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

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 aggreement to multi-atlas approaches. Thus, our method acts as an ~~accurate~~ reliable multi-atlas approach when using special ~~,~~ or hard-to-define atlases that are laborious to construct.

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

We conduct two separate Monte Carlo cross-validation experiments comparing MAGeT-Brain and multi-atlas whole hippocampal segmentation using differing atlas and template library sizes, and registation and label fusion methods. The first experiment is a 10-fold validation (per parameter setting) over 60 subjects taken from the Alzheimer's Disease Neuroimaging Database (ADNI), and the second is a five-fold validation over 81 subjects having had a first episode of psychosis. In both cases, automated segmentations are compared with manual segmentions following the Pruessner-protocol. Using

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

the best settings found from these experiments, we segment 246 images of the ADNI1:Complete 1Yr 1.5T dataset and compare these with segmentations from existing automated methods: FSL FIRST, FreeSurfer, MAPER, and SNT. Finally, we conduct a leave-one-out cross-validation (LOOCV) of hippocampal subfield segmentation in standard 3T T1-weighted images, using five high-resolution manually segmented atlases (Winterburn et al., 2013).

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

Conclusion: We demonstrate that MAGeT-Brain produces ~~accurate~~ consistent whole hippocampal segmentations using only 9 atlases, or fewer, with various hippocampal definitions, disease populations, and image acquisition types. Additionally, we show that MAGeT-Brain identifies hippocampal subfields in standard 3T T1-weighted images with overlap scores comparable to competing methods.

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; Jeneson and Squire, 2012; Wixted and Squire, 2011; 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~~ reliable and faithful 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 image 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~~ by the 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 error 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~~ reliability (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~~ favourable segmentations of the hippocampus. Also, several methods have been explored for label fusion. For example, the STAPLE algorithm (Simultaneous Truth And Performance Level Estimation; Warfield et al. (2004)) uses an expectation-maximization framework to compute a probabilistic segmentation from a set of competing regmentations, or the work of Coupé et al. (2012) who show that a subset of segmentations can be estimated using metrics, such as the sum of squared differences in the regions of interest to be segmented.

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~~ reliable 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 segmentation of hippocampus. Additionally, in this paper we rigorously explore the effects of using multiple input atlases, of varying the size of the template library constructed, and registration and label fusion methods. As a result, we aim to demonstrate that it is indeed possible to reliably apply the segmentation represented in a very small set of segmented input atlases to an unlabelled target image set.

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 ADNI1 baseline dataset, achieving a mean Dice score of 0.85 against semi-automated segmentations.

Also of interest to this manuscript are the methods that attempt to define hippocampal subfields using standard T1- or T2-weighted data, of which there are few. Van Leemput et al. (2009) demonstrate that the applicability of hippocampal subfield segmentation in T1-weighted images by Bayesian techniques using Markov random field shape priors learned from 10 manual segmentations. This work, available as part of the FreeSurfer package, is limited as the segmentation omits the tail of the hippocampus and the protocol has yet to be fully validated. Yushkevich et al. (2009) manually segment hippocampal subfields on high-resolution (either $0.2mm$-isotropic or $0.2mm \times 0.3mm \times 0.2mm$ resolution voxels) T2-weighted MR images acquired from five post-mortem medial temporal lobe samples. Then, using nonlinear registration guided by shape-based models of the subfield segmentations ~~,~~and manually derived hippocampus masks of the target images, the authors demonstrate accurate parcellation of hippocampal subfields, with respect to manual segmentations, in clinical 3T T1-weighted MRI volumes. Using multi-atlas with bias correction techniques, Yushkevich et al. (2010) demonstrate a semi-automated method of subfield segmentation on in vivo focal T2-weighted MR acquisitions of the temporal lobe. Manual input is only needed to mark divisions between the head, body and tail of the hippocampus on target images.

In this paper we describe a thorough validation of the MAGeT-Brain algorithm for the fully automatic segmentation of the hippocampus and ~~its subfields~~a proof-of-concept validation of its application to the segmentation of hippocampal subfields in standard T1-weighted images. First, we address the very idea of ~~boostrapping~~ generating a template library from a limited number of input atlases (Chakravarty et al., 2013) for whole hippocampus segmentation by conducting a multi-fold validation experiment over a range of atlas and template library sizes, registration and label fusion methods. This type of validation is done

first on a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with manual segmentations Pruessner-protocol, and then replicated on a first episode psychosis patient dataset to determine the behaviour of MAGeT-Brain when segmenting younger and differently diseased subjects. Next, we compare MAGeT-Brain with other popular segmentation algorithms (FreeSurfer, FSL FIRST, MAPER, and SNT) on all the images available in the ADNI1:Complete 1Yr 1.5T sample. Lastly, using the optimal parameter settings for MAGeT-Brain found from the previous experiments, we investigate hippocampal subfield segmentation by conducting a leave-one-out validation using the Winterburn et al. (2013) manually segmented high-resolution MR atlases.

2 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.

The MAGeT-Brain (**M**ultiple **A**utomatically **G**enerated **T**emplates) ~~bootstraps the creation of~~ algorithm creates a large template library given a ~~limited~~ much smaller sized input atlas library ~~,~~ and then uses ~~the~~ this template library in basic multi-atlas segmentation ~~. Images for~~ to segment a set of input target images. The images used in the template library are selected from ~~a set of input target~~ the input 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. Effectively, basic multi-atlas segmentation is then conducted using the template library to segment the entire set of target images (including the target images used in the construction of the template library). Since each template library image has multiple labels (one from each atlas), the final number of labels to be fused for each target may be quite large (i.e. # of 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`.

