Elsevier Editorial System(tm) for NeuroImage
Manuscript Draft

Manuscript Number: NIMG-13-2171R1

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

Article Type: Regular Article

Section/Category: Methods & Modelling

Corresponding Author: Mr. Jonathan Pipitone, MSc

Corresponding Author's Institution: Centre for Addiction and Mental Health

First Author: Jonathan Pipitone, MSc

Order of Authors: Jonathan Pipitone, MSc; Min Tae M Park, BSc; Julie Winterburn, BSc; Tristram A Lett, BSc; Jason P Lerch, PhD; Jens C Pruessner, PhD; Martin Lepage, PhD; Aristotle N Voineskos, PhD, MD; Mallar M Chakravarty, PhD

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 (Multiple Automatically Generated Templates), 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 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 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 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: 80mm3 , 1600mm3 , and 800mm3 , 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 250mm3 across the entire dataset, compared with FreeSurfer and FSL FIRST which both overesimate volume differences by 2600mm3 and 2800mm3 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 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.

camh
Centre for Addiction and Mental Health

February 18, 2014

Editorial Board, Neuroimage

Dear Sir or Madam:

Please find enclosed our manuscript entitled, *Bootstrapping Multi-atlas Segmentation Using Multiple Automatically Generated Templates for the Segmentation of the Whole Hippocampus and Subfields* by Jon Pipitone and colleagues.

In this manuscript we describe a novel automated MRI hippocampal segmentation algorithm optimised to perform well with only a small number of manually segmented training images. We believe this is an important contribution because the expertise and effort needed to perform manual segmentation can be prohibitive for many clinicians and researchers, and yet existing automated methods generally require between 30 to 80 training segmentations. This situation makes it infeasible to use these methods for segmentations based on histological-based digital segmentations (because of their rarity), high-resolution digital atlases (because of the time needed to segment these images), or when exploring new protocols or the effect of variations on a segmentation protocol.

It is for these reasons that we developed the automated segmentation algorithm, mischievously named MAGeT-Brain, which takes advantage of the neuroanatomical variability that exists in the target population being studied to bootstrap a large template library from a small set of manually segmented images. In this manuscript we rigorously validate this approach on multiple disease populations, and compare our segmentations with existing popular methods (e.g. FSL and FreeSurfer). We find that MAGeT-Brain produces very reliable and consistent segmentations of the whole hippocampus when compared with manual segmentations, and is competitive with exiting methods. Finally, we have made our algorithm available publically online for use by other groups, and are pursuing the contribution of our segmentations to the Alzheimer's Disease Neuroimaging Initiative image database.

We believe that the technique we have developed and our findings are a significant contribution to the neuroimaging community, specifically for those

researchers interested in large scale studies of the hippocampus in the context of normal brain function and different forms of brain dysfunction.

We hope you find the enclosed manuscript meets the high standards of NeuroImage.

Sincerely,

Jon Pipitone
Kimel Family Translational
Imaging-Genetics Laboratory
Research Imaging Centre
Centre for Addiction and Mental
Health
Toronto, Ontario

encl: Manuscript

## Highlights

- We propose an automated MR image hippocampus (and subfield) segmentation method
- Our method is optimised for use with a small number (< 10) of training images
- Consistent, accurate identification of the whole hippocampus and subfields
- Validated on healthy, Alzheimer's disease, and first episode schizophrenia subjects
- Source code and high-resolution training subfield atlases available online

We thank the reviewers for their careful review of our manuscript and their detailed and insightful comments. We have addressed their concerns below and believe that we have greatly improved the scientific quality of the our manuscript as a result.

Before addressing each reviewers' comments individually, in summary we made the following major revisions to the paper:

