brain using cross-validation designs. 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 assesses hippocampal subfield segmentation quality in a leave-one-out cross-validation design. The last two experiments assess the validity of the MAGeT-Brain 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 methods and assessed segmentation bias.
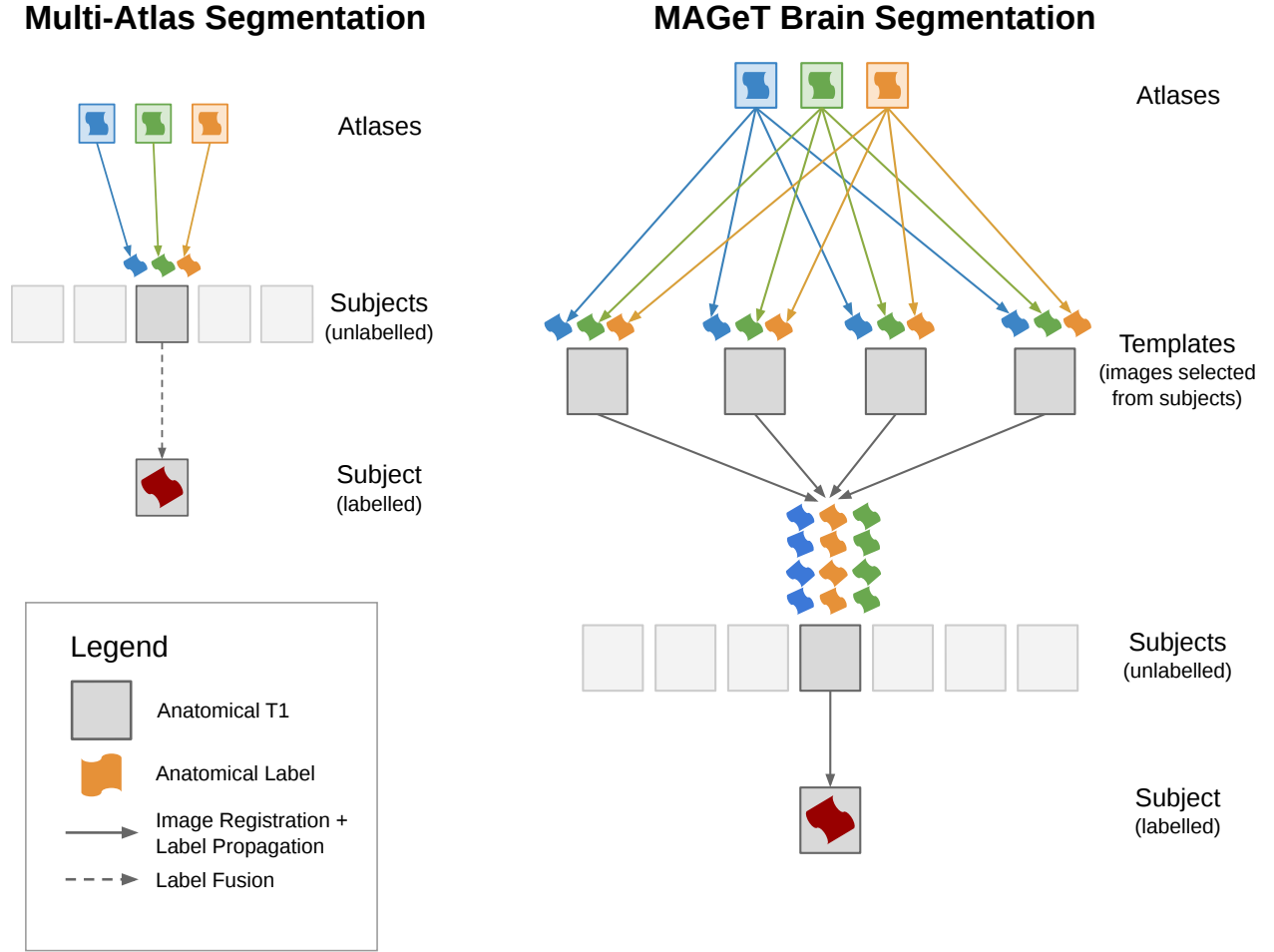
Figure 1: A schematic illustration of basic multi-atlas segmentation and MAGeT-Brain segmentation. In multi-atlas segmentation, manual labels from atlas images are warped (propagated) into subject space by applying the transformations estimated from non-linear image registration. The resulting candidate 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 the labels for a subject are then fused into a final segmentation.

2.2.1   Experiment 1: Whole Hippocampus Cross-Validation

Monte Carlo Cross-Validation (MCCV) (Shao, 1993) was performed using images from the ADNI dataset (Hsu et al., 2002), and manual whole hippocampus segmentations following the protocol laid out in (Pruessner et al., 2000).

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: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.ucla.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), other biological markers, and clinical and neuropsychological assessment 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.

60 1.5T images were arbitrarily selected from the baseline scans in the *ADNI1:Complete 1Yr 1.5T* standardized dataset. 20 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.ucla.edu) between March 2012 and August 2012. The image dataset used was the "ADNI1:Complete 1Yr 1.5T" standardized dataset available from ADNI [1] (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 = 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 dimensions

---

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

Table 1: **ADNI-1 cross-validation subset demographics.** CN - Cognitively Normal. LMCI - Late-onset Mild Cognitive Impairment. AD - Alzheimer's Disease. Hisp - Hispanic. CDR-SB - Clinical Dementia Rating-Sum of Boxes. ADAS - Alzheimer's Disease Assessment Scale. MMSE - Mini-Mental State Examination.

	CN $N = 20$	LMCI $N = 20$	AD $N = 20$	Combined $N = 60$
Age at baseline ₛYears	72.2 75.5 80.3	70.9 75.6 80.4	69.4 74.9 80.1	70.9 75.2 80.2
Sex : Female	50% (10)	50% (10)	50% (10)	50% (30)
Education	14.0 16.0 18.0	13.8 16.0 16.5	12.0 15.5 18.0	13.0 16.0 18.0
CDR-SB	0.00 0.00 0.00	1.00 2.00 2.50	3.50 4.00 5.00	0.00 1.75 3.62
ADAS 13	6.00 7.67 11.00	14.92 20.50 25.75	24.33 27.00 32.09	9.50 18.84 26.25
MMSE	28.8 29.5 30.0	26.0 27.5 28.2	22.8 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.

of $1.25mm \times 1.25mm \times 1.2mm$.

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 normalized mutual information weighted majority vote, described below). In each validation round, both a MAGeT-Brain and multi-atlas segmentation is produced. A total of $10 \times 60 \times 9 \times 20 \times 2 \times 3 = 6.48 \times 10^5$ validation rounds were conducted and resulting segmentations analysed.

Before registration, all images underwent preprocessing with the N3 algorithm (Sled et al., 1998) to minimize intensity nonuniformity. In this experiment we compared two non-linear 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 maximizes the correlation between blurred MR 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 larger deforma-

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

234 tions are then input to subsequent steps where the fit is refined by estimating smaller deformations on data
235 blurred with a Gaussian kernel with a smaller full width at half maximum (FWHM). The final transfor-
236 mation is a set of local translations defined on a bed of equally spaced nodes that were estimated through
237 the optimization of the correlation coefficient. For the purposes of this work we used the regularization
238 parameters optimized in Robbins et al. (2004), displayed in Table 2.

239 **Automatic Normalization Tools (ANTS)**   ANTS is a diffeomorphic registration algorithm which
240 provides great flexibility over the choice of transformation model, objective function, and the consistency of
241 the final transformation (Avants et al., 2008). The transformation is estimated in a hierarchical fashion where
242 the MRI data is subsampled, allowing large deformations to be estimated and successively refined at later
243 hierarchical stages (where the data is subsampled to a finer grid). The deformation field and the objective
244 function are regularized with a Gaussian kernel at each level of the hierarchy. The ANTS algorithm is
245 freely available `http://www.picsl.upenn.edu/ANTS/`. We used an implementation of the ANTS algorithm
246 compatible with the MINC data format, mincANTS `https://github.com/vfonov/mincANTS`.

247 We used the following command line when running ANTS:

248 ```
mincANTS 3 -m PR[target_file.mnc,source_file.mnc,1,4]
```
249  ```
  --number-of-affine-iterations 10000x10000x10000x10000x10000
```
250 ```
  --affine-gradient-descent-option 0.5x0.95x1.e-4x1.e-4
```
251  ```
  --use-Histogram-Matching --MI-option 32x16000
```
252 ```
  -r Gauss[3,0] -t SyN[0.5] -i 100x100x100x20
```
253  ```
  -o transformation.xfm
```
254

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

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

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

261 **Cross-correlation Weighted Majority Vote** An optimal combination of targets from the template li-
262   brary has previously been shown to improve segmentation accuracy (Aljabar et al., 2009; Collins and

---

[2]`https://sourceforge.net/projects/advants/files/Documentation/`

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 (Chakravarty et al., 2013). 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 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 accuracy, 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 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 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.

2.2.2 Experiment 2: Hippocampal Subfield Cross-Validation

The previous experiment assesses MAGeT-Brain peformance on whole hippocampus segmentation. In this experiment, we assess MAGeT-Brain hippocampal subfield segmentation of standard 3T T1-weighted images using a leave-one-out cross-validation (LOOCV) design.

Winterburn Atlases The Winterburn atlases (Winterburn et al., 2013) are digital hippocampal segmentations of five in-vivo $0.3mm$-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 29-57 years (mean age of 37), and include two males and three females.

---

[3]`https://github.com/vfonov/EZminc`

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 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.

In this experiment, the high-resolution $0.3mm^3$ voxel Winterburn atlases are used as the MAGeT-Brain atlas library, but in each round of LOOCV a $0.9mm^3$ voxel target image corresponding to the left-out atlas is used so as to evaluate MAGeT-Brain on standard 3T T1-weighted resolution images.

Specifically, this experiments has two parts, each using different sets of $0.9mm^3$ voxel target images: trilinear subsampled versions of the Winterburn atlas images (Experiment 2a), and separately acquired T1 BRAVO acquisitions of four of the five Winterburn atlas subjects (Experiment 2b). Image subsampling of the Winterburn atlas images was performed using trilinear subsampling techniques.

The 3T T1 BRAVO acquisitions of four of the five Winterburn atlas subjects were separately obtained within a short time of the original atlas image acquisitions. 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, EFGRE-BRAVO, with the following parameters: $TE/TR/TI = 3.0ms/6.7ms/650ms$, flip angle $= 8°$, $FOV = 15.8cm$, slice thickness$= 0.9mm$, 176 in-plane steps for an approximate isotropic resolution of 0.9mm dimension voxels.