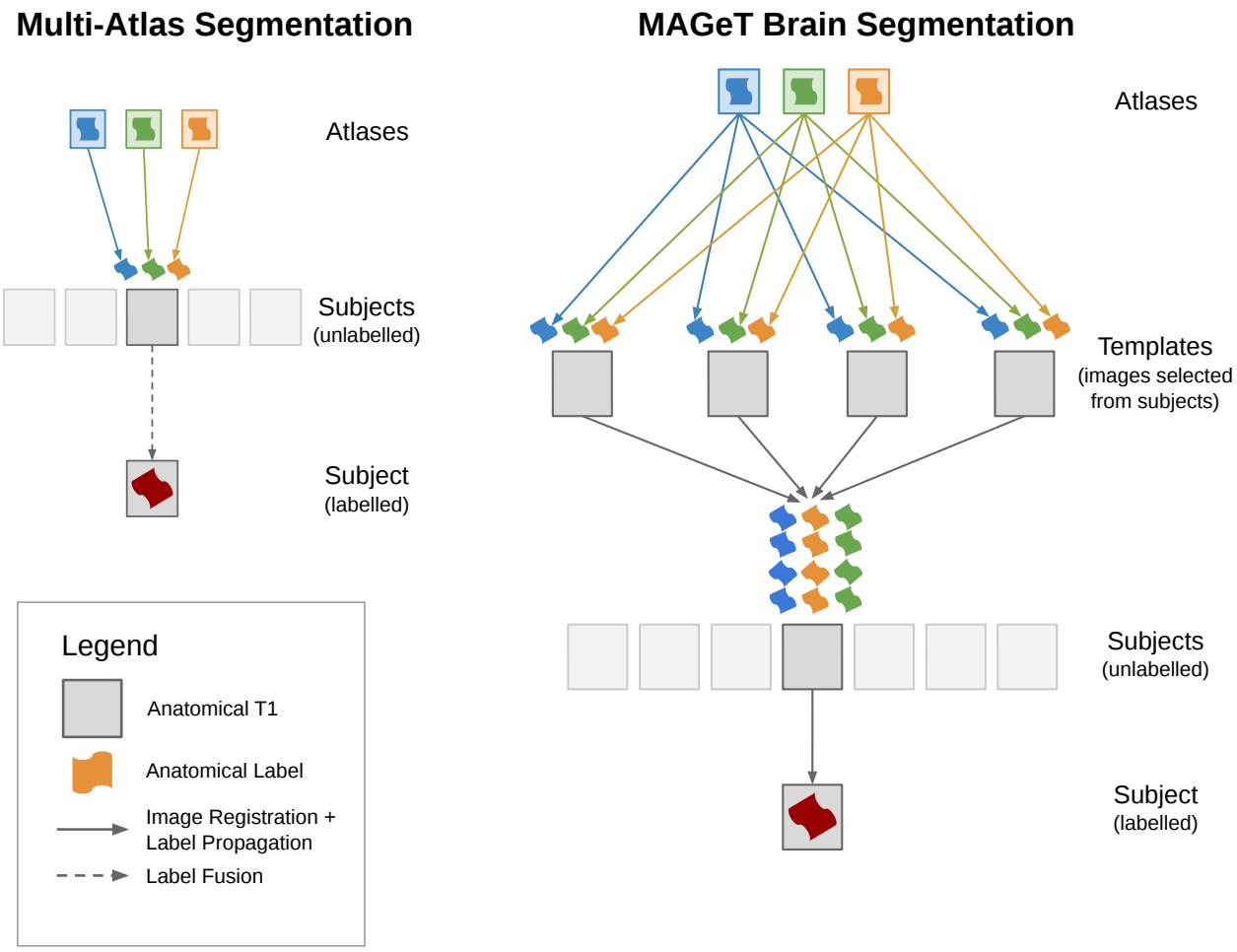


Figure 1: A schematic illustration of basic multi-atlas segmentation and MAGeT-Brain segmentation. In multi-atlas segmentation, manual labels from atlas images are warped (propagated) into subject space by applying the transformations estimated from nonlinear image registration. The resulting candidate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

3 Experiments

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

- Experiment 1 investigates MAGeT-Brain whole hippocampus segmentation of aging and Alzheimer's diseased subjects over a wide range of parameter settings using a Monte Carlo cross-validation design. The results of this experiment enable us to choose the parameter settings offering the best performance for use in subsequent experiments.

- Experiment 2 is a similar cross-validation to explore MAGeT-Brain segmentations on the brain images of young, first episode psychosis patients. In addition, MAGeT-Brain segmentations with two different atlas segmentation protocols are compared to automated segmentations by the FSL FIRST and FreeSurfer algorithms. The results of this experiment combined with the previous experiment establishes parameter settings that do not overfit to the neuroanatomical features of a specific patient cohort.

- Experiment 3 bridges MAGeT-Brain with the existing segmentation literature by comparing MAGeT-Brain whole hippocampus segmentations with those of several well-known automated methods (FreeSurfer, FSL FIRST, MAPER, SNT) on the entire ADNI1:Complete 1Yr 1.5T image dataset consisting of 246 brain images of subjects diagnosed as cognitively normal, having mild cognitive impairment, or Alzheimer's disease.

- Experiment 4 assesses hippocampal subfield segmentation quality in a leave-one-out cross-validation on the five high-resolution manually segmented Winterburn MR atlases (Winterburn et al., 2013).

3.1 Experiment 1: Whole Hippocampus Segmentation Cross-Validation — Alzheimer's Disease

In this experiment we explore the very idea of ~~bootstrapping~~ generating a template library for multi-atlas-based segmentation from a small number of input atlases. To do so, we conduct repeated cross-validations of MAGeT-Brain whilst varying the composition and sizes of the atlas and template libraries used, as well as varying the registration algorithm and label fusion method. The dataset used in this experiment is images from the ADNI dataset (Jack et al., 2008) along with whole hippocampus labels manually segmented following the Pruessner-protocol (Pruessner et al., 2000).

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

3.1.1 Experiment 1: Materials and Methods

ADNI1:Complete 1Yr 1.5T dataset Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other

Table 1: **ADNI1 cross-validation subset demographics.** CN - Cognitively Normal. LMCI - Late-onset Mild Cognitive Impairment. AD - Alzheimer's Disease. 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 Years	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.

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.

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

Clinical, demographic and pre-processed T1-weighted MRI were downloaded by the authors from the ADNI database (adni.loni.usc.edu) between March 2012 and August 2012. The image dataset used was the ADNI1:Complete 1Yr 1.5T standardized dataset available from 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 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

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

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

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 nonlinear image registration methods:

Automatic Normalization and Image Matching and Anatomical Labeling (ANIMAL) The ANIMAL algorithm carries out image registration in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that 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 deformations are then input to subsequent steps where the fit is refined by estimating smaller deformations on data blurred with a Gaussian kernel with a smaller full width at half maximum (FWHM). The final transformation is a set of local translations defined on a bed of equally spaced nodes that were estimated through the optimization of the correlation coefficient. For the purposes of this work we used the regularization parameters optimized in Robbins et al. (2004), displayed in Table 2.

Automatic Normalization Tools (ANTS) ANTS is a diffeomorphic registration algorithm which provides great flexibility over the choice of transformation model, objective function, and the consistency of the final transformation (Avants et al., 2008). The transformation is estimated in a hierarchical fashion where the MRI data is subsampled, allowing large deformations to be estimated and successively refined at later hierarchical stages (where the data is subsampled to a finer grid). The deformation field and the objective

function are regularized with a Gaussian kernel at each level of the hierarchy. The ANTS algorithm is freely available `http://www.picsl.upenn.edu/ANTS/`. We used an implementation of the ANTS algorithm compatible with the MINC data format, mincANTS `https://github.com/vfonov/mincANTS`.

We used the following command line when running ANTS:

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

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

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

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

Cross-correlation Weighted Majority Vote An optimal combination of targets from the template library has previously been shown to improve segmentation accuracy with respect to manual segmentations (Aljabar et al., 2009; Collins and Pruessner, 2010). In this method, each template library image is ranked in similarity to each unlabelled image by the 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 agreement, and we use it as the basis of comparison in this experiment.

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

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

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.

3.1.2 Experiment 1: Results

We find that for MAGeT-Brain segmentations, similarity score increases as atlas and template library size is increased, although with diminishing returns and an eventual trend towards a plateau (Figure 2a). For instance, with 9 atlases and using ANTS for registration and majority vote fusion, the mean DSC scores for 1, 5, 9 and 17 templates are 0.845, 0.865, 0.867, 0.869, respectively. A maximum similarity score of 0.869 is found when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion.