1. We re-ran the comprehensive cross-validation experiment (Experiment 1) using fully manually segmented whole hippocampus labels (following the Pruessner protocol) for training and validation on ADNI subjects. The original analysis using the SNT semi-automated labels has been moved to Supplementary Materials. This should address the concerns around the validity of using the semi-automated SNT labels as "ground truth" in this experiment.
2. We ran a similar cross-validation experiment on the First Episode Psychosis dataset to ensure that the parameter settings found in Experiment 1 also apply to younger patients and subjects.
3. We add hippocampal volumes of the First Episode psychosis dataset as computed by FSL FIRST and FreeSurfer to serve as comparisons for MAGeT-brain volumes when assessing agreement and biases.
4. Simplified the subfield cross validation experiment so as to maximise performance and make our results more easily comparable to Yushkevich et al. 2010. These experiments now include Dice Similarity Coefficients for the automated segmentations and bias estimates with respect to a ground truth.
5. We carefully rewrote and reworded parts of the paper to make it clear that the experiments (in particular, the subfield segmentation experiment) evaluate precision/reliability of segmentation with respect to an atlas segmentation, not accuracy with respect to ground truth (like a histological ground truth).
6. Rather than default to the standard organization of many manuscripts (Introduction, Materials and Methods, Results, Discussion), given the large amount of data that is now included, we have chosen the following reorganization to improve readability:

    - Introduction
    - Experiment 1 - 10-fold cross-validation using manual segmentations of ADNI data
    - Experiment 2 - 5-fold cross-validation using manual segmentations of first episode psychosis subjects, and comparison to other standard techniques
    - Experiment 3 - Comparison of our method against other standard methods
    - Experiment 4 - Validation of subfield segmentations
    - Discussion
    - Supplementary Materials
        i. Experiment 5 - 10-fold cross-validation using SNT segmentations of ADNI

    Note that the Sections for Experiments 1-4 include a targeted Materials and Methods and Results section so as to be self-contained. Many of the experiments proposed depend on the results of Experiments 1 and 2, and so this organization should improve the logical

flow of the manuscript. The original 10-fold cross validation using the semi-automatically generated SNT labels has now been labelled Experiment 5 and is moved to the Supplementary Materials. We have chosen to include this experiment so that others who have used these labels for evaluation (of which there are several) can directly compare their results to ours.

What follows is a point-by-point response to the reviewers' comments. The reviewers' comments are in bolded text, our responses to the comments in regular weight text, and additions to the manuscript indented and italicized.

## Reviewer #1

**The manuscript by Pipitone et al. presents a novel method for performing automatic segmentation of the hippocampus and hippocampal subfields. Using relatively few traced hippocampi, the authors derive fairly good reliability ratings for hippocampus and subfields compared to manually traced values. The authors also show (with the exception of one subfield) that automated segmentations exceed what one might expect from shuffling of voxels (an estimate, to my read, of chance performance).**

We thank the reviewer for the thorough and complimentary appraisal of our work.

**Overall, the method presented is important as it will likely save time compared to other methods with overall segmentation of the hippocampus. I was less convinced of its utility for hippocampal subfields though and to me this part of the manuscript requires some rethinking and revision. The authors may even consider taking out this part of the manuscript all together because to my read, at least, this part seemed to stand on somewhat shakier ground. Thus, the majority of my comments below are focused on segmentation of subfields with the exception of some of the minor comments toward the end. One last consideration is that the authors need to include statistical results within the Results rather than referring to figures in which statistical differences are visually evident but not numerically provided.**

We have attempted to address these comments based on the reviewer's criticisms in the newly revised version of the manuscript. More detailed responses are provided below.

**MAJOR**

**1. In many places, I noted some confusion between the ideas of validity and actual ground truth hippocampal subfield neuroanatomy. The authors have shown convincingly that the reliability values are good in the hippocampal subfields, with the exception of the CA4/DG subfield. Because the tracing method used though is based on the Winterburn et al., a method which, unlike the Yuskevich method, is not based directly on actual histological data, it seems difficult to make strong statements about**

**"ground" truth.**

We agree, and also feel this confusion between the ideas of validity and ground truth was present in the other experiments on whole hippocampal volumetry. To address this confusion we have thoroughly changed language throughout the paper so as to make clear that in the experimental evaluations we are not assessing ground truth validity but are instead assessing the reliability/precision of our segmentations with respect to the manual segmentation protocol used to segment the atlases and target images (see Minor issues below for specific instances in the paper). We have completely removed the use of the phrase “ground truth” (since, as we in the paper, and reviewer #1 points out, ground truth in hippocampal segmentation is ill defined even in histological studies).