The template library is composed of all $0.9mm^3$ voxel target images (either the *BRAVO* or *subsampled* sets), plus an additional set of 16 of healthy subjects 3T T1 images acquired separately (Table 3). 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, EFGRE-BRAVO, with the following parameters: $TE/TR/TI = 3.0ms/6.7ms/650ms$, flip angle$=8°$, $FOV = 15.3cm$, slice thickness$= 0.9mm$, 170 in-plane steps for an approximate isotropic resolution of 0.9mm dimension voxels.

The optimal size of template library, registration method, and label fusion method found in Experiment 1 are used. Figure 2 illustrates schematically the experimental set up.

Evaluation method Evaluating the accuracy of hippocampal subfields for T1 images at $0.9mm^3$-isotropic voxels is inherently ill-defined since there are no manual protocols for segmentation at this resolution to use as a reference. Instead, we can evaluate the *precision* with which MAGeT-Brain produces hippocampal subfields that correspond to the segmentation protocol used by the given atlas library images. Specifically, we compute the relative percent error in volume of the MAGeT-Brain-produced hippocampal subfields with respect to the full-resolution Winterburn atlas segmentations. In addtion, by directly resampling the Winterburn atlas segmentations to $0.9mm^3$ voxels (using standard nearest-neighbour image resampling techniques) we obtain a scaled down version of the labels preserving the original segmentation protocol within the limits of error from rounding and interpolation. The relative percent error in volume of these labels provides a baseline to compare MAGeT-Brain segmentations against.

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

To validate that the MAGeT-Brain works effectively in the context of other neurological disorders, in this experiment we use the Winterburn atlases to derive whole hippocampal segmentations of a dataset of patients

Figure 2: A schematic illustration of the experimental set up of the leave-one-out cross-validation (LOOCV) subfield experiment.

Table 3: **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.

	N	Control
		$N = 16$
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.

with Schizophrenia. 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 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 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.

Manual segmentation of each subject whole hippocampus is produced 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) ignoring subfield delineations. 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 Pruessner and Winterburn hippocampal segmentation protocols differ slightly in the neuroanatomical features that are delineated (Winterburn et al., 2013). This difference 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 (Pearson) correlation in average bilateral hippocampal volume of the subjects produced by each method.

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

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

ADNI1 dataset revisited The *ADNI1:Complete 1Yr 1.5T* standardized dataset contains 1919 images in total. SNT, MAPER, 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)). Clinical and demographic data for the entire ADNI1:Complete 1Yr 1.5T dataset are shown in Table 5.

Table 4: **Schizophrenia First Episode Patient Demographics.** ambi - ambidextrous. SES - Socioeconomic Status score. FSIQ - Full Scale IQ.

	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.

374   Semi-automated segmentations of the left and right whole hippocampi are made available with a subset
375 of ADNI images (Hsu et al., 2002). These labels have been generated using the SNT tool from Medtronic
376 Surgical Navigation Technologies, Louisville, CO (see Supplementary Materials for detailed discussion of the
377 segmentation process used).

378 **Experiment details**   MAGeT-Brain was configured with an atlas library composed of the five Winterburn
379 T1 atlases (described in Experiment 2), and size of template library, registration method, and label fusion
380 method were determined by the optimal settings found in Experiment 1. The FSL FIRST segmentation
381 method was used via the `run_first_all` script according to the FIRST user guide [4]. All images in the
382 ADNI1:Complete 1Yr 1.5T dataset were segmented by both methods.

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

388 **Evaluation method**   As in Experiment 3, the SNT and Winterburn hippocampal segmentation protocols
389 differ in the neuroanatomical features delineated, and so we assessed MAGeT-Brain by the degree of (Pearson)
390 correlation of average hippocampal volume across subjects. We also computed the correlation in hippocam-
391 pal volume between existing, established automated segmentation methods – FSL FIRST, FreeSurfer, and
392 MAPER, and SNT semi-automated segmentations. Additionally, we evaluate the volume-related fixed and
393 proportional biases in all segmentation methods using Bland-Altman plots (Bland and Altman, 1986).

---

[4]http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST/UserGuide

Table 5: **ADNI1 1.5T Complete 1Yr dataset demographics.** CN - Cognitively Normal. LMCI - Late-onset Mild Cognitive Impairment. AD - Alzheimer's Disease. Hisp - Hispanic. CDR-SB - Clinical Dementia Rating-Sum of Boxes. ADAS - Alzheimer's Disease Assessment Scale. MMSE - Mini-Mental State Examination.

	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.

3   Results

3.1   Experiment 1 Results: Whole Hippocampus Cross-Validation

In this experiment we conduct 10 rounds of MAGeT-Brain and multi-atlas segmentation of of 60 subjects using atlas and template library sizes varying from 1-9 and 1-20 images respectively, two registration algorithms (ANTS or ANIMAL), and three label fusion techniques (unweighted, cross-correlation, and normalised mutual information weighted majority vote). Hippocampal MAGeT-Brain-based segmentations using both ANIMAL and ANTS registration algorithms demonstrate good overlap with manual segmentations (maximum mean DSC of 0.87 when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion); Figure 3). 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 plateaus with template libraries larger than ten images.

The use of MAGeT-Brain with ANTS registration shows a pronounced increase in segmentation accuracy over MAGeT-Brain with ANIMAL registration, across all other variable settings we tested. Additionally, by itself, using a weighted voting strategy did not significantly improve segmentation accuracy, contrary to the findings of Aljabar et al. (2009) in basic multi-atlas segmentation. Given these findings, 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 similarity score over multi-atlas-based segmentation for the same size of atlas library and voting method (Figure 4). The magnitude of improvement increases with the size of template library, and shows diminishing returns with larger atlas libraries. Peak improvement 0.025 DSC) is found with a single atlas and template library of 19 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 5). The size of template library necessary to reach a decrease ($p < 0.05$) 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 1-9 atlases.