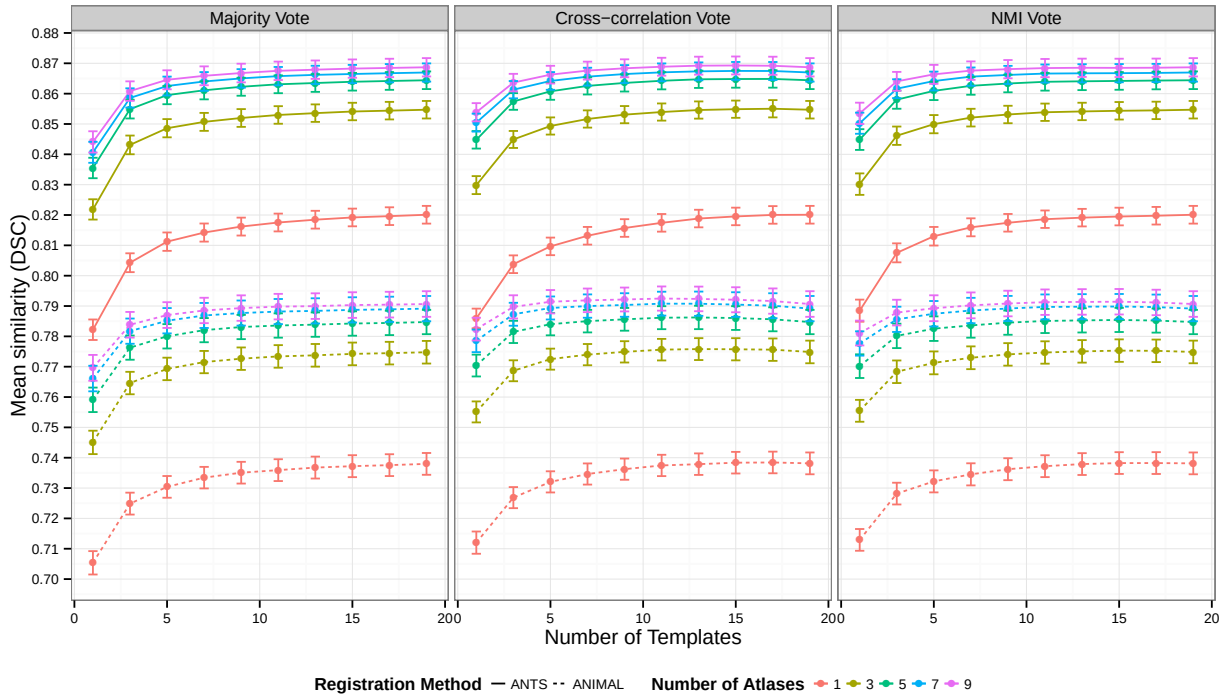
The ANTS registration method consistently outperforms ANIMAL registration over all variable settings we tested (mean increase in DSC is 0.079). Pearson correlations of MAGeT-Brain DSC scores when using weighted voting and when using non-weighted majority vote label fusion (with ANTS registration) for all combinations of atlases and templates are $r > 0.899$, p¡0.001, with a mean difference in DSC score of 0.002. This result suggests that using a weighted voting strategy does not significantly improve MAGeT-Brain segmentation agreement, contrary to the findings of Aljabar et al. (2009) for basic multi-atlas segmentation. Thus, in the remainder of our experiments only results using the ANTS registration algorithm and majority vote fusion will be shown.

With at least five templates, MAGeT-Brain consistently shows a higher DSC score than multi-atlas segmentation wth the same number of atlases: $r = 0.94$, $p < 0.001$, mean DSC increase = 0.008 (Figure 2b). The magnitude of DSC increase grows with template library size but shows diminishing returns with larger atlas libraries. Peak increase (+0.025 DSC) is found with a single atlas and template library of 19 images.

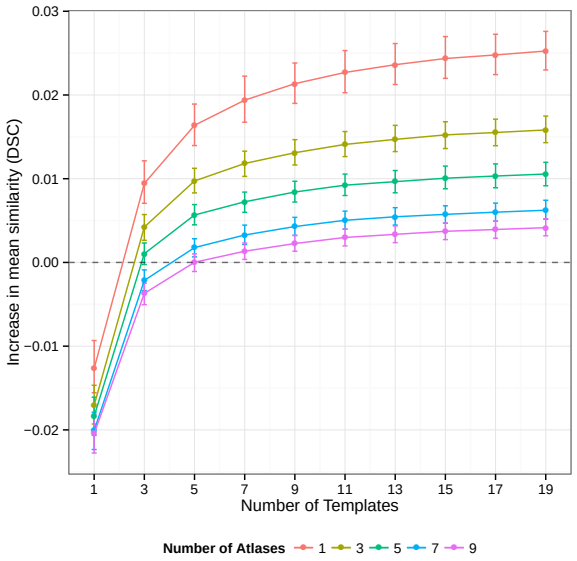
In addition to a mean increase in similarity score over multi-atlas-based segmentation, MAGeT-Brain also shows more consistency in similarity scores across all subjects and validation folds (Figure 2c). A template library of at least 13 images is sufficient to show significant ($p < 0.05$) decrease in variance for all sizes of atlas library tested (1-9 images).

We find similar behaviour with respect to optimal parameter settings and increased consistency of MAGeT-Brain segmentations in the replication of this experiment (Experiment 5, Supplementary Materials) where a different hippocampal definition is used (SNT labels available with the ADNI datasets). This strongly suggests that these results are independent of the segmentation protocol used and are, instead, features of the MAGeT-Brain algorithm.

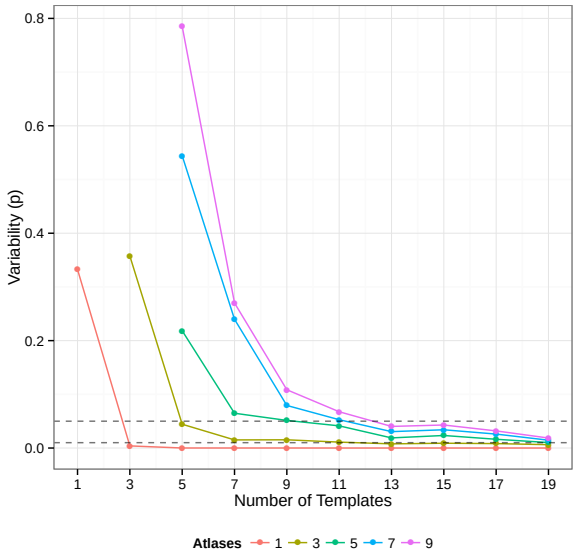
We have omitted results obtained when using an even number of atlases or templates since with these configurations we found significantly decreased performance. We believe this results from an inherent bias in the majority vote fusion method used (see Discussion).



(a) DSC vs. atlas and template library size



(b) Increase in similarity score over multi-atlas



(c) Difference in variability with multi-atlas

Figure 2: **Whole hippocampus segmentation cross-validation on ADNI subjects with Pruessner-protocol manual segmentations.** (2a) Average DSC score of MAGeT-Brain with manual segmentations for 60 ADNI subjects taken over 10 folds of cross-validation at each parameter setting. Error bars indicate standard error. (2b) Increase in DSC of MAGeT-Brain over multi-atlas segmentations. (2c) shows the significance of t-tests comparing the variability in DSC scores of MAGeT-Brain and multi-atlas across validation folds. Only points where MAGeT-Brain mean variability is lower than multi-atlas are shown. Dashed lines indicate p-values of 0.05 and 0.01.

3.2 Experiment 2: Whole Hippocampus Segmentation Cross-Validation — First Episode of Psychosis

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

3.2.1 Experiment 2: Materials and Methods

First Episode Psychosis (FEP) Dataset All patients were recruited and treated through the Prevention and Early Intervention Program for Psychoses (PEPP-Montreal), a specialized early intervention service at the Douglas Mental Health University Institute in Montreal, Canada. People aged 14 to 35 years from the local catchment area suffering from either affective or non-affective psychosis who had not taken antipsychotic medication for more than one month with an IQ above 70 were consecutively admitted as either in- or out-patients. Of those treated at PEPP, only patients aged 18 to 30 years with no previous history of neurological disease or head trauma causing loss of consciousness were eligible for the neuroimaging study; only those suffering from schizophrenia spectrum disorders were considered for this analysis. For complete 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 3.

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

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.

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

In addition to the MCCV, we segment the entire first episode psychosis dataset using MAGeT-Brain using two different atlases, as well as two popular automated segmentation packages, FSL FIRST and FreeSurfer. Specifically, MAGeT-Brain is run once with the five Winterburn atlas images and labels as atlases and a

Table 3: **First Episode Psychosis Subject 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.

randomly selected subset of 19 target images as templates. MAGeT-Brain is run a second time using the same template images, but we using five additional first episode psychosis subjects and corresponding manual segmentations (not included above) as atlases. FSL FIRST and FreeSurfer are run with the default settings: FSL FIRST `run_first_all` script was used according to the FIRST user guide [4], and FreeSurfer was run with the command `recon-all -all`.

Evaluation method Manual and automated segmentations are directly compared using Dice's similarity coefficent (DSC). In the MCCV, the per-subject DSC value is computed as the average value over the five rounds of validation for a given atlas and template library size. The reported average DSC value per given atlas and template library size is the average DSC value over all subjects segmented.