With respect to the question of the basis for the tracing method used in Winterburn et al. we note that while the contrast differences were used to guide the manual labels in Winterburn atlases that these manual definitions are, in fact, completely informed by the Durvenoy (2005) and the Mai et al. (2008) histological atlases. Yushkevich et al. (2009) conduct a study of post-mortem acquisitions and base their manual segmentation protocol on the Duvernoy (2005) and the Fatterpekar (2002) MRI/histological atlases. The segmentation protocol used in Yushkevich et al. (2010) is based on an 4T acquisition MRI-protocol (Mueller and Wiener, 2009). Thus, the Winterburn atlases have as much grounding in histology as either Yushkevich study. We’ve added a more detailed summary of these works to the introduction in order to put our use of the Winterburn atlases in context (see Page 4, lines 122-135).

It is also worthwhile noting that since the initial submission of our manuscript, there have been recent advances from the Yushkevich group that revise their initial labelling based on serial histological acquisitions (Adler et al., 2014). While this refined reconstruction and labelling is impressive, to the best of our knowledge this work has yet to be integrated and validated in the context of a large scale fully automated pipeline. We have also added this note to the Discussion.

Also, in order to improve the rigour of our subfield segmentation validation, we have now completely revised the evaluation section for Experiment 3 (the new section for the evaluation of subfield segmentation). In this current experiment we perform a multi-fold leave-one-cross validation in order to getter a better sense of the reliability of the protocol. This now gives us 20 segmentations with which to evaluate our segmentations. The resulting Kappa distributions are now presented in Fig 5, Bland Altman plots of our segmentations are presented in Fig 6, and qualitative visualization of the subfields are given in Fig 7. We believe that this approaches the evaluation stringency that we have put in place for the other experiments (see below).

**Simply put, there is currently no consensus amongst neuroanatomists exactly how subfields such as CA4/DG differ from CA3/2, especially on lower resolution MRIs and in the hippocampal head, differentiate. In fact, many methods lump these subfields together, esp. at 3T. Even histological reconstructions depend primarily on one or two**

**(postmortem) brains and there has yet to be any evidence that individual variability in these fairly small subfields can actually accurately be identified on relatively low-resolution (.3 mm +) scans (see my later comments on this as well). Thus, it is difficult not to attribute the low reliability of CA4/DG not just to its small volume, as the authors do, but the fact that there is no consensus in the field exactly where this area is, leading to likely variability in tracings.**

In addition to the language changes mentioned above, to highlight the issue of definitions we have added the following paragraph to the conclusion of the paper (Page 30, lines 703-715):

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

To highlight the definitional ambiguities we have included a table contrasting the segmentations of several manual MR subfield segmentation protocols. (Page 29, Table 9):

Table 9: **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/dentate gyrus, subiculum, entorhinal cortex |
| Van Leemput et al. (2009) | CA1, CA2/CA3, CA4/dentate gyrus, 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 |
| Mueller et al. (2007) | CA1, CA2, CA3/CA4 & dentate gyrus, Sibiculum, entorhinal cortex |

Despite definitional disagreements, previous work from our group (Winterburn et al. 2013) show

that the subfield tracing protocol is highly reliably (Page 262, Winterburn et al. 2013):

**Table 2**
Test–retest results for the whole hippocampus and each of the sub-fields. Simulated test–retest results are also given for original labels that were translated by one voxel (i.e.: 0.3 mm in all directions), scaled by 0.99, and scaled by 1.01 (representing a 1% shrinkage or scale in all three cardinal directions). Values are given as mean Kappa (range).

| Region | Manual | 0.3 mm translation | 0.99 scaling | 1.01 scaling |
|---|---|---|---|---|
| CA1 | 0.78 (0.77–0.79) | 0.66 (0.65–0.68) | 0.64 (0.59–0.71) | 0.67 (0.60–0.76) |
| CA2/CA3 | 0.64 (0.56–0.73) | 0.54 (0.47–0.64) | 0.35 (0.23–0.46) | 0.43 (0.34–0.54) |
| CA4/dentate gyrus | 0.83 (0.81–0.85) | 0.70 (0.65–0.75) | 0.68 (0.6–0.74) | 0.67 (0.58–0.72) |
| SR/SL/SM | 0.71 (0.68–0.73) | 0.69 (0.66–0.72) | 0.74 (0.69–0.79) | 0.72 (0.67–0.8) |
| Subiculum | 0.75 (0.72–0.78) | 0.60 (0.52–0.66) | 0.59 (0.42–0.67) | 0.60 (0.41–0.68) |
| Whole hippocampus | 0.91 (0.90–0.92) | 0.85 (0.84–0.86) | 0.87 (0.86–0.91) | 0.88 (0.86–0.91) |

**I think the authors need to be clearer about the fact that only the Yushkevich et al. method builds subfield segmentation atlases based on post-mortem brains, something that the Winterburn atlas does not do.**

This is true of Yushkevich et al. (2009) only whereas Yushkevich et al. (2010) studies manually segmented in vivo focal MR images. As mentioned above, we have included more detail about these studies in the introduction. In addition, we have also discussed the recent work from that group from Adler et al. (2014) in the Discussion to highlight the differences between our work and the Yushkevich work.