⁴¹⁹   We have omitted results obtained when using an even number of atlases or templates since with this ⁴²⁰ configuration we found significantly decreased performance. We believe this is as a result of an inherent bias ⁴²¹ in the majority vote fusion method used (see Discussion).
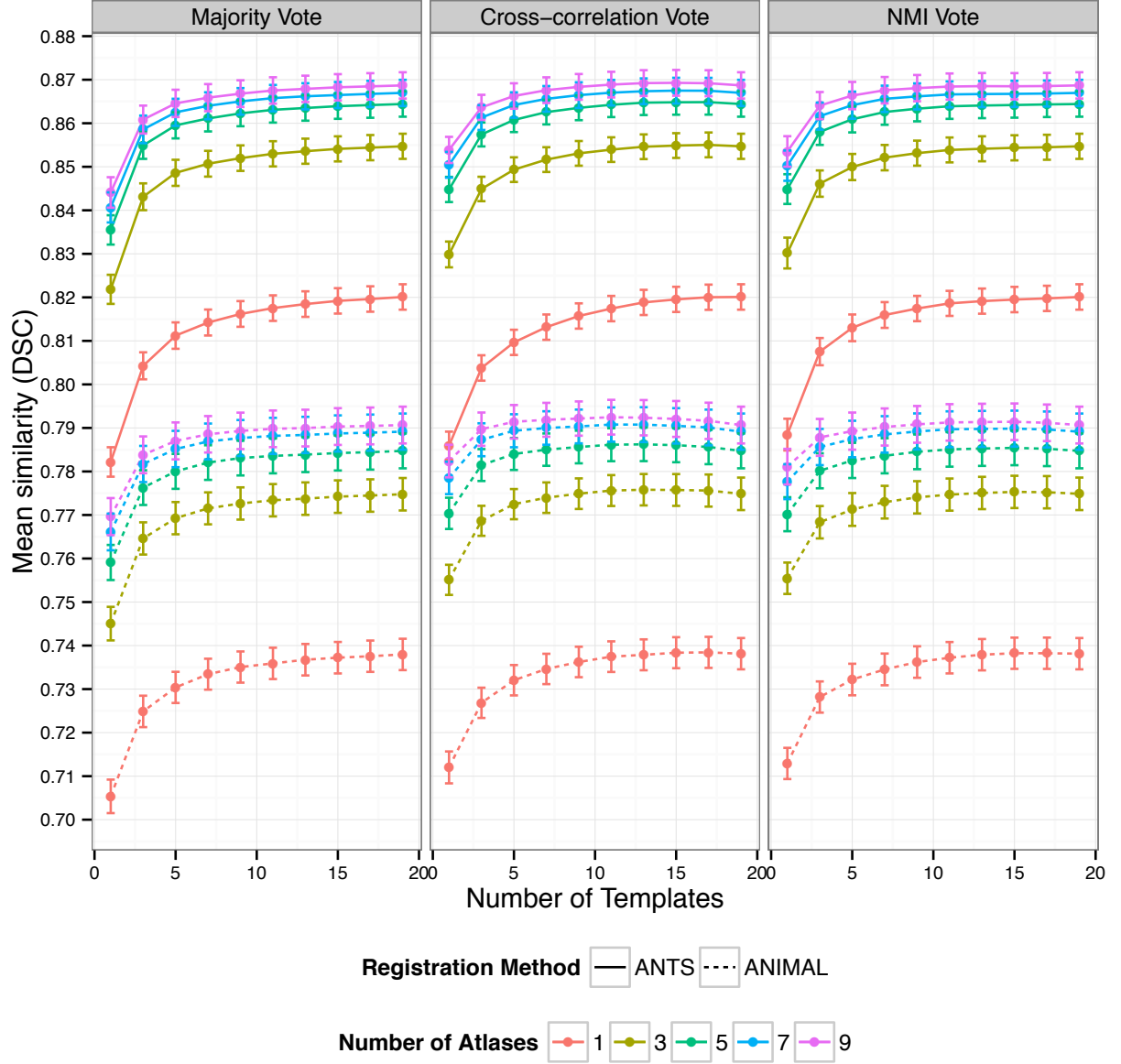
Figure 3: Mean Dice's Similarity Coefficient of MAGeT-Brain segmentations relative to Pruessner-protocol manual segmentations for 60 ADNI subjects vs. atlas and template library size, registration algorithm, and label fusion method. Error bars indicate standard error.
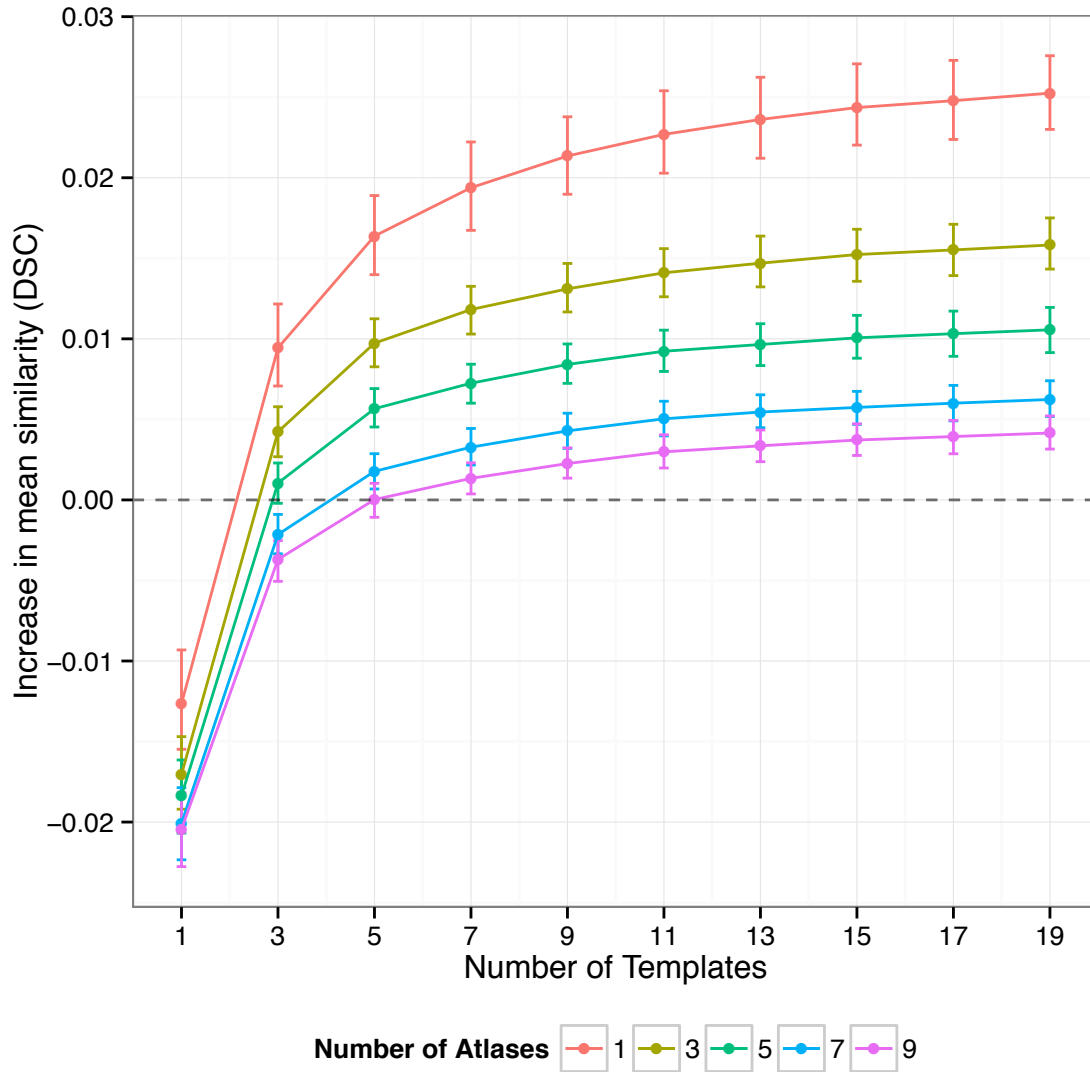
Figure 4: Increase in mean Dice's Similarity Coefficient of MAGeT-Brain over multi-atlas segmentations vs. atlas and template library size when using the ANTS registration method, and majority-vote label fusion. Segmentation similarity is computed against Pruessner-protocol manual segmentations. Error bars indicate standard error.
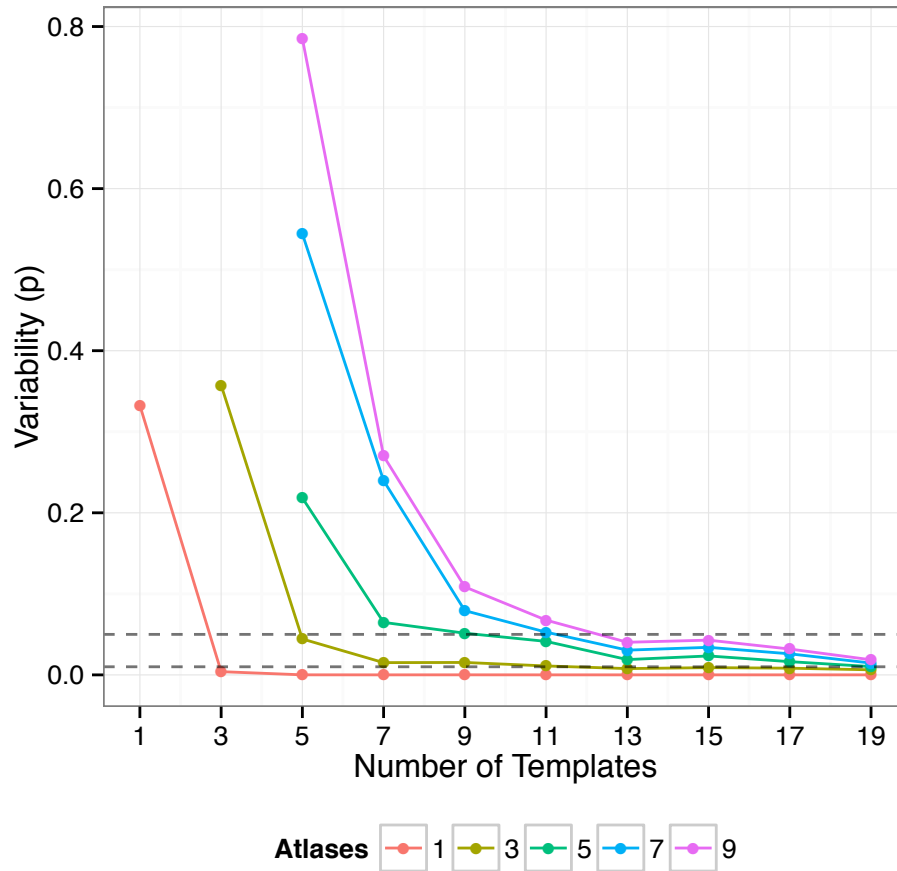
Figure 5: **Difference in variability of MAGeT-Brain vs. multi-atlas segmentation with respect to manual segmentation.** Variance of segmentation accuracy between MAGeT-Brain and multi-atlas segmentation is computed for each subject across all ten rounds of validation. Shown on the y-axis is the p-value 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. Dashed lines indicate a p-value of 0.05 and 0.01.

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 was conducted in which lower-resolution images ($0.9mm^3$ voxels) of each Winterburn atlas subject were segmented using the remaining high-resolution Winterburn atlas subjects' images.

In general, across hippocampal subfields the percent error in volume of MAGeT-Brain segmentations (relative to the full-resolution Winterburn atlas segmentation) compares favourably to the error resulting from image resampling (Figure 6). In particular, there is no significant difference between percent error of MAGeT-Brain volumes compared to image resampling error for the CA1, Subiculum, and CA2/CA3 subregions. The MAGeT-Brain volumes of CA4/DG and SR/SL/SM subregions show both significantly different (lower) percent error (p ¡ 0.001).