The Pruessner segmentation protocol differs slightly from the Winterburn protocol, and those used by FreeSurfer and FSL FIRST, in the inclusion of neuroanatomical features and the manner they are delineated (see Winterburn et al. (2013), and Table 9 in the Discussion below). This variation in protocol poses a problem if an overlap measure is used for evaluation: 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 assess the degree of (Pearson) correlation in average bilateral hippocampal volume produced by each method. Additionally, we evaluate the volume-related fixed and proportional biases in all segmentation methods using Bland-Altman plots (Bland and Altman, 1986).

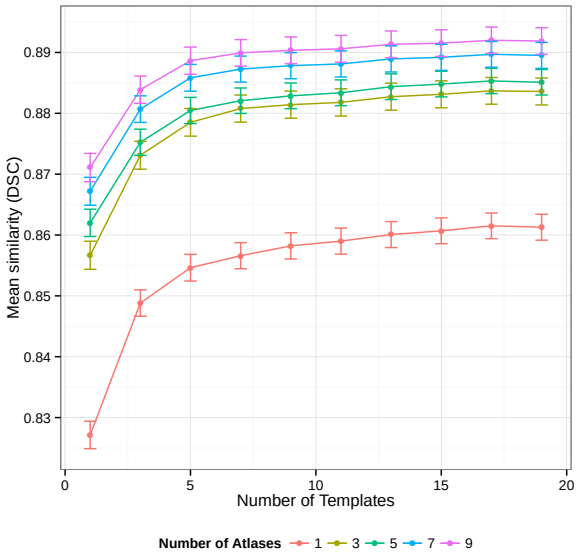
3.2.2 Experiment 2: Results

As in Experiment 1, we find that similarity score increases with a greater number of atlases or templates but quickly plateaus (Figure 3a). A maximum similarity score of 0.892 is found when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion.

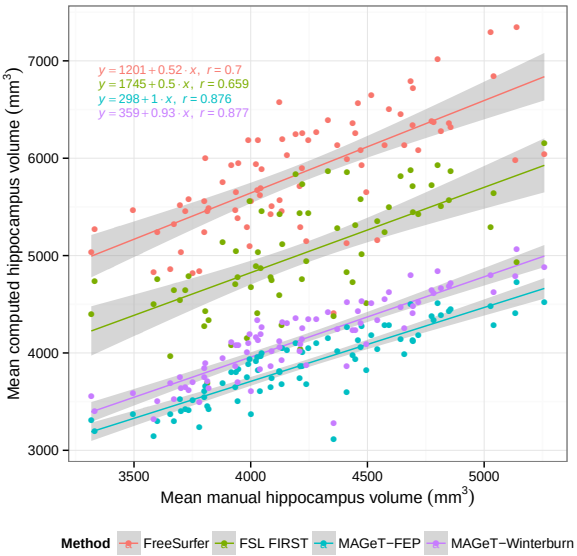
We found a close relationship in average hippocampal volume between the manual label volumes and MAGeT-Brain when using the Winterburn atlases, or manually segmented FEP subjects as atlases (Figure 3b). Both sets of volumes are correlated with Pearson $r > 0.88$. FreeSurfer and FSL FIRST volumes are both correlated with manual volumes at Pearson $r > 0.7$.

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

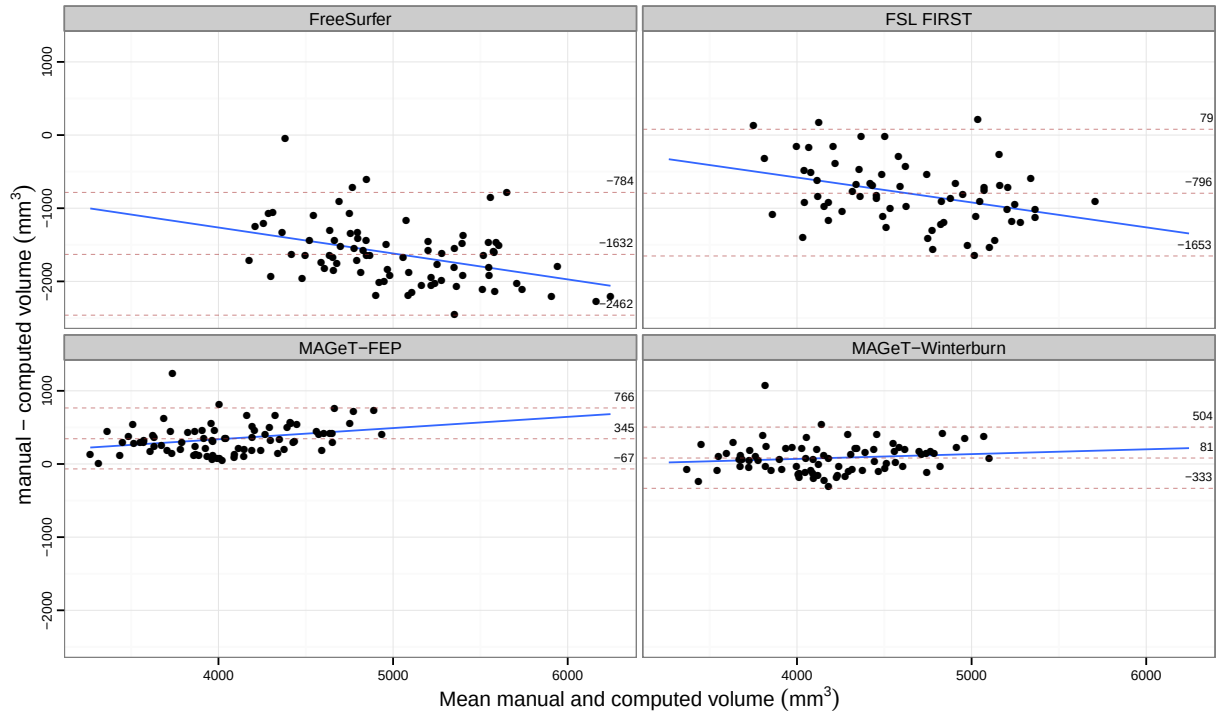
As Bland and Altman (1986) noted, high correlation amongst measures of the same quantity does not necessarily imply agreement (as correlation can be driven by a large range in true values, for instance). Figure 3c shows Bland-Altman plots illustrating the level of agreement of each method with manual volumes. All methods show an obvious proportional bias: FreeSurfer and FSL FIRST markedly underestimate smaller hippocampi and over-estimate large hippocampi (the limits of agreement are between $-2482mm^3$ and $-784mm^3$, and between $-1653mm^3$ and $79mm^3$, respectively), whereas both MAGeT-Brain methods show a much less exaggerated, but conservative bias (limits of agreement between $-67mm^3$ and $766mm^3$ when using FEP atlases, and between $-333mm^3$ and $504mm^3$ when using Winterburn atlases). On average, FreeSurfer and FSL FIRST overestimate hippocampal volume by about $1600mm^3$ and $800mm^3$, respectively. In contrast, on average MAGeT-Brain underestimates volumes by about $300mm^3$ when using FEP atlases and by about $80mm^3$ when using Winterburn atlases (compared to the Pruessner-protocol manual segmentations).



(a) Dice's similarity score vs. atlas and template library size



(b) Computed vs. manual hippocampus volume



(c) Bland-Altman plots of computed vs. manual hippocampus volume

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

3.3 Experiment 3: Whole Hippocampus Segmentation Comparison — ADNI1 Complete 1Yr

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

3.3.1 Experiment 3: Materials and Methods

ADNI1:Complete 1Yr 1.5T dataset The *ADNI1:Complete 1Yr 1.5T* standardized dataset contains 1919 images in total. SNT, MAPER, and FreeSurfer hippocampal volumes for a subset of images were provided by ADNI, along with quality control data for each FreeSurfer segmentation (guidelines described in (Hartig et al., 2010)). See Section 3.1.1 for study details, inclusion criteria and imaging characteristics.