**My suggestion would be for the authors to consider other segmentation methods which are more widely accepted in the field, such as those of Mueller et al. 2007. Does this approach lead to better reliability? My guess is that it would. The methods proposed here would appear equally compatible with other tracing methods and it would be helpful to know if this methodology would also work [with] more conservative approaches or those approaches better rooted in ultra-high-resolution (post-mortem) templates.**

We thank the reviewer for raising this point and we do believe that this is actually one of the great strengths of our methodology. As we have demonstrated in the segmentation work for the whole hippocampus (Experiments 1, 2, 3, 5) we can easily accommodate other methods. However, one of big differences between our approach and the approach of others is how we have chosen to define our subfields. So rather than try to estimate volumes on T2-weighted data that is of high-resolution in the coronal plane and highly anisotropic along the anterior-posterior axis, we asked where it was possible to actually identify subfields in data that approximates that T1-weighted data that are used in standard research practices. In addition, our algorithm relies on whole brain registration. It is unclear at the moment how the registration methodologies would have to be changed to accommodate data that has voxel dimensions of 0.4mm x 0.5mm x 2mm. Further to the best of our knowledge, Mueller's work has not been made publicly available, which would require that we adapt her segmentation protocol to our data, which has higher resolution in the anterior-posterior. It is unclear at the present time how to do this. We do believe, however, that the reviewer's comment is an extremely valuable one, and we are considering how to do this in future work. To this end, we have added the following the Discussion section of our manuscript:

*Experiments 1, 2, and 5 have demonstrated that our algorithm flexibly accommodates different whole hippocampus manual segmentation methodologies. We have not explicitly evaluated a subfield definition other than the Winterburn protocol, and therefore it is possible that using an alternate subfield definition could improve the reliability of our automated subfield definitions. For example, established definitions such as those from Mueller et al. (2007) could be a prime candidate for further exploration. In addition, the conservative nature of the Mueller definition (labelling of the 5 slices in the hippocampus body only) would likely further aid in reliability measurement. However, there are two main logistical problems that we would have to overcome prior to implementation. The first is that these definitions were developed for data that is highly anisotropic (0.4mm × 0.5mm × 2mm), and it is unclear how our algorithms would deal with such atlases used as input. The second is that, since these atlases are not publicly available, we would be have to re-implement the protocol using our atlases. At the present time it is unclear how we would adapt these protocol to data that we used, where subfield segmentations are defined on 0.3mm$^3$ voxels. However, the impact of subfield definitions in the context of our work is an important one and should be considered in subsequent studies.*
(See Page 29, Lines 677-689).

**While the Yushkevich et al. method does segment the DG from CA fields, this is based on post-mortem atlases, a fact that is not discussed or mentioned in the current manuscript.**

As mentioned above, we compare and contrast subfield segmentation protocols from Yushkevich and others in Discussion, and in the Introduction we now explicitly discuss the post-mortem basis for the Yushkevich atlases (Page 4, Lines 127-132):

*Yushkevich et al. (2009) manually segment hippocampal subfields on high-resolution (either 0.2mm$^3$ isotropic or 0.2mm × 0.3mm × 0.2mm resolution voxels) T2-weighted MR images acquired 127 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 in clinical 3T T1-weighted MRI volumes.*

**2. Although the methods here are interesting and promising, I thought more comparison was needed with the Yushkevich et al. methods. Along the lines of point #1, it might be helpful to directly compare reliability for different subfields reported here with those of Yushkevich et al. I realize the authors did this to some extent but I think they could go even further. Again, given that the "ground" truth is not based on post-mortem data with this methodology (or at least is several steps removed from it), it is important to compare with methods that derive from MRIs collected on scans in which these segmentation can be made with significantly more accuracy (and subsequently validated**