Figure 7 shows a qualitative comparison of MAGeT-Brain subfield segmentations for a single subject.
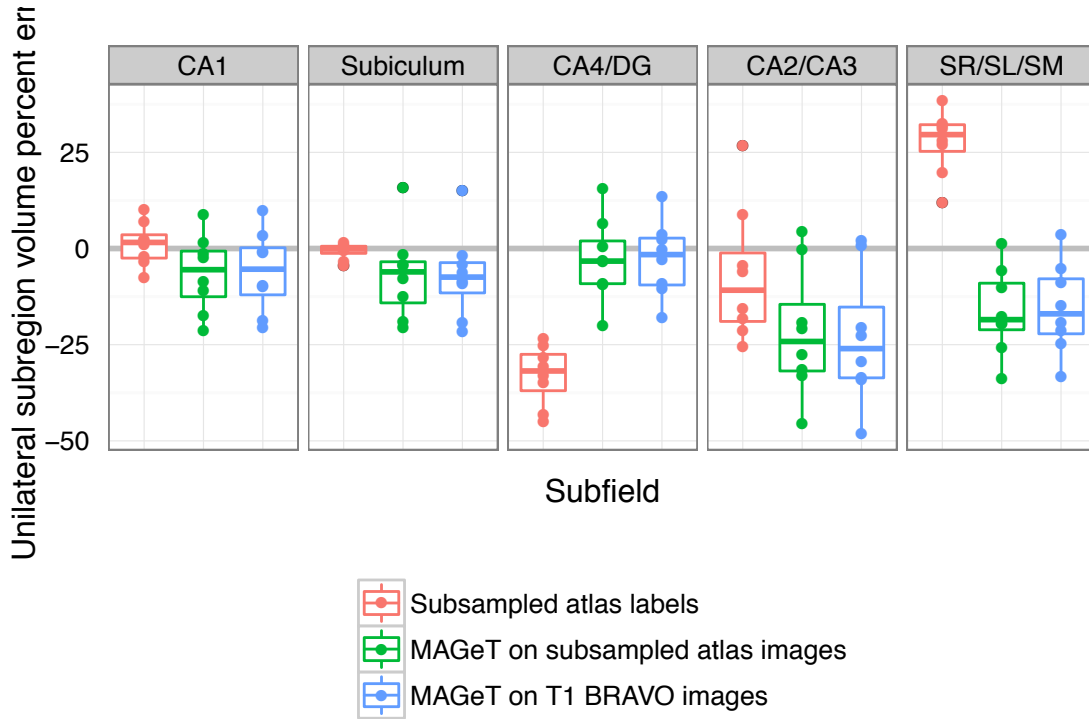
Figure 6: **Percent error in unilateral subfield volume.** Percent error is measured against the volumes of the unmodified Winterburn atlas subfield segmentations. **Subsampled atlas labels** volumes of the manual segmentations of the Winterburn atlases after resampling to $0.9mm^3$ voxels. **MAGeT on subsampled atlas images** volumes are MAGeT-Brain segmentations of the Winterburn atlas images after resampling to $0.9mm^3$ voxels. **MAGeT on T1 BRAVO** volumes are MAGeT-Brain segmentations of T1 BRAVO images ($0.9mm^3$ voxels) acquired separately of four of the five Winterburn atlas subjects.
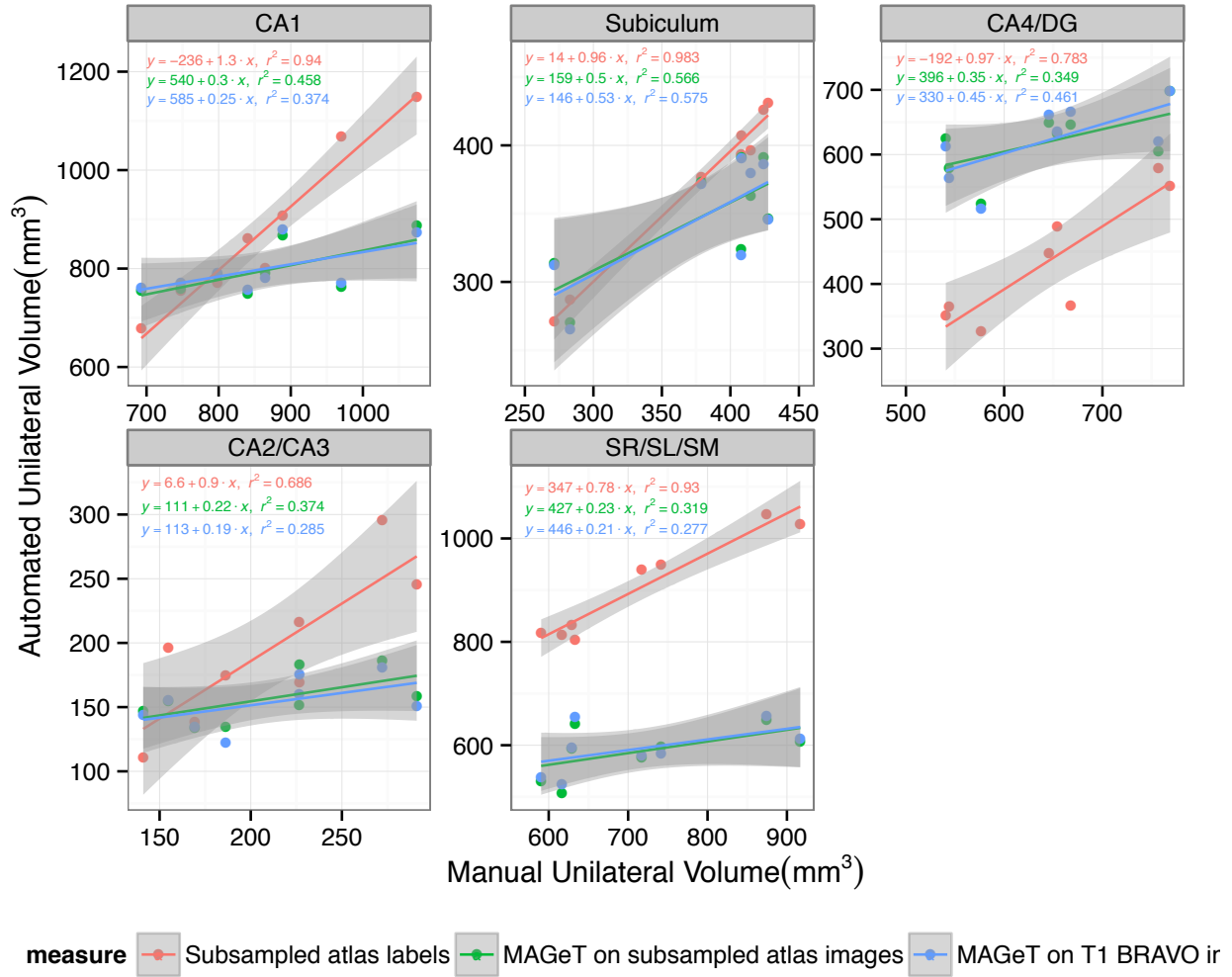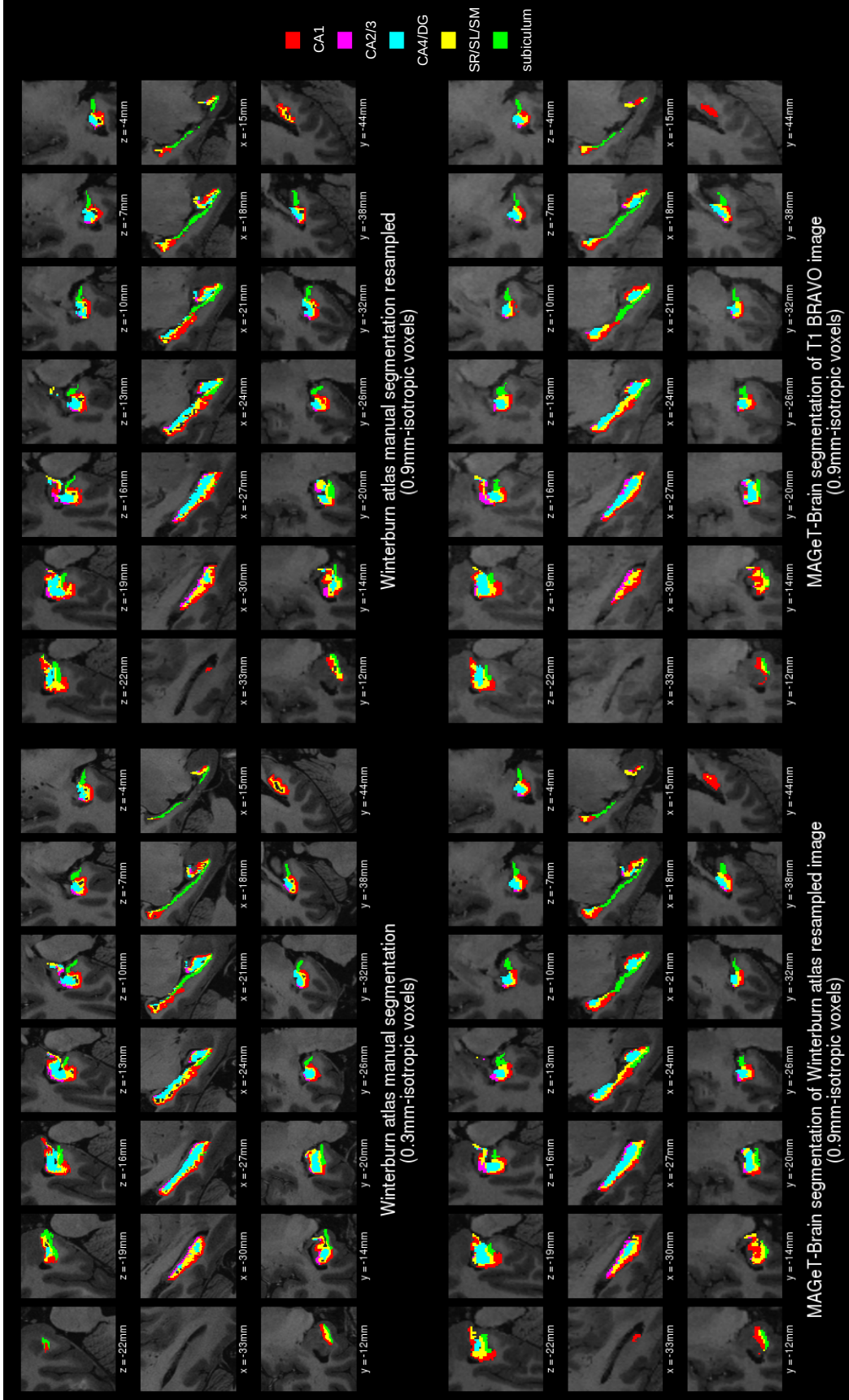
Figure 7: Detailed subfield segmentation results for a single subject. In the upper left corner is the original high-resolution Winterburn atlas manual subfield segmentation; in the upper right corner is the Winterburn atlas segmentation subsampled from 0.3mm- to 0.9mm-isotropic voxels; in the lower left corner is the MAGeT-Brain segmentation of the subsampled Winterburn atlas images; in the lower right corner is the MAGeT-Brain segmentation of a separately acquired T1 BRAVO image of the same subject. 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.

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 a gold standard.

MAGeT-Brain produces hippocampal volumes that are highly correlated with manual segmentation volumes (Pearson $r = 0.877$, $t = 16.244$, $p < 0.001$; Figure 8).
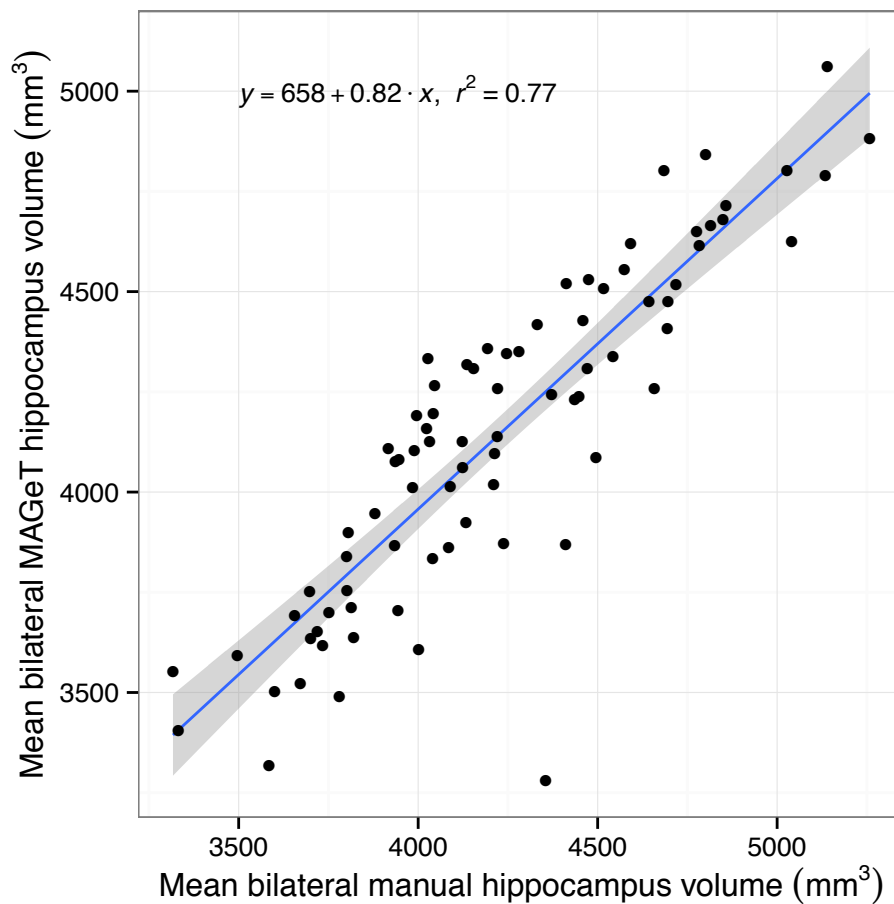
Figure 8: Mean bilateral hippocampus volume as measured by MAGeT-Brain vs. manually segmented volumes from the First Episode Patients with Schizophrenia dataset. A linear fit line is shown, with shaded region showing standard error.