For a subset of the ADNI images, semi-automated segmentations of the left and right whole hippocampi generated using the SNT tool from Medtronic Surgical Navigation Technologies, Louisville, CO (see Supplementary Materials for detailed discussion of the segmentation process) are made available (Hsu et al., 2002). These labels are used as the reference labels in several other studies of (semi-)automated segmentation methods (see Discussion). In addition, ADNI also distributes hippocampal segmentations and volumes determined using MAPER (Heckemann et al., 2011), a multi-atlas segmentation tool, and the FreeSurfer tool (including quality control data, with guidelines described in Hartig et al. (2010)).

Experiment details MAGeT-Brain was configured with an atlas library composed of the five Winterburn atlas images (Experiment 2, section 3.2) and segmentations. A template library of 19 images were randomly selected from the target dataset of ADNI subjects, and ANTS registration and majority vote label fusion were used as these were found to perform favourably in earlier experiments.

FSL FIRST segmentation was performed using the `run_first_all` script according to the FIRST user guide [5]. All images in the ADNI1:Complete 1Yr 1.5T dataset were segmented by both methods.

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

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

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

Table 4: **ADNI1 1.5T Complete 1Yr dataset demographics.** CN - Cognitively Normal. LMCI - Late-onset Mild Cognitive Impairment. AD - Alzheimer's Disease. 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.

Table 5: **Number of segmented images and quality control failures of ADNI1:Complete 1Yr 1.5T dataset by method.**label

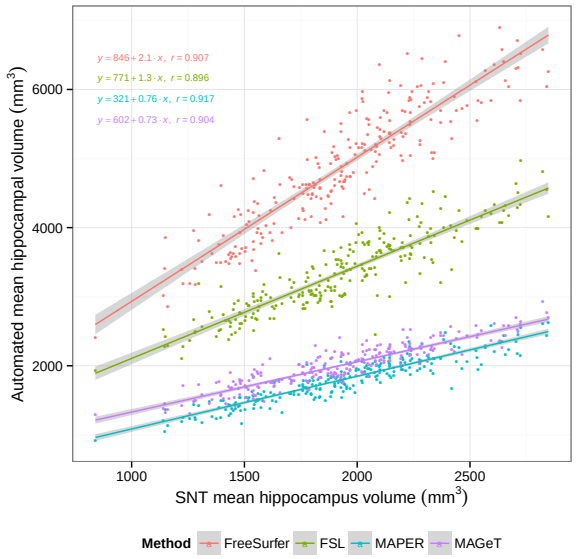
X	SNT	MAGeT	MAPER	FSL	FS
Images	368	368	368	368	368
Failures	n/a	30	n/a	20	88

3.3.2 Experiment 3: Results

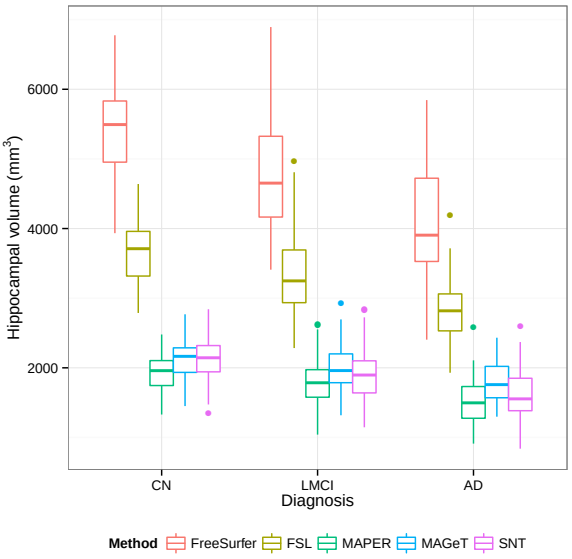
We found a close relationship in total bilateral hippocampal volume between all methods and the SNT semi-automated label volumes (Figure 4a). Volumes are well correlated ($r > 0.78$)c for all methods, and across disease categories. Within disease categories (Figure 4b), MAGeT-Brain is consistently well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

Bland-Altman plots illustrate the level of agreement of each method with SNT segmentation hippocampal volumes (Figure 4c). All methods show an obvious proportional bias: FreeSurfer and FSL FIRST markedly under-estimate smaller hippocampi and over-estimate large hippocampi, whereas MAPER and MAGeT-Brain show a reverse, conservative bias (Figure 4c). 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.

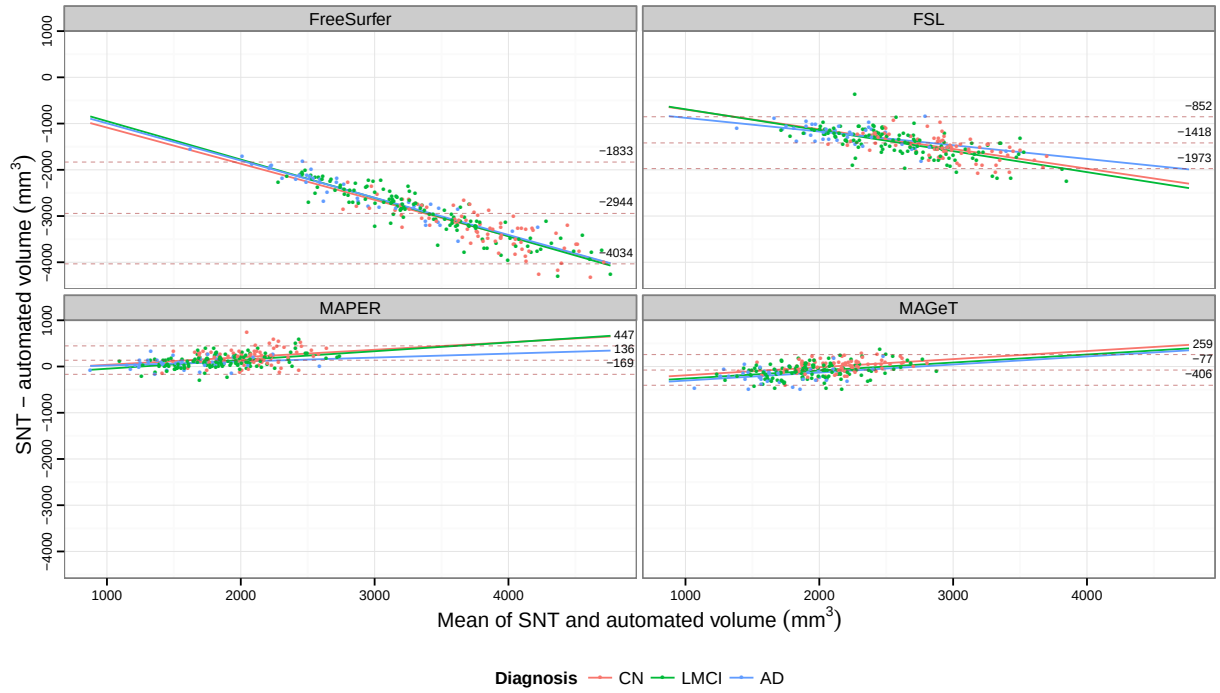
Figure 5 shows a qualitative comparison of MAGeT-Brain and SNT hippocampal segmentations for 10 randomly selected subjects in each disease category, and illustrates some of the common errors found during visual inspection. Mostly frequently, we found MAGeT-Brain improperly includes the vestigial hippocampal sulcus and, although not anatomically incorrect, MAGeT-Brain under-estimates the hippocampal body in comparison to the SNT segmentation.