**with histological staining).**

In the discussion section (Page 30, Table 10) we have included a table which includes a direct comparison of the overlap similarity scores (Dice's) for our method, Yushkevich et al. (2010) and Van Leemput et al. (2009):

Table 10: **A comparison of subfield segmentation overlap similarity with manual raters.**

| Subfield | MAGeT-Brain | Van Leemput et al. (2009) | Yushkevich et al. (2010) |
|---|---|---|---|
| CA1 | 0.563 | 0.62 | 0.875 |
| CA2/3 | 0.412 | 0.74 | $CA2 = 0.538, CA3 = 0.618$ |
| CA4/DG | 0.647 | 0.68 | $DG = 0.873$ |
| presubiculum | — | 0.68 | — |
| subiculum | 0.58 | 0.74 | 0.770 |
| hippocampal fissure | — | 0.53 | — |
| SR/SL/SM | 0.428 | — | — |
| fimbria | — | 0.51 | — |
| head | — | — | 0.902 |
| tail | — | — | 0.863 |

As mentioned previously, in addition, we have replaced our original subfield validation experiment with one that is more similar to that of Yushkevich et al, (2010). Our updated experiment is a modified leave-one-out cross-validation with the Winterburn atlas images and segmentations, described as follows:

> *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 4 segmentations are evaluated, resulting in a combined total of 5 × 4 = 20 segmentations evaluated overall. We chose an atlas library with an odd number of images so as to ensure unbiased label fusion when using majority voting (see Discussion).*
> (Experiment 4, Page 21, Lines 493-501)

MINOR

**1. Intro: "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 authors should also probably include reference to other work more specifically focusing on the hippocampus (the Scoville and Milner work involved the whole medial temporal lobe, not hippocampus exclusively). The authors might consider citing work by Squire or others in this context.**

Agreed. We have included citations for the following works:

Jeneson, A., & Squire, L. (2012). Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory*, *19*(1), 15–25. doi:10.1101/lm.024018.111

Wixted, J., & Squire, L. (2011). The medial temporal lobe and the attributes of memory. *Trends in cognitive sciences*, *15*(5), 210–217. doi:10.1016/j.tics.2011.03.005

**2. "accurate identification of the hippocampal subfields is indeed possible using this methodology."**

**Use of accurate here is misleading as there is no consensus regarding exactly where subfields are, particular in areas like the head, using MRI. The gold standard for accuracy is comparing an in vivo MRI with a histological reconstruction in which staining is possible. This current method does not involve this and the authors should probably make this issue clear. Accuracy is of course only as good as the underlying knowledge of exactly where the subfields are. Because in vivo and ex vivo scans from the same participants are currently lacking, which would represent the true "gold standard," statements about "accuracy" should be tempered and considered carefully here.**

Agreed. This sentence is now rewritten to read: "... reliable reproduction of hippocampal subfield segmentations in standard 3T T1-weighted images is possible." (Page 30, Line 719-720)

In addition, as noted earlier, we have removed mention of "gold standard" throughout the manuscript and instead refer to reliability and overlap similarity with the atlas segmentation protocol.

**3. "We then performed a leave-one-out validation to determine if hippocampal subfields can be accurately identified using our multi-atlas framework."**

**Again, accuracy is relative here. I think what the authors mean is that the algorithm produced a good match to manual tracing. This should probably be stated as otherwise, accuracy comes across as misleading.**

Agreed. This sentence in the Introduction (Page 5, Line 146-147) now omits "accuracy" and defers discussion about evaluation to section 3.4, Experiment:

> *...we investigate hippocampal subfield segmentation by conducting a leave-*
> *147 one-out validation using the Winterburn et al. (2013) manually segmented*

*high-resolution MR atlases.*

**4. The authors should probably clarify upfront (in the Intro) that atlas refers to manually segmented images based on a recent method developed by Winterburn et al. rather than something like the Duvernoy atlas (although that in turn is derived, in part, from this atlas).**

We have specifically defined the term *atlas*, as well as other commonly confused terms such as *template* and *label* in Section 2, Page 5. Specifically, on line 151, we define an atlas: ,

*We use the term atlas to mean a manually segmented image ...*

In addition, we have also been careful to include a subsection for each experiment describing the atlases and manual segmentation methods used in each.