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	368	368	368	368	368
Failures	n/a	30	n/a	20	88

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

Based on the results from Experiment 1, in this experiment MAGeT-Brain was configured with a template library of 21 randomly chosen subject images (7 from each disease class) and majority vote label fusion. The entire ADNI1:Complete 1Yr 1.5T dataset was segmented by MAGeT-Brain, and the resulting volumes compared with those obtained by expert semi-automated segmentation (SNT), and three other automated segmentation techniques: MAPER, FreeSurfer, and FSL FIRST. Table 6 shows the total count of segmentations available, including a count of those which have failed a quality control inspection. A total of 246 images are included in the following analysis, having met quality criteria and having segmentations from every method.

We found a close relationship in total bilateral hippocampal volume between all methods and the SNT semi-automated label volumes (Figure 10). Volumes are correlated with Pearson $r > 0.78$ for all methods across disease categories. Within disease categories (Figure 9), 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.

Bland-Altman plots illustrate the level of agreement of each method with SNT segmentation hippocampal volumes (Figure 11). 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). 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 (Figure 11). Additionally, all methods show a fixed volume bias, with FreeSurfer and FSL FIRST 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 12 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 visual inspection. Mostly frequently, we found MAGeT-Brain improperly includes the vestigial hippocampal sulcus and, although not anatomically incorrect, MAGeT-Brain under-estimates the hippocampal body in comparison to the SNT segmentation.
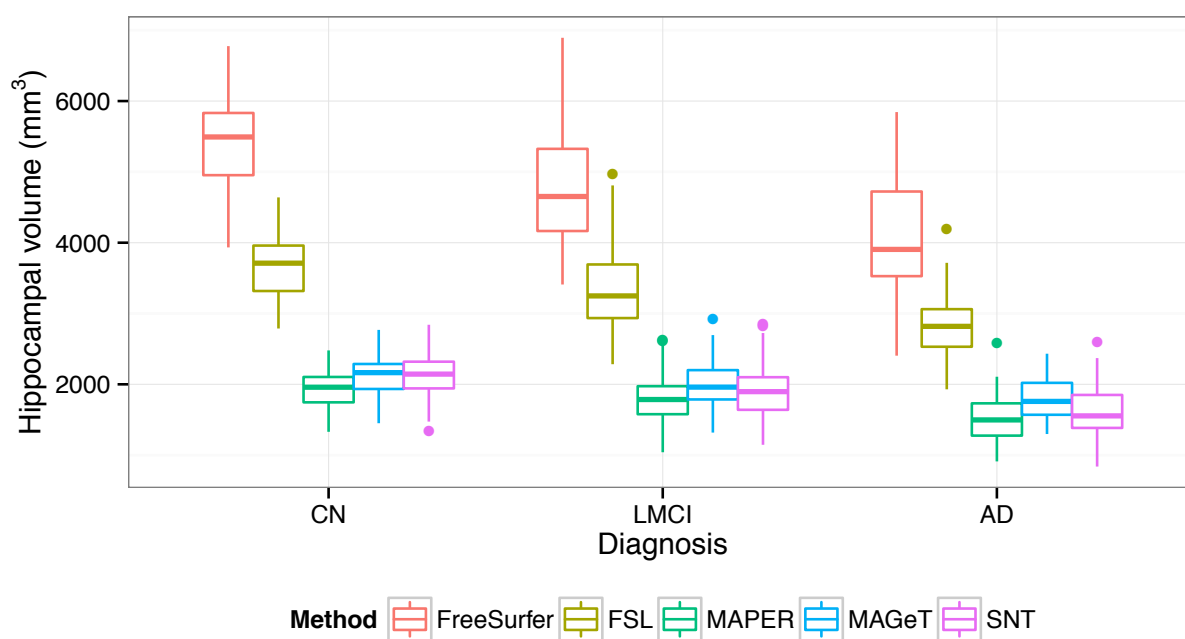
25