(a) Computed vs. semi-automated (SNT) segmentation volume



(b) Hippocampal volume by diagnosis group and segmentation method



(c) Bland-Altman plots of computed vs. SNT hippocampus volume

Figure 4: **ADNI1:Complete 1Yr 1.5T dataset segmentation.** (4a) Subject mean hippocampal volume as measured by each of the four automated methods (FreeSurfer (FS), FSL FIRST, MAPER, MAGeT-Brain) versus the semi-automated SNT segmentation volumes. Linear fit lines and Pearson correlations with SNT labels are shown for each method. (4b) Mean hippocampal volume by method and disease category. AD = Alzheimer's disease, LMCI = late-onset mild cognitive impairment, and CN = cognitively normal. (4c) Bland-Altman plots shows the agreement between computed and SNT hippocampus volume. The overall mean difference in volume, and limits of agreement ($\pm 1.96SD$) are shown by dashed horizontal lines. Linear fit lines are shown for each diagnosis group. Note, points below the mean difference indicate overestimation of the volume with respect to the SNT volume, and vice versa.

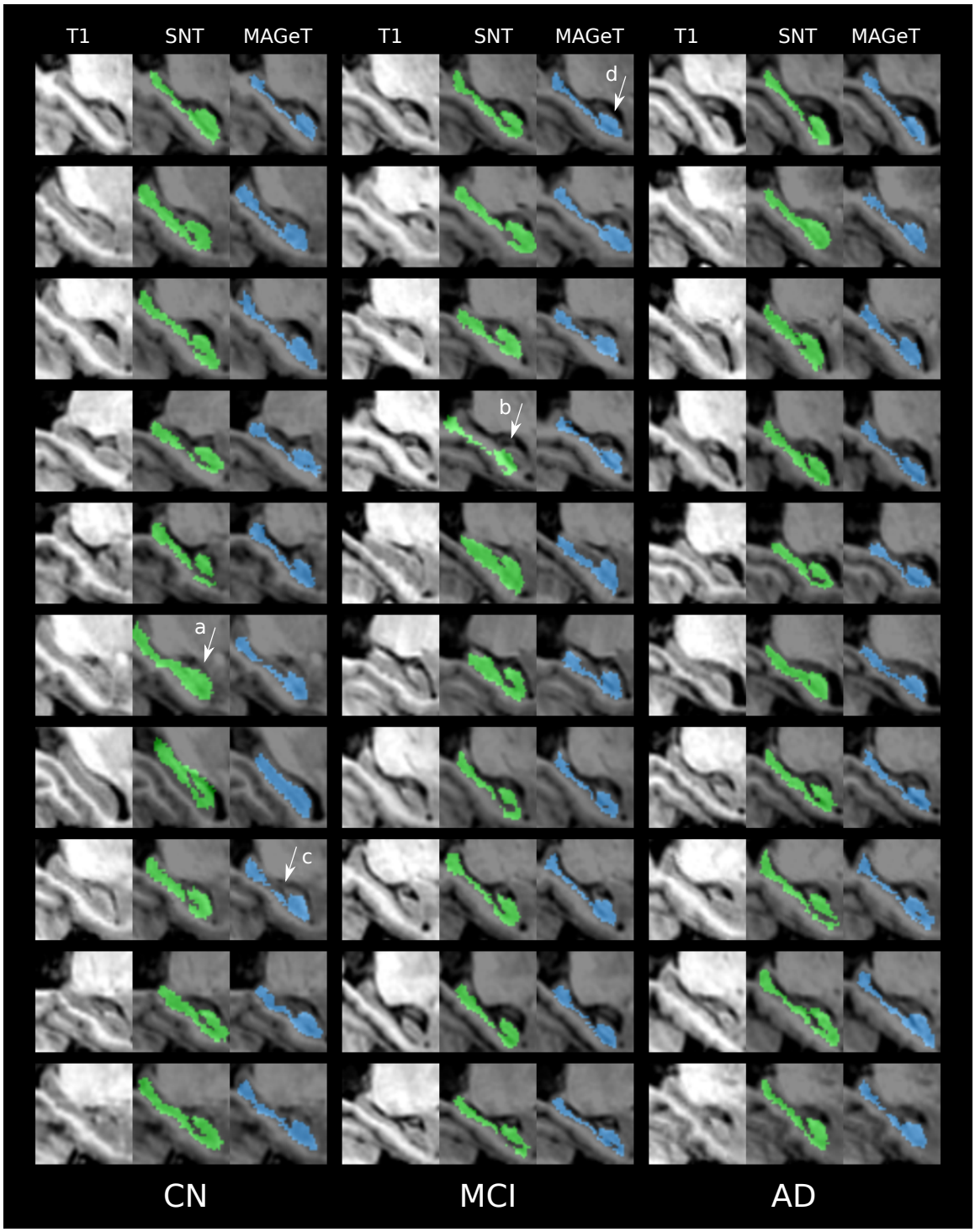


Figure 5: SNT and MAGeT-Brain segmentations for 30 ADNI subjects — 10 subjects randomly selected from each disease category in the subject pool used in Experiment 1 (Section 3.1. Sagittal slices are shown for each unlabelled T1-weighted anatomical image. SNT labels appear in green, and MAGeT-Brain labels 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.

3.4 Experiment 4: Hippocampal Subfield Segmentation Cross-Validation

The previous experiment assesses MAGeT-Brain performance on whole hippocampus segmentation. In this experiment, we ~~evaluate~~ conduct a proof-of-concept evaluatation of MAGeT-Brain hippocampal subfield segmentation of standard 3T T1-weighted images at $0.9mm$-isotropic voxels. We use a modified leave-one-out cross-validation (LOOCV) design.

3.4.1 Experiment 4: Materials and Methods

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

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

In this experiment, the Winterburn atlases (Experiment 2, section 3.2) are resampled to $0.9mm$-isotropic voxel resolution to simulate standard 3T T1-weighted resolution images. Image subsampling is performed using trilinear subsampling techniques. In each round of LOOCV, a single atlas image is selected and treated as a target image to be segmented by MAGeT-Brain. So as to have an odd-sized atlas library, atlas image is segmented once using each possible triple of atlas images, and corresponding manual segmentations, from the remaining four unselected atlases. Thus, for each of the five atlases, a total of $\binom{3}{4} = 4$ segmentations are evaluated, resulting in a combined total of $5 \times 4 = 20$ segmentations evaluated overall. We chose an atlas library with an odd number of images so as to ensure unbiased label fusion when using majority voting (see Discussion).

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

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

Table 6: **Demographics for the hippocampal subfield cross-validation healthy control subject sample used in the template library (excluding the Winterburn atlas subjects).** Education is shown in years.

	N	Control $N = 14$
Age	14	34.5 53.0 62.0
Sex : Male	14	43% (6)
Education : 12	13	15% (2)
13		8% (1)
14		23% (3)
16		15% (2)
18		38% (5)
Handedness : R	14	93% (13)

$_a$ 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.

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

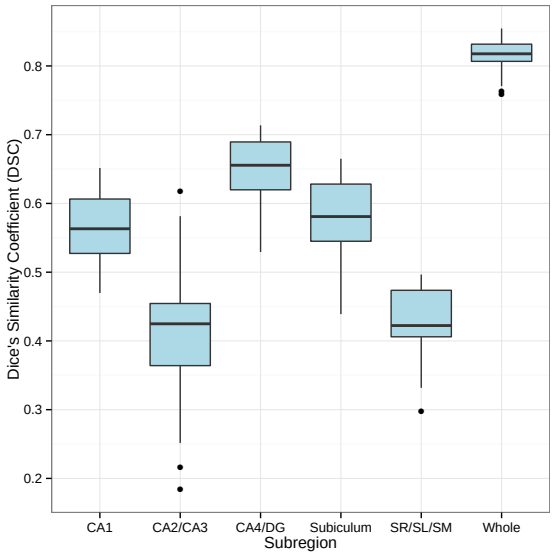
Additionally, by shifting the original manual $0.3mm$-isotropic voxel segmentations by one voxel in the x, y, and z direction and then resampling it to $0.9mm$-isotropic voxels we obtain a simulated manual segmentation having a small amount of error. We can compare the DSC overlap score of the shifted labels (relative to the directly resampled labels) with the DSC score of the MAGeT-Brain generated labels in order to evaluate their relevance.