More specifically, in the introduction to Section 3, which summarises the experiments described in our paper (Page 7), we now state (lines 191-192):

*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)***

(Emphasis not included in the manuscript)

**5. T1 resolution: 1: 25mm _ 1: 25mm _ 1: 2mm**

**This seems somewhat low to make statements about subfields. This should probably be clarified.**

We apologize if this was unclear, but the T1-weighted images used for the segmentation of the subfields were 0.9 mm isotropic voxels (resampled from the original 0.3mm isotropic voxel Winterburn atlas images). We suspect that the reviewer is referring to voxel-dimensions of the ADNI data that was used in Experiments 1 and 3. We trust the reorganization of our work has helped in keeping track of the different data that were used.

**6. What is the full voxel resolution of the EFGRE-BRAVO sequence?**

We have included this in our manuscript, page 21, line 486-488:

*... 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*.

**7. "Hippocampal MAGeT-Brain-based segmentations using both ANIMAL and ANTS**

**registration algorithms demonstrate good overlap with SNT Gold Standard 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 authors should report actual statistics here**

This paragraph now includes mean DSC scores to illustrate our claim (Experiment 1, Page 11, lines 318-322):

> *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 0.844, 0.865, 0.867, 0.868, respectively. A maximum similarity score of 0.869 is found when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion.*

**8. "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." Again, these statements are lacking any statistical data to back them up.**

We have included a summary of Pearson correlations between ANTS and ANIMAL-based results (Experiment 1, Page 11, lines 323-330):

> *The ANTS registration method consistently outperforms ANIMAL registration over all variable settings we tested (mean increase in DSC is 0.078). Pearson correlations of 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.*

**9. "With an increasing number of templates, MAGeT-Brain shows improvement over**

**multi-atlas-based segmentation in overlap accuracy when using the same number of atlases and voting method (Figure 4). The two methods converge in accuracy when using seven atlases. Peak improvement in MAGeT-Brain accuracy ( 0.02 DSC) is found when one atlas is used with a template library of 19 images."**

**Once again, some statistical tests are needed to back up these statements. Just showing errorbars on plots is not sufficient.**

We have included a Pearson correlation MAGeT and multiatlas scores and a mean increase over all atlas and template library size (Experiment 1, Page 11, Lines 331-334):

> *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.*

**10. "In general, across hippocampal subregions 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, the CA1, CA4/DG, and SR/SL/SM subregions all show a percent error in volume that is at or lower than resampling error. The Subiculum and CA2/CA3 subregions show distinctly larger percent error in volume than is found through resampling."**

**Again, statistics are needed to back up statements here.**

The updated results section for this experiment now includes (Page 22, Lines 523-531):

> *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 of 0.816±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 over-estimated 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$^3$).*

## Reviewer #2

**This paper describes the MAGeT procedure that maps the hippocampus from an adaptive set of templates based on mapping of pre-existing atlases, and establishes an optimal set of atlases and templates to be used in applications of schizophrenia and AD datasets. This work is clearly important as it creates a framework where existing atlases can be used to create templates in populations under study without the need to create manual segmentations which is time consuming and not achievable by many groups. The second contribution of this work is that it shows large hippocampal subfields can be segmented in 3T scans with acceptable errors.**

**In general, the validation process and results of experiments 1, 2 and 4 are thoroughly described. Some issues with regard to experiment 3 need clarification (below).**

**Major concerns:**

**1. Experiment 1 uses SNT segmentations in 69 scans to cross validate the MAGeT procedure. The issue with using SNT segmentation is two-fold:**

**a. SNT segmentations come from semi-automated maps of one atlas. This means that all 69 segmentations contain an inherent common characteristic from the same atlas subject's segmentation. Mapping these segmentations back onto each other may in fact create a biased result (i.e., artificially improved accuracy) than if all 69 segmentations were manually delineated.**

**b. Even though the SNT segmentations have been shown to have ICC of 0.94, they are still not the same as manual segmentations, as shown in Figure 12 some of them have inaccurate labeling.**

**Therefore, a better approach would be to use the manual segmentations that the SNT segmentations compared to, if it is possible to obtain them. Or use a dataset that has manual segmentations to start with.**

The reviewer raises an excellent point about the semi-automated nature of SNT labels potentially leading to biased MAGeT-brain segmentation results either because of inherent bias in automated procedure (leading to higher reported accuracy) or because of poor segmentation

consistency (leading to lower reported accuracy).