Figure 9: Subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by FreeSurfer, FSL FIRST, MAPER, MAGeT-Brain, and SNT vs. disease category.
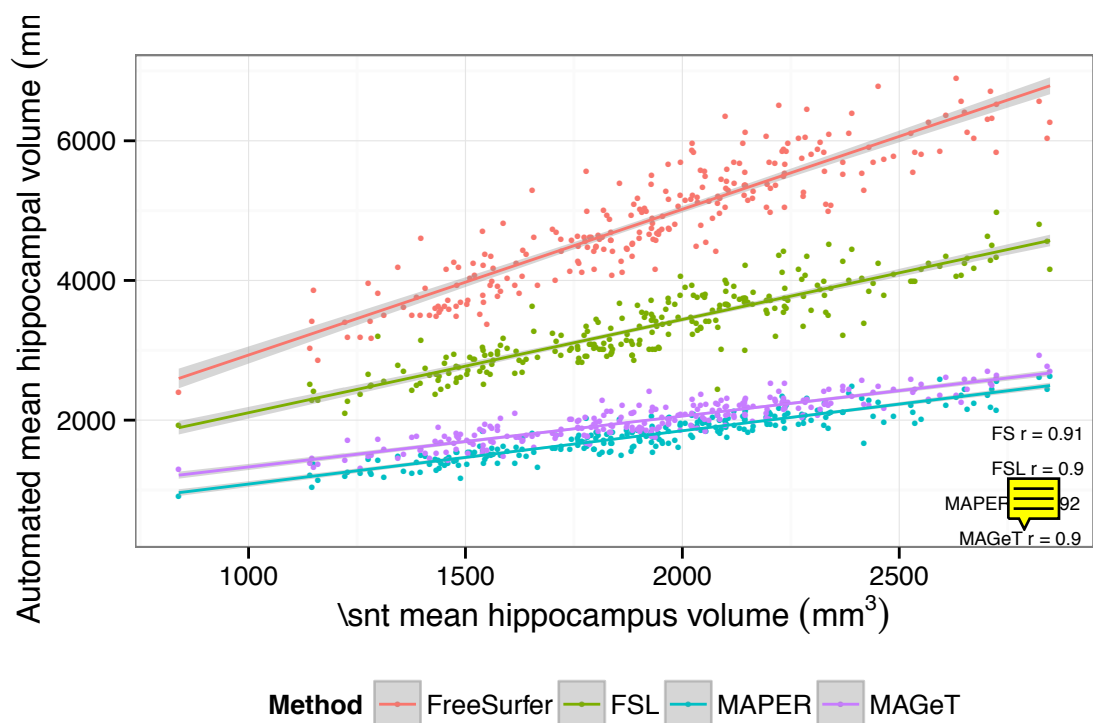
Figure 10: Subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by each of the four automated methods investigated (FreeSurfer (FS), FSL FIRST, MAPER, MAGeT-Brain) vs. SNT. Linear fit lines and Pearson correlations are shown for each method.
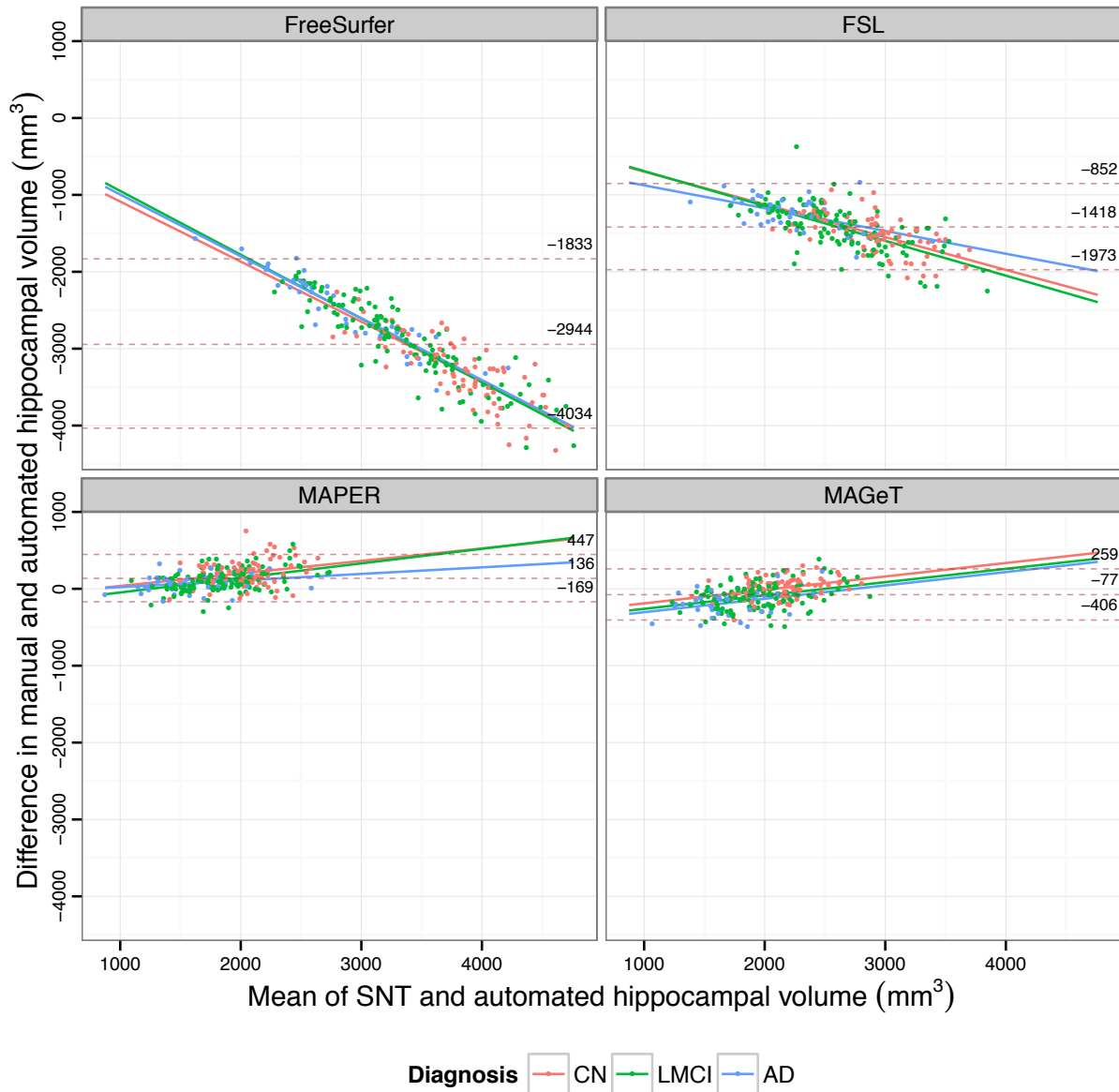
Figure 11: Bland-Altman plots comparing subject mean hippocampal volume as measured in the ADNI1:Complete 1Yr 1.5T dataset by SNT segmentation and each of the four automated methods investigated (FreeSurfer, FSL FIRST, MAPER, MAGeT-Brain). The overall mean difference in volume, and limits of agreement ($\pm 1.96 SD$) 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 SNT volume, and vice versa.
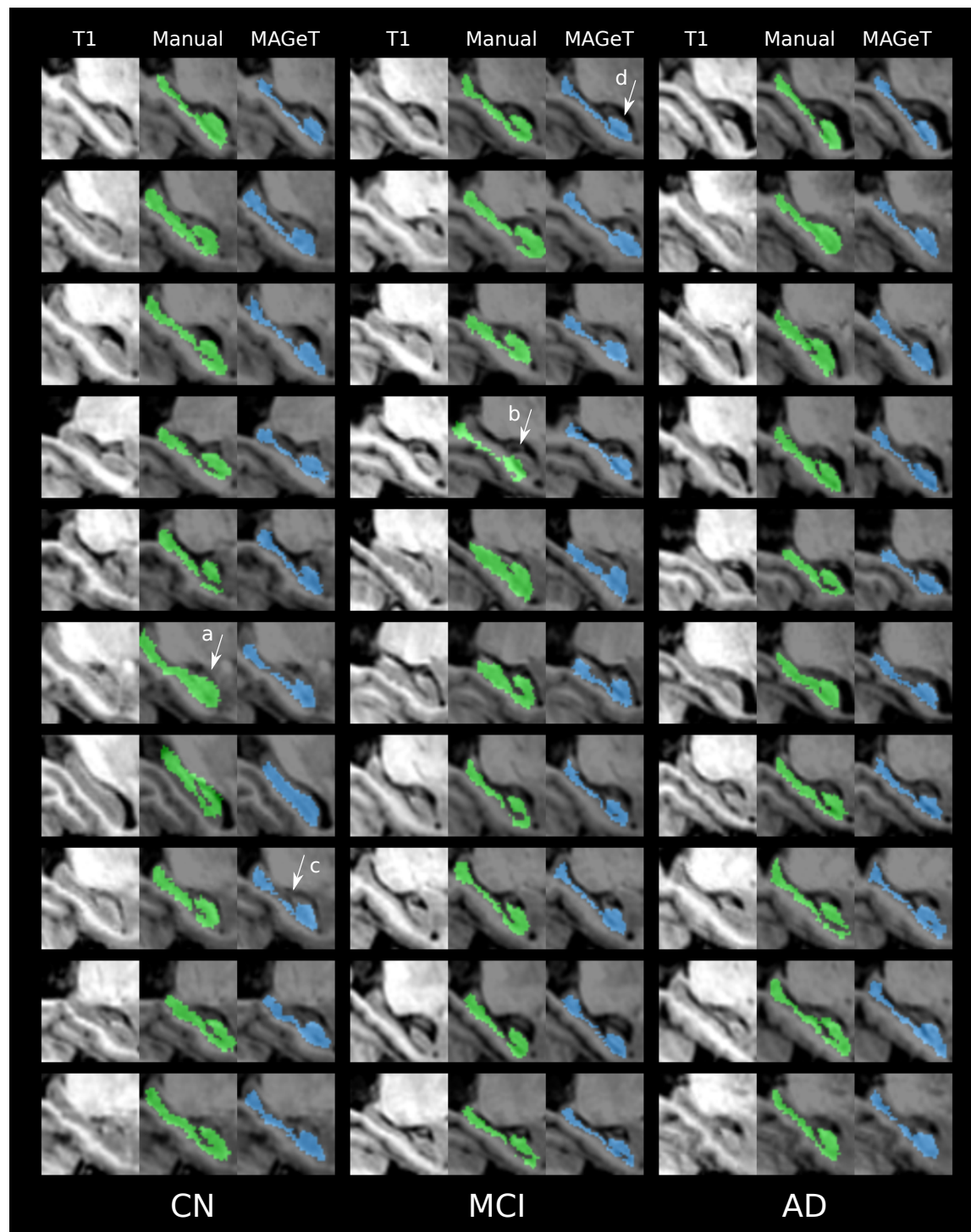
Figure 12: SNT and MAGeT-Brain segmentations for 30 ADNI subjects (10 subjects randomly selected from each disease category in the subject pool used in Experiment 1). Sagittal slices are shown for each unlabelled T1-weighted anatomical image. SNT labels appear in green, and MAGeT-Brain labels appear 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.

Table 7: **Automated segmentation accuracy (overlap with SNT labels) of the ADNI dataset.** For each method, the number of 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 automated 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
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
MAGeT-Brain	9	0.869		10 rounds of Monte Carlo CV on a pool of 69 subjects
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. Experiment 1 establishes MAGeT-Brain as effective and as comparable to multi-atlas segmentation. This experiment also rigorously characterizes the behaviour of MAGeT-Brain whole hippocampal segmentation under various parameter settings, allowing us to choose an optimal setting for subsequent experiments. Experiment 2 demonstrates the accuracy of MAGeT-Brain hippocampal subfield identification despite contrast and resolution limitations in standard T1-weighted image volumes. Experiments 2 and 3 validate MAGeT-Brain whole hippocampal segmentation accuracy and consistency on populations with different ageing and neuropsychiatric characteristics. Furthermore, taken together, these experiments demonstrate 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).

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.

Previous work on multi-atlas segmentation methods (Aljabar et al., 2009; Collins and Pruessner, 2010) has found that cross-correlation and normalized mutual information-based weighted label fusion improves segmentation accuracy over simple majority vote label fusion. 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 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 weights 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 accuracy and 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.

To this end, 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 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)).

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 similarly rigorous validation a 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 consistently segmented).

On that note, one author (JW), an expert manual ra identified regular inconsistencies in the SNT segmentations: occurrences of over- and under-estimation, as well as misalignments of the entire segmentation volume (Figure 12). Although the SNT segmentations are used as gold standards for validation in many other studies (Table 7) providing a benchmark for comparison, their use in Experime and 4 presents the possibility that the DSC scores and correlations we obtained are lower than what would be found with a more accurate and consistent gold standard.

In comparison to other methodologies in the field MAGeT-Brain performs favourably. Table 7 surveys some of the most recent reported DSC values reported on ADNI dataset, using SNT segmentations for the atlas 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

require more atlases (between 16-55) than our MAGeT-Brain implementation with the Winterburn atlases (Winterburn et al., 2013).

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 DSC scores for the segmentation of ADNI data). While this method is effective 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, and then use this method to segment an ADNI sample, with a hierarchical experimentation protocol. 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).

Among the automated segmentation methods we compared in this paper (FreeSurfer, MAPER, FSL FIRST-FIRST), we find extremely variable performance of all methods. With the exception of FSL FIRST all methods correlate well with the SNT volumes provided in the ADNI database. However, FreeSurfer and FSL 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, MAGeT-Brain and MAPER are far more conservative in volume estimation, suggesting 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 also demonstrated that 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 3T data (Yushkevich et al., 2009; Van Leemput et al., 2009). While recent work demonstrates that subfield segmentations can be used for classification of AD, MCI, and CN, 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.

Table 8: **Summary of labelled subfields of the Hippocampus from recent MRI segmentation protocols.**

Protocol	Labelled Subfields
Winterburn et al. (2013)	CA1, CA2/CA3, CA4/dentate gyrus, strata radiatum/lacunosum/moleculare, subiculum
Wisse et al. (2012)	CA1, CA2, CA3, CA4/DG, subiculum, entorhinal cortex
Van Leemput et al. (2009)	CA1, CA2/CA3, CA4/DG, presubiculum, subiculum, hippocampal fissure, fimbria, hippocampal tail, inferior lateral ventricle, choroid plexus
Yushkevich et al. (2009)	CA1, CA2/CA3, dentate gyrus (hilus), dentate gyrus (stratum moleculare), strata radiatum/lacunosom/moleculare/vestigial hippocampal sulcus

The fact that CA2/CA3 and molecular layers cannot be reasonably identified should not be surprising as these are extremely thin and spatially convoluted regions that originally required high-resolution MRI for identification. It is likely that the extents of these regions are well below the resolution offered by standard T1-weighted images and, in fact, many manual segmentation methodologies do not attempt to parse these regions either (Wisse et al., 2012; Mueller and Weiner, 2009).

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

5   Acknowledgements

6 Supplementary Materials

6.1 ADNI SNT Hippocampal Labels

⁶¹⁴ Semi-automated hippocampal volumetry was carried out using a commercially available high dimensional
⁶¹⁵ brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO), that has previously been
⁶¹⁶ validated and compared to manual tracing of the hippocampus (Hsu et al., 2002). Measurement of hippocam-
⁶¹⁷ pal volume is achieved first by placing manually 22 control points as local landmarks for the hippocampus on
⁶¹⁸ the individual brain MRI data: one landmark at the hippocampal head, one at the tail, and four per image
⁶¹⁹ (i.e., at the superior, inferior, medial and lateral boundaries) on five equally spaced images perpendicular
⁶²⁰ to the long axis of the hippocampus. Second, fluid image transformation is used to match the individual
⁶²¹ brains to a template brain (Christensen et al., 1997). The pixels corresponding to the hippocampus are then
⁶²² labeled and counted to obtain volumes. This method of hippocampal voluming has a documented reliability
⁶²³ of an intraclass coefficient better than .94 (Hsu et al., 2002).