3.4.2 Experiment 4: Results

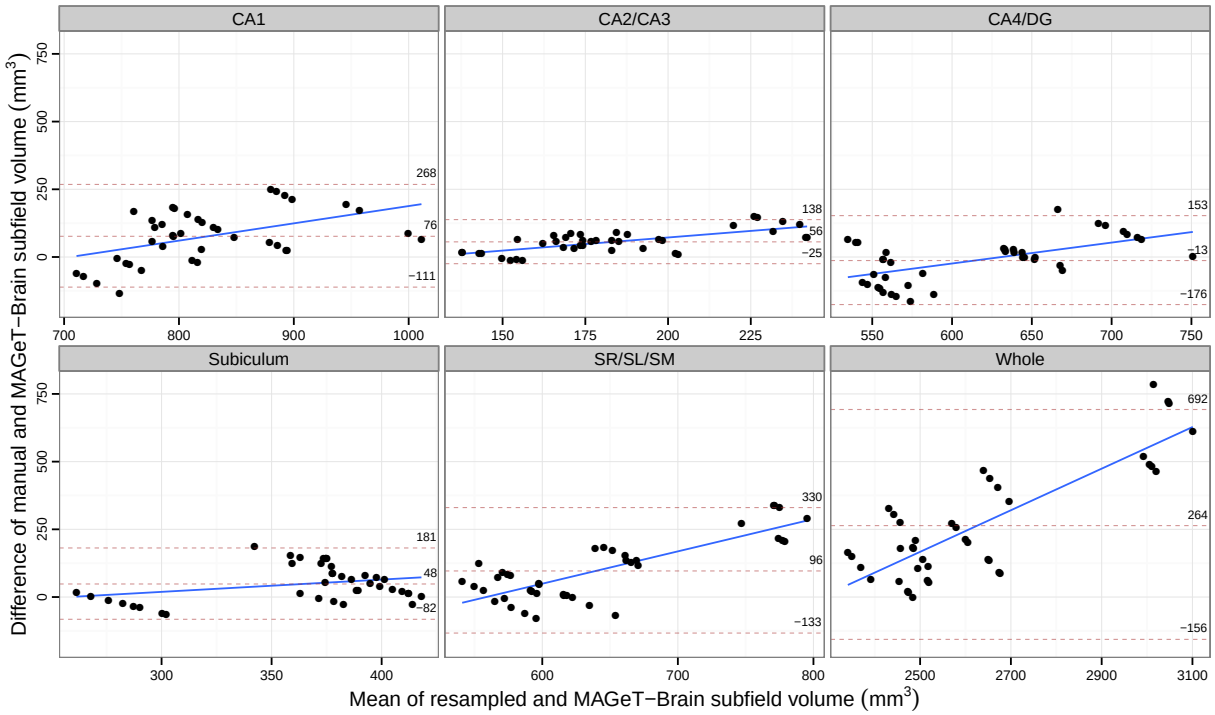
Figure 6a shows the overlap similarity scores between the MAGeT-Brain segmentations and the resampled Winterburn atlases for each hippocampal subfield across all subjects and folds of the validation. Mean and standard deviation DSC scores of the subfields are shown in Table 7, along with DSC scores for the resampled atlas segmentations when perturbed slightly and compared to the originals. We find that the CA4/DG subfield shows the highest mean DSC score of 0.647±0.051, followed by the Subiculum and CA1 subfields having scores of 0.563±0.046 and 0.58±0.057, respectively. Both the CA4/DG and molecular regions score below 0.5. These scores may seem low but not when taken in context and compared to existing (semi-)automated methods (see Discussion). The whole hippocampus is segmented with a mean DSC score of0.816±0.023.

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

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



(a) DSC score by subfield



(b) Bland-Altman plots of computed vs. manual subfield volumes

Figure 6: **Hippocampal subfield cross-validation.** (6a) Similarity of MAGeT-Brain segmentation of subfields and the resampled Winterburn atlas segmentations at $0.9mm^3$ voxel resolution, over all validaiton folds. Overlap score for each hemisphere is measured separately. (6b) shows the agreement, by subfield, of computed and manual volumes across all validation folds. The overall mean difference in volume, and limits of agreement ($\pm 1.96SD$) are shown by dashed horizontal lines. Linear fit lines are shown. Note, points below the mean difference indication overestimation of the volume with respect to the resampled volume, and vice versa.