Therefore, we chose to recreate our original cross-validation experiment using 60 manually segmented ADNI baseline images segmented by co-author Jens Pruessner (Experiment 1). The original SNT cross-validation experiment with 69 subjects has been moved to the Supplementary Materials (Experiment 5), since comparisons to SNT segmentations are widely reported in the literature and so serve as an informal benchmark between algorithms. Note, due to the lack of manual segmentation available for the entire ADNI dataset, we chose to use the SNT labels as our reference labels for this work. We have also discussed the pitfalls of using this segmentation approach on pages 27-28, lines 613-620:

> *On that note, one author (JW), an expert manual rater (Winterburn et al., 2013), identified regular inconsistencies in the SNT segmentations: occurrences of over- and under-estimation, as well as misalignments of the entire segmentation volume (Figure 5). Although the SNT segmentations are used as benchmarks for validation in many other studies (Table 8), these segmentation inconsistencies present the possibility that a more accurate and consistent benchmark segmentation protocol ought to be used in order to truly understand the results of such validations. Indeed, our replication of the 10-fold cross-validation using SNT segmentations (Experiment 5, Supplementary Materials) shows noticeably poorer mean similarity scores for both MAGeT-Brain and multi-atlas.*

**Another option would be to use the FEP dataset of 81 scans all with manual segmentations.**

In addition to the redesign of Experiment 1, we also chose to replicate the experiment using the FEP dataset. In this revision to the FEP experiment (Experiment 2), we include a 5-fold Monte Carlo cross-validation using all 81 scans with manual segmentations (see point 2, below).

**2. For experiment 3, the optimal parameters (number of atlases, templates) for other experiments were determined on experimenting with AD (with elderly normal) data. How exactly they translate into a younger cohort is not clear. It would be beneficial to see that the same cross validation procedure on the FEP data would yield similar parameters or plots (as figures 3-5).**

As noted above, we included a cross-validation in the FEP experiment similar to the cross-validation performed with 69 ADNI subjects, but on the younger first episode psychosis dataset. To simplify things, we used ANTS for registration and majority vote fusion, but varied the number of atlases (1-9) and templates (1-19) and report the change in Dice overlap score.

**3. The FEP experiment is not described well to be included in the manuscript in its current form. Further experiments should be conducted. For example:**

**a. The evaluations should follow those used in experiment 4: gold standard, FSL,**

**MAPER.**

**b. The evaluation should also include controls as did for the ADNI data.**

In addition to the cross-validation using the FEP dataset described above, we have also expanded upon the analysis of the single MAGeT-brain segmentation of the entire FEP dataset. Specifically, we evaluate our segmentations with those produced by FSL and FreeSurfer (MAPER is unavailable as it is not distributed nor open-source). See Experiment 2 results (Section 3.2.2, Page 18), and Figure 3, page 16.

**Minor concerns:**

**1. For experiment 2, the errors obtained on subfields should be compared with those obtained based on FreeSurfer subfield labels as well.**

Since the subfield segmentation protocol differs between the method used by FreeSurfer (van Leemput et al. 2009) and that of the Winterburn atlases, we felt it would not be meaningful to directly compare segmentations of the same images. Instead, as noted above (in response to comments from Reviewer #1), in the discussion section we now include a comparison of subfield segmentation protocols (Table 9, Page 29), and a comparison of reported subfield segmentation accuracy from Van Leemput et al. 2009, Yushkevich et al. 2010, and from our subfield segmentation validation in Experiment 4.

**2. In Figure 12, the SNT is labeled as "Manual." SNT segmentation, as mentioned in the manuscript, is derived from semi-automated mapping of an atlas, not manually performed.**

Fixed. All references to the SNT segmentations now refer to these as SNT labels, and that they are semi-automated segmentations, not manual.

# 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 aggreement 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 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

*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

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 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 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 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 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 (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. 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 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 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. First, we address the very idea of boostrapping 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.

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-atlas segmentation. Images for the template library are selected from a set of input target images, either arbitrarily or so as to reflect the neuroanatomy or demographics of the target set as a whole (for instance, by sampling equally from cases and controls). The template library images are 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`.