6.2 Experiment 5: Whole Hippocampus Segmentation Cross-Validation with SNT Segmentations

⁶²⁶ Monte Carlo Cross-Validation (MCCV) (Shao, 1993) was performed using a pool of images and SNT semi-
⁶²⁷ automated segmentations from the ADNI dataset (Hsu et al., 2002). This form of cross-validation allows
⁶²⁸ us to rigorously validate a large number of parameter settings of MAGeT-Brain (atlas and template library
⁶²⁹ sizes, registration algorithm, and label fusion method) and select the best parameters to use in subsequent
⁶³⁰ experiments.

6.2.1 Methods

ADNI1:Complete 1Yr 1.5T dataset 69 1.5T images were arbitrarily selected from the baseline scans
⁶³³ in the *ADNI1:Complete 1Yr 1.5T* standardized dataset. 23 subjects were chosen from each disease category:
⁶³⁴ cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). Demographics for
⁶³⁵ this subset are shown in Table 1.

⁶³⁶ Clinical, demographic and pre-processed T1-weighted MRI were downloaded by the authors from the
⁶³⁷ ADNI database (adni.loni.ucla.edu) between March 2012 and August 2012. The image dataset used was

Table 9: **ADNI1 SNT cross-validation subset demographics.** CN - Cognitively Normal. LMCI - Late-onset Mild Cognitive Impairment. AD - Alzheimer's Disease. Hisp - Hispanic. CDR-SB - Clinical Dementia Rating-Sum of Boxes. ADAS - Alzheimer's Disease Assessment Scale. MMSE - Mini-Mental State Examination.

	CN	LMCI	AD	Combined
	$N = 23$	$N = 23$	$N = 23$	$N = 69$
Age at baseline ₓYears	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.

the "ADNI1:Complete 1Yr 1.5T" standardized dataset available from ADNI [5] (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 = 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 dimensions of $1.25mm \times 1.25mm \times 1.2mm$.

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 SNT segmentations. In each validation round, the dataset is partitioned into a training set consisting of images and SNT segmentations used as an atlas library, and a validation set consisting of the remaining images segmented by both MAGeT-Brain and multi-atlas. The computed segmentations were compared to the SNT segmentations (see 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 normalized mutual information weighted majority vote, described below). 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 compared two non-linear image registration methods, three label fusion methods.

The automatically produces labels are compared to the SNT labels provided by ADNI. The segmentation

---

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

accuracy reported for a parameter setting is the average over ten validation rounds.

6.2.2 Results

In this experiment we conduct 10 rounds of MAGeT-Brain and multi-atlas segmentation of of 69 subjects using atlas and template library sizes varying from 1-9 and 1-20 images respectively, two registration algorithms (ANTS or ANIMAL), and three label fusion techniques (unweighted, cross-correlation, and normalised mutual information weighted majority vote). Hippocampal MAGeT-Brain-based segmentations using both ANIMAL and ANTS registration algorithms demonstrate good agreement with SNT segmentations (maximum mean DSC of 0.84 when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion); Figure 3). 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 plateaus with template libraries larger than ten images.

The use of MAGeT-Brain with ANTS registration shows a pronounced increase in segmentation accuracy over MAGeT-Brain with ANIMAL registration, across all other variable settings we tested. Additionally, by itself, using a weighted voting strategy did not significantly improve segmentation accuracy, contrary to the findings of Aljabar et al. (2009) in basic multi-atlas segmentation. Given these findings, 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 similarity score over multi-atlas-based segmentation for the same size of atlas library and voting method (Figure 14). The magnitude of improvement increases with the size of template library, and shows diminishing returns with larger atlas libraries. Peak improvement 0.023 DSC) is found with a single atlas and template library of 19 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 15). The size of template library necessary to reach a decrease ($p < 0.05$) 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).
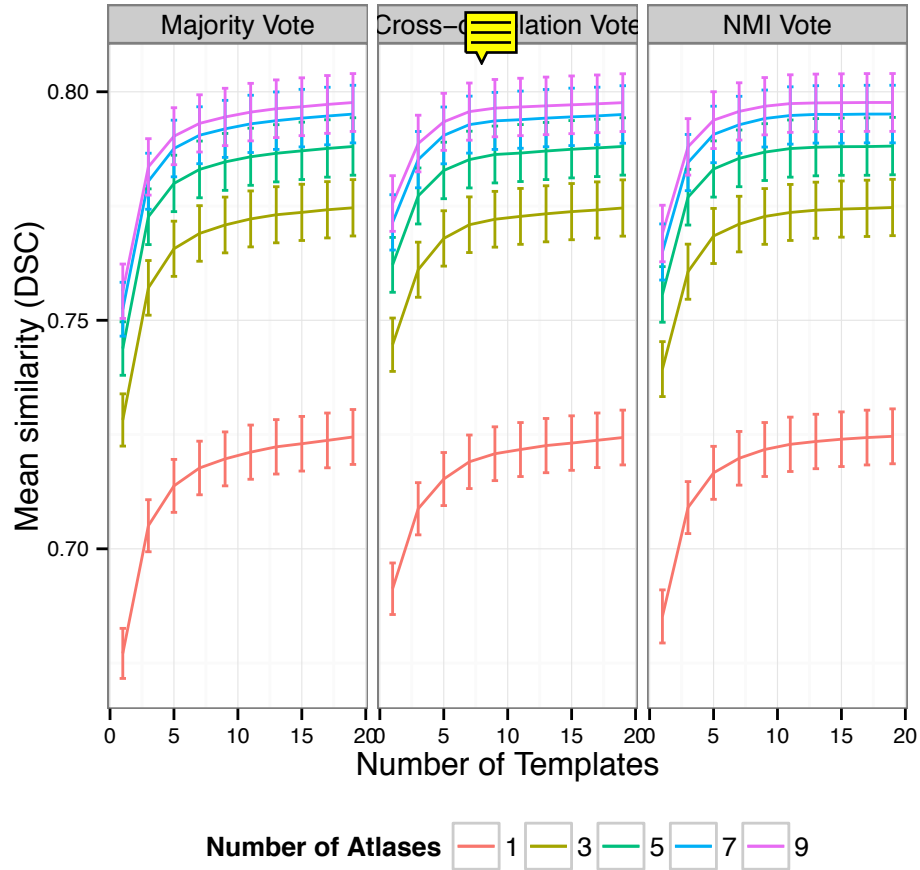
Figure 13: Mean Dice's Similarity Coefficient of MAGeT-Brain segmentations relative to SNT segmentations for 69 ADNI subjects vs. atlas and template library size, registration algorithm, and label fusion method. Error bars indicate standard error.
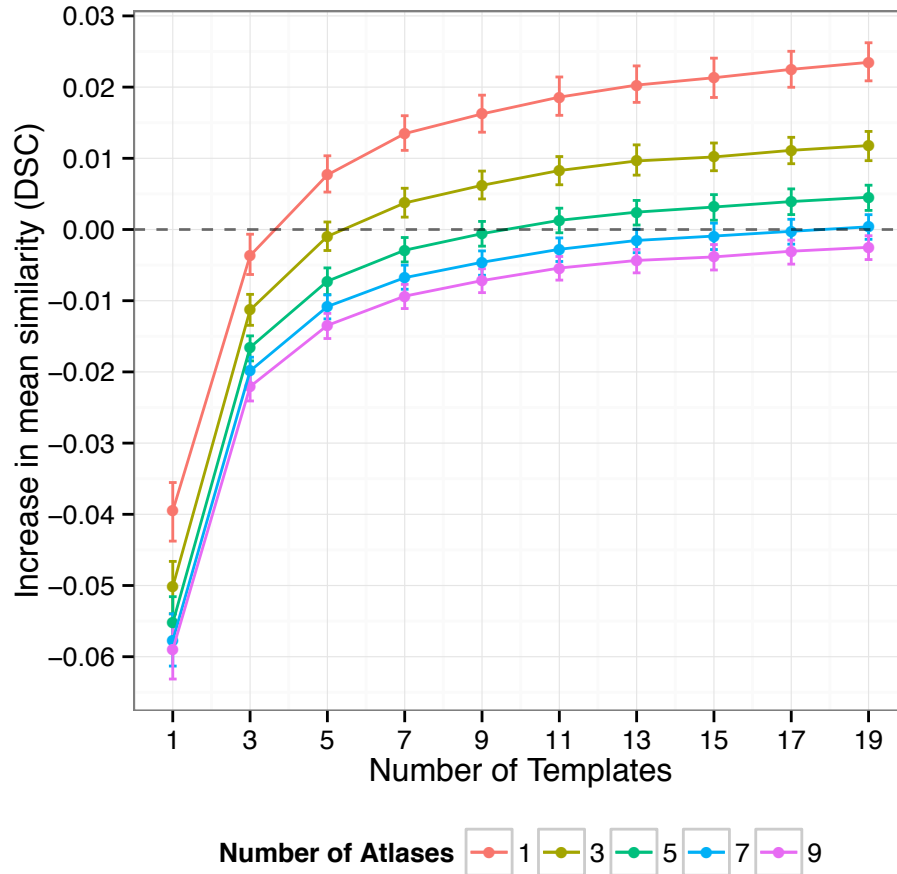
Figure 14: Increase in mean Dice's Similarity Coefficient of MAGeT-Brain over multi-atlas segmentations vs. atlas and template library size when using the ANTS registration method, and majority-vote label fusion. Segmentation similarity is computed against SNT segmentations. Error bars indicate standard error.
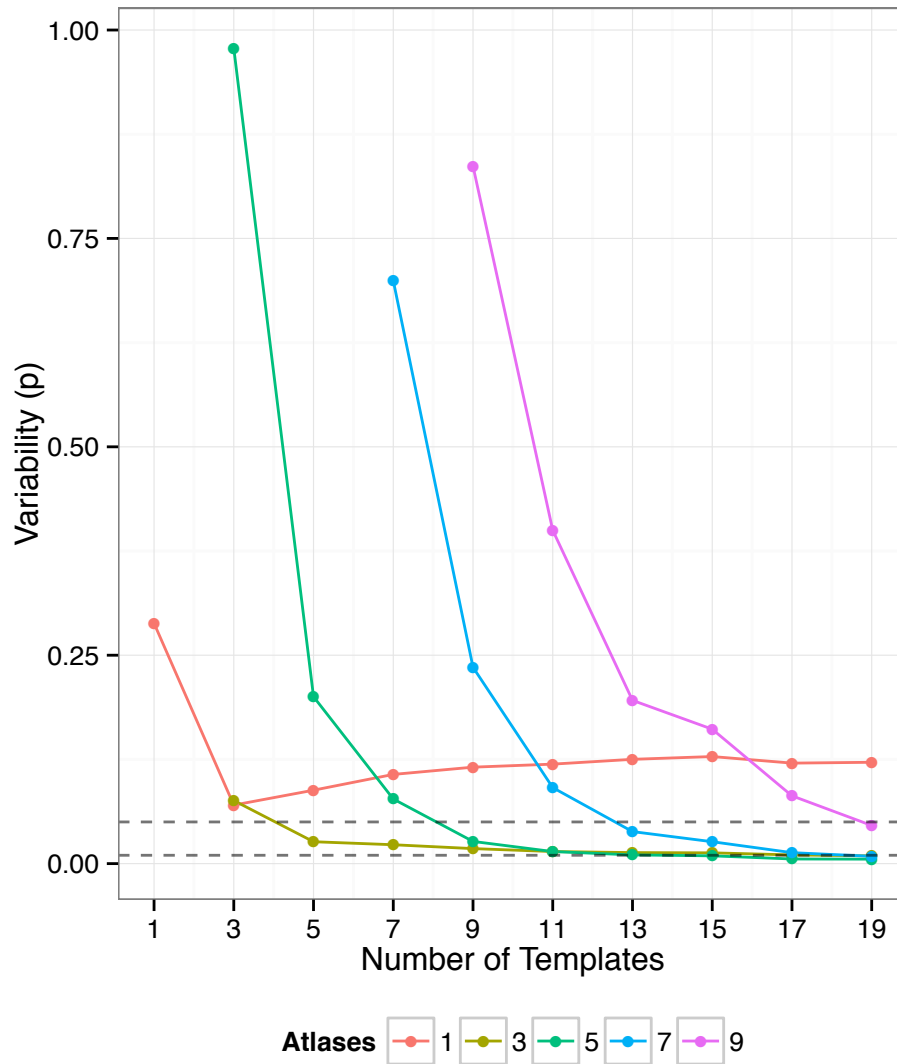
Figure 15: **Difference in variability of MAGeT-Brain vs. multi-atlas segmentation with respect to SNT segmentations.** Variance of segmentation accuracy between MAGeT-Brain and multi-atlas segmentation is computed for each subject across all ten rounds of validation. Shown on the y-axis is the p-value 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. Dashed lines indicate a p-value of 0.05 and 0.01.