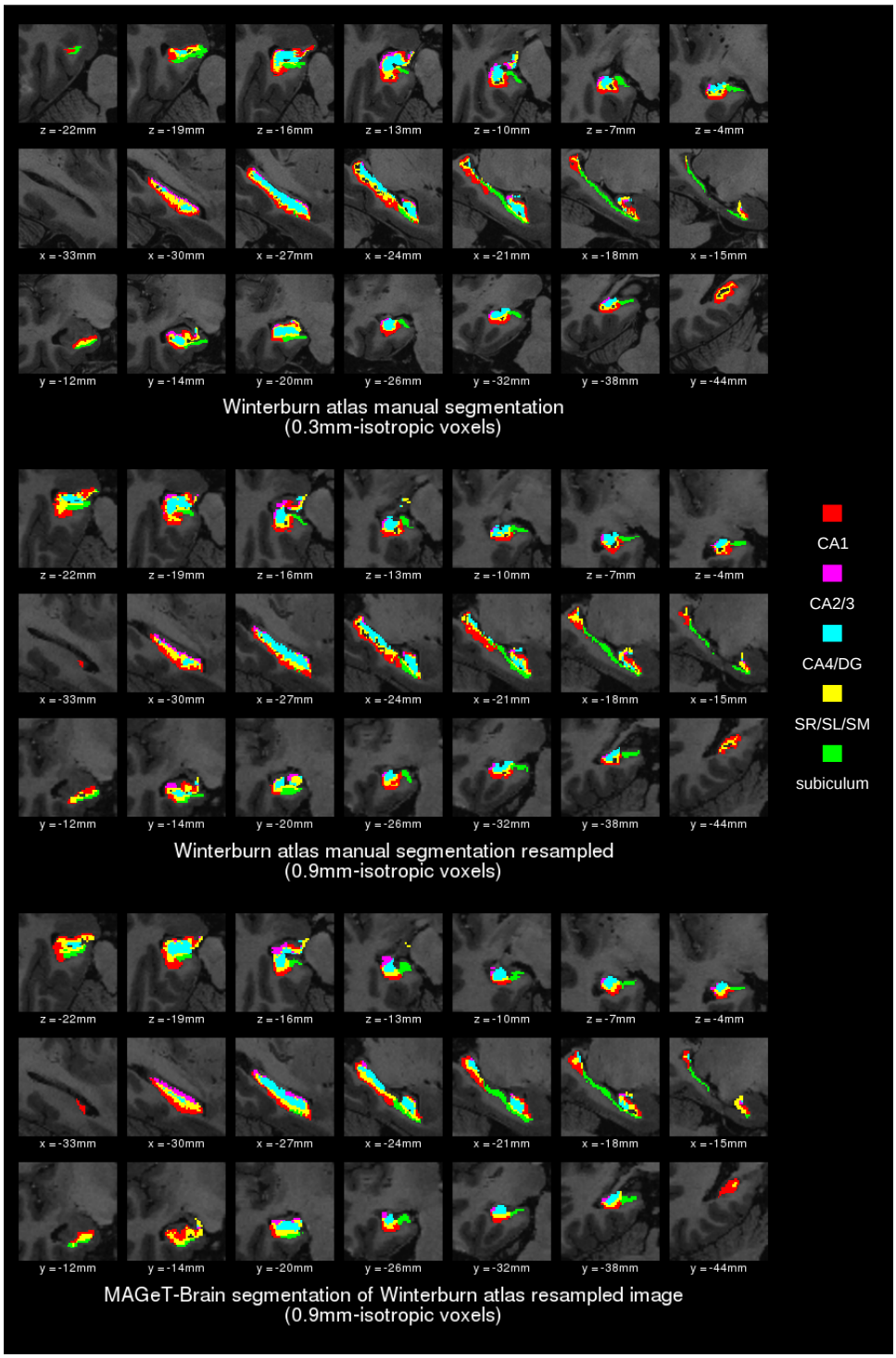


Figure 7: **Detailed subfield segmentation results for a single subject.** In the upper left corner is the original high-resolution Winterburn atlas manual subfield segmentation; in the upper right corner is the Winterburn atlas segmentation subsampled from 0.3mm- to 0.9mm-isotropic voxels; in the lower left corner is the MAGeT-Brain segmentation of the subsampled Winterburn atlas image from a single fold of the cross-validation. In each segmentation, slices from the left hemisphere are shown in Talairach-like ICBM152 space: the first row shows axial slices from inferior to superior; the second row shows sagittal slices from lateral to medial; the third row shows coronal slices from anterior to posterior.

Table 7: Overlap similarity results for the each of the subfields of the hippocampus. Simulated overlap similarity results are also given for manual labels that were translated by one voxel (i.e.: $0.3mm$) in all directions and then resampled. Values are given as mean Dice's Similarity Coefficient (DSC) $\pm$ standard deviation.

Subfield	MAGeT	0.9mm translation
CA1	0.56 ± 0.05	0.27 ± 0.03
CA2/CA3	0.41 ± 0.10	0.12 ± 0.05
CA4/DG	0.65 ± 0.05	0.42 ± 0.05
SR/SL/SM	0.43 ± 0.05	0.19 ± 0.04
Subiculum	0.58 ± 0.06	0.14 ± 0.04

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

The core claim the MAGeT-Brain method is based on – that ~~we can meaningfully bootstrap a template library~~ a useful template library can be generated from a small set of labelled atlas images – is validated in the ~~cross validation~~ cross-validation conducted in Experiment 1 (and the replication in Experiment 2 and Experiment 5, Supplementary Materials). We find that both increasing the number of atlases and the number of templates used improves MAGeT-Brain segmentation over and above basic-multi-atlas segmentations using the same number of atlas images. That is, by taking the extra step of generating a template library using target images, MAGeT-Brain is able to improve the overlap between the automatically generated segmentations and manually generated "gold standard" segmentations. The magnitude of this improvement is greatest with a small number of atlases, but even with larger atlas libraries we have found that generating a template library reduces the variability in segmentation precision (i.e. MAGeT-Brain more consistently produces high quality segmentations than does basic-multi-atlas segmentation over repeated randomized trials). These effects do not appear dependant on the hippocampal segmentation protocol used.

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

Table 8: ~~Automated segmentation accuracy of the Hippocampus~~ **Summary of automated segmentation methods of the Hippocampus.** This table ~~shows the best~~ summarizes published Dice's overlap measure between automated and ~~ground truth~~ manual segmentations of the hippocampus. Unless otherwise specified, validation datasets are composed equally of cases and control subjects, and use manual segmentation labels as ground truth in computing DSC scores. AD = Alzheimer's Disease; MCI = Mild Cognitive Impairment; CN = Cognitively Normal (CN); FEP = First Episode of Psychosis; LOOCV = Leave-one-out cross-validation; MCCV = Monte Carlo cross-validation; SNT = Surgical Medtronic Navigation Technologies semi-automated labels. Some studies of automated segmentation of ADNI images are excluded because they do not not provide overlap measures for the hippocampus (Heckemann et al., 2011; Chupin et al., 2009).

Method	Atlases	DSC mean (AD; MCI; CN)	Reference	Validation	Dataset (Truth)
MAGeT-Brain	9	0.841		10-fold MCCV on 69 subjects	ADNI (SNT)
Patch-based label fusion	16	0.861 (0.838; —; 0.883)	Coupe et al. (2011)	LOOCV	ADNI (SNT)
Multi-atlas	20	0.848 (—; 0.798, 0.898)	Wang et al. (2011)	10-fold MCCV on 20 of 139 subjects	ADNI (SNT)
ACM (AdaBoost-based)	21	0.862	Morra et al. (2008)	LOOCV	ADNI (SNT)
LEAP	30	0.848	Wolz et al. (2010)	Segmentation of 182 subjects	ADNI (SNT)
Multi-atlas	30	0.885	Lötjönen et al. (2010)	Segmentation of 60 subjects	ADNI (SNT)
Multi-atlas (MAPS)	55	0.890	Leung et al. (2010)	Segmentation of 30 subjects (10 AD, MCI, and CN)	ADNI (SNT)
MAGeT-Brain	9	0.869		10-fold MCCV on 60 subjects	ADNI (Pruessner)
MAGeT-Brain	9	0.892		5-fold MCCV on 81 subjects	FEP subjects
Neural nets	10	0.740	Powell et al. (2008)	Segmentation of 5 subjects	controls
Probablistic atlas	11	0.852	van der Lijn et al. (2008)	11 atlases used in 100 rounds of LOOCV on 20 elderly subjects	elderly controls
Probabilistic Atlas	16	0.860	Chupin et al. (2009)	LOOCV	AD subjects
Anatomically-guided EM	17	0.812	Pohl et al. (2007)	LOOCV on 17 controls, segmentation of 33 mixed subjects	mixed diagnosis
Multi-atlas	30	0.820	Heckemann et al. (2006a)	LOOCV	controls
Multi-atlas	30	0.880	Gousias et al. (2008)	30 adult atlas used, segmentation of 33 2yr old subjects	2yr old controls
Multi-atlas	80	0.890	Collins and Pruessner (2010)	LOOCV	controls
Multi-atlas	55	0.860	Barnes et al. (2008)	LOOCV	controls and AD
Multi-atlas	275	0.835	Aljabar et al. (2009)	LOOCV	controls

not see a significant indication of this effect in the MAGeT-Brain segmentations. Selectively filtering out atlases with lower image similarity is believed to reduce sources of error from estimating deformations via nonlinear registration, partial volume effects from nearest neighbour image resampling, and neuroanatomical mismatch between atlases and subjects. That MAGeT-Brain does not see the same boost in performance from weighted voting may suggest that the neuroanatomical variability of a template library 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. However, it is also important to note that even in the absence of a more sophisticated label fusion algorithm, MAGeT Brain performs reasonably well in comparison to other groups that have tested new segmentation algorithm with Alzhiemer disease, mild cognitive impairment, and cognitively normal data from the ADNI database (Table 8. In addition, our validation in Experiment 2 (with the first episode psychosis subjects) yields DSC's that are amongst the highest reported. Thus, more work is required to determine the extent to which label fusion will improve the ~~accuracy~~ reliability of our algorithm.

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

Despite MAGeT-Brain achieving segmentation results which are competitive with the rest of the field (Table 8), a concern may be raised over the modest improvement in segmentation agreement observed using MAGeT-Brain over multi-atlas, with the same number of atlases (Experiment 1). As we have shown in that same experiment, the benefit in using MAGeT-Brain is both an increase in the overlap agreement and also in the improved consistency of the labelling regardless of atlas or template choice. Reducing the variability in segmentation agreement is an important consideration that few have touched on previously. In addition, the Monte Carlo cross-validations that we present in Experiment 1 and Experiment 2 are amongst the most stringent performed in the multi-atlas segmentation literature. To the best of our knowledge, with the exception of (Wang et al., 2011), other groups do at most a single round of leave-one-out-validation (Table 8). Thus, 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.