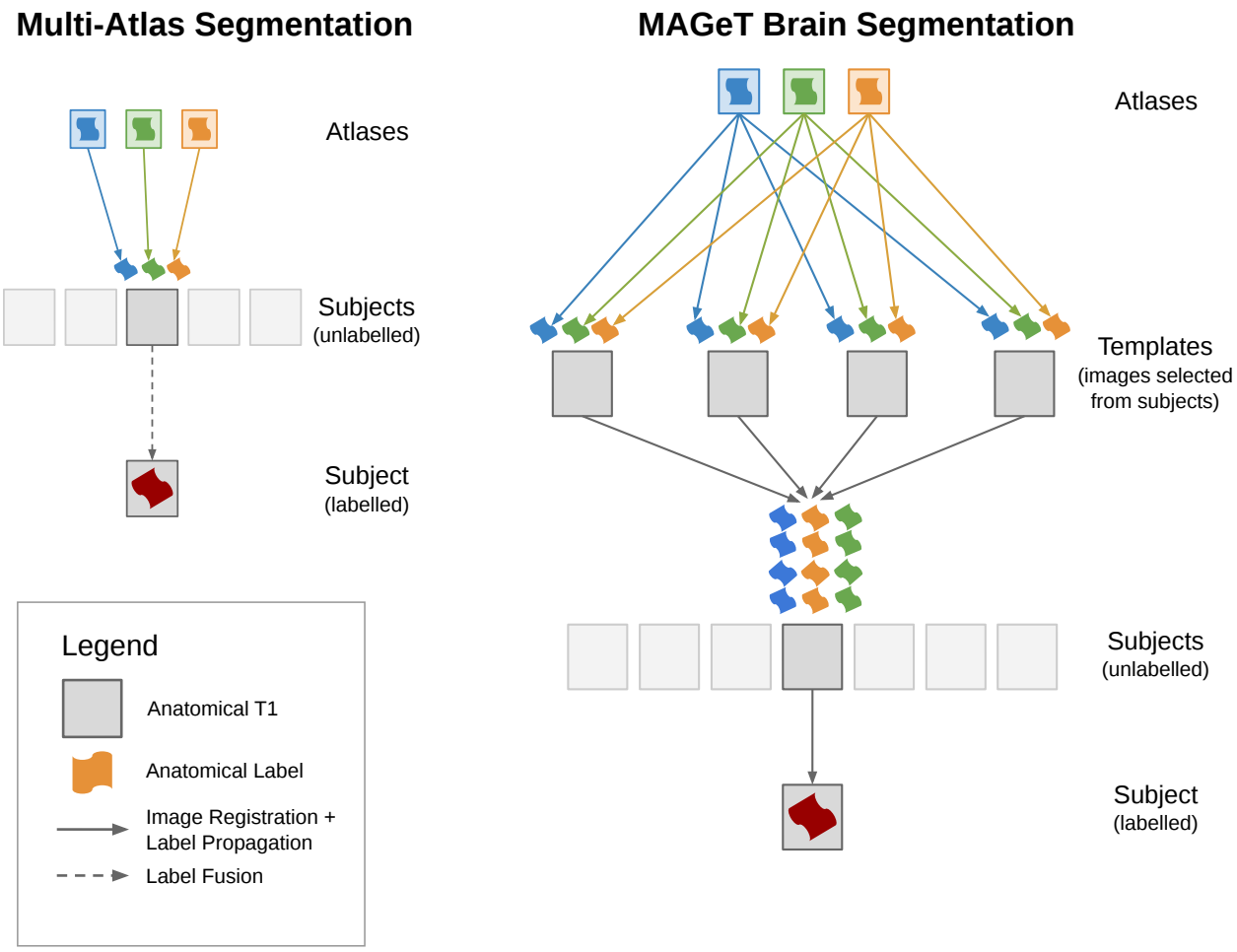



Figure 1: A schematic illustration of basic multi-atlas segmentation and MAGeT-Brain segmentation. In multi-atlas segmentation, manual labels from atlas images are warped (propagated) into subject space by applying the transformations estimated from nonlinear image registration. The resulting candidate 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 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 round, the dataset is randomly partitioned into a training set and a validation set. The method to be

[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 |

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 (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 0.844, 0.865, 0.867, 0.868, respectively. A maximum similarity score of 0.869 is found when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion.