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.

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.

J. M. Bland and D. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, pages 307–310, 1986.

M. Boccardi, M. Bocchetta, L. G. Apostolova, G. Preboske, N. Robitaille, P. Pasqualetti, L. D. Collins, S. Duchesne, C. R. Jack, and G. B. Frisoni. Establishing Magnetic Resonance Images Orientation for the EADC-ADNI Manual Hippocampal Segmentation Protocol. *Journal of neuroimaging : official journal of the American Society of Neuroimaging*, pages 1–6, Nov. 2013a. ISSN 1552-6569. doi: 10.1111/jon.12065.

M. Boccardi, M. Bocchetta, R. Ganzola, N. Robitaille, A. Redolfi, S. Duchesne, C. R. Jack, and G. B. Frisoni. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, pages 1–11, May 2013b. ISSN 1552-5279. doi: 10.1016/j.jalz.2013.03.001.

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.

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.

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.

M. M. Chakravarty, P. Steadman, M. C. van Eede, R. D. Calcott, V. Gu, P. Shaw, A. Raznahan, D. L. Collins, and J. P. Lerch. Performing label-fusion-based segmentation using multiple automatically generated templates. *Human brain mapping*, 34(10):2635–54, Oct. 2013. ISSN 1097-0193. doi: 10.1002/hbm.22092.

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.

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

726 cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–87, June 2009. ISSN 1098-1063.
727 doi: 10.1002/hipo.20626.

728 D. L. Collins and J. C. Pruessner. Towards accurate, automatic segmentation of the hippocampus and
729 amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52
730 (4):1355–66, Oct. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.04.193.

731 D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR
732 volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192–205,
733 1994. ISSN 0363-8715.

734 D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3-D model-based neuroanatomical
735 segmentation. *Human Brain Mapping*, 3(3):190–208, Oct. 1995. ISSN 10659471. doi: 10.1002/hbm.
736 460030304.

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

740 P. Coupé, S. F. Eskildsen, J. V. Manjón, V. S. Fonov, and D. L. Collins. Simultaneous segmentation and
741 grading of anatomical structures for patient's classification: application to Alzheimer's disease. *NeuroIm-*
742 *age*, 59(4):3736–47, Feb. 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.10.080.

743 J. G. Csernansky, S. Joshi, L. Wang, J. W. Haller, M. Gado, J. P. Miller, U. Grenander, and M. I. Miller. Hip-
744 pocampal morphometry in schizophrenia by high dimensional brain mapping. *Proceedings of the National*
745 *Academy of Sciences of the United States of America*, 95(19):11406–11411, 1998.

746 T. den Heijer, F. V. der Lijn, M. W. Vernooij, M. de Groot, P. J. Koudstaal, a. V. der Lugt, G. P.
747 Krestin, a. Hofman, W. J. Niessen, and M. M. B. Breteler. Structural and diffusion MRI measures of
748 the hippocampus and memory performance. *NeuroImage*, 63(4):1782–9, Dec. 2012. ISSN 1095-9572. doi:
749 10.1016/j.neuroimage.2012.08.067.

750 B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany,
751 D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation:
752 automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–55, Jan. 2002.
753 ISSN 0896-6273.

754 E. Geuze, E. Vermetten, and J. D. Bremner. MR-based in vivo hippocampal volumetrics: 2. Findings in
755 neuropsychiatric disorders. *Molecular Psychiatry*, 10(2):160, Sept. 2004. doi: 10.1038/sj.mp.4001579.

756 J. W. Haller, A. Banerjee, G. E. Christensen, M. Gado, S. Joshi, M. I. Miller, Y. Sheline, M. W. Van-
757 nier, and J. G. Csernansky. Three-dimensional hippocampal MR morphometry with high-dimensional
758 transformation of a neuroanatomic atlas. *Radiology*, 202(2):504–510, 1997.

759 M. Hartig, D. Truran-sacrey, S. Raptentsetsang, N. Schuff, and M. Weiner. USCF FreeSurfer Overview and
760 QC Ratings. 2010.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

799 J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike,
800 C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-
801 Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma,
802 H. Hulshoff Pol, T. Cannon, R. Kawashima, and B. Mazoyer. A four-dimensional probabilistic atlas of
803 the human brain. *Journal of the American Medical Informatics Association : JAMIA*, 8(5):401–30. ISSN
804 1067-5027.

805 J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson,
806 B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts,
807 N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, D. Boomsma,
808 T. Cannon, R. Kawashima, and B. Mazoyer. A probabilistic atlas and reference system for the hu-
809 man brain: International Consortium for Brain Mapping (ICBM). *Philosophical transactions of the Royal*
810 *Society of London. Series B, Biological sciences*, 356(1412):1293–322, Aug. 2001. ISSN 0962-8436. doi:
811 10.1098/rstb.2001.0915.

812 J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas of the human brain:
813 theory and rationale for its development. The International Consortium for Brain Mapping (ICBM).
814 *NeuroImage*, 2(2):89–101, June 1995. ISSN 1053-8119.

815 J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, A. W.
816 Toga, C. R. Jack, M. W. Weiner, and P. M. Thompson. Validation of a fully automated 3D hippocampal
817 segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly
818 controls. *NeuroImage*, 43(1):59–68, Oct. 2008. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.07.003.

819 S. G. Mueller and M. W. Weiner. Selective effect of age, Apo e4, and Alzheimer's disease on hippocampal
820 subfields. *Hippocampus*, 19(6):558–64, June 2009. ISSN 1098-1063. doi: 10.1002/hipo.20614.

821 K. L. Narr, P. M. Thompson, P. Szeszko, D. Robinson, S. Jang, R. P. Woods, S. Kim, K. M. Hayashi,
822 D. Asunction, A. W. Toga, and R. M. Bilder. Regional specificity of hippocampal volume reductions
823 in first-episode schizophrenia. *NeuroImage*, 21(4):1563–75, Apr. 2004. ISSN 1053-8119. doi: 10.1016/j.
824 neuroimage.2003.11.011.

825 S. M. Nestor, E. Gibson, F.-Q. Gao, A. Kiss, and S. E. Black. A Direct Morphometric Comparison of Five
826 Labeling Protocols for Multi-Atlas Driven Automatic Segmentation of the Hippocampus in Alzheimer's
827 Disease. *NeuroImage*, Nov. 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2012.10.081.

828 Z. Pausova, T. Paus, M. Abrahamowicz, J. Almerigi, N. Arbour, M. Bernard, D. Gaudet, P. Hanzalek,
829 P. Hamet, A. C. Evans, M. Kramer, L. Laberge, S. M. Leal, G. Leonard, J. Lerner, R. M. Lerner, J. Math-
830 ieu, M. Perron, B. Pike, A. Pitiot, L. Richer, J. R. Séguin, C. Syme, R. Toro, R. E. Tremblay, S. Veillette,
831 and K. Watkins. Genes, maternal smoking, and the offspring brain and body during adolescence: design
832 of the Saguenay Youth Study. *Human brain mapping*, 28(6):502–18, June 2007. ISSN 1065-9471. doi:
833 10.1002/hbm.20402.

834 J. Poppenk and M. Moscovitch. A Hippocampal Marker of Recollection Memory Ability among Healthy
835 Young Adults: Contributions of Posterior and Anterior Segments. *Neuron*, 72(6):931–937, Dec. 2011.
836 ISSN 0896-6273. doi: 10.1016/j.neuron.2011.10.014.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

J. L. Winterburn, J. C. Pruessner, S. Chavez, M. M. Schira, N. J. Lobaugh, A. N. Voineskos, and M. M. Chakravarty. A novel in vivo atlas of human hippocampal subfields using high-resolution 3 T magnetic

resonance imaging. *NeuroImage*, 74:254–65, July 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013. 02.003.

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.

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.

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.

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.

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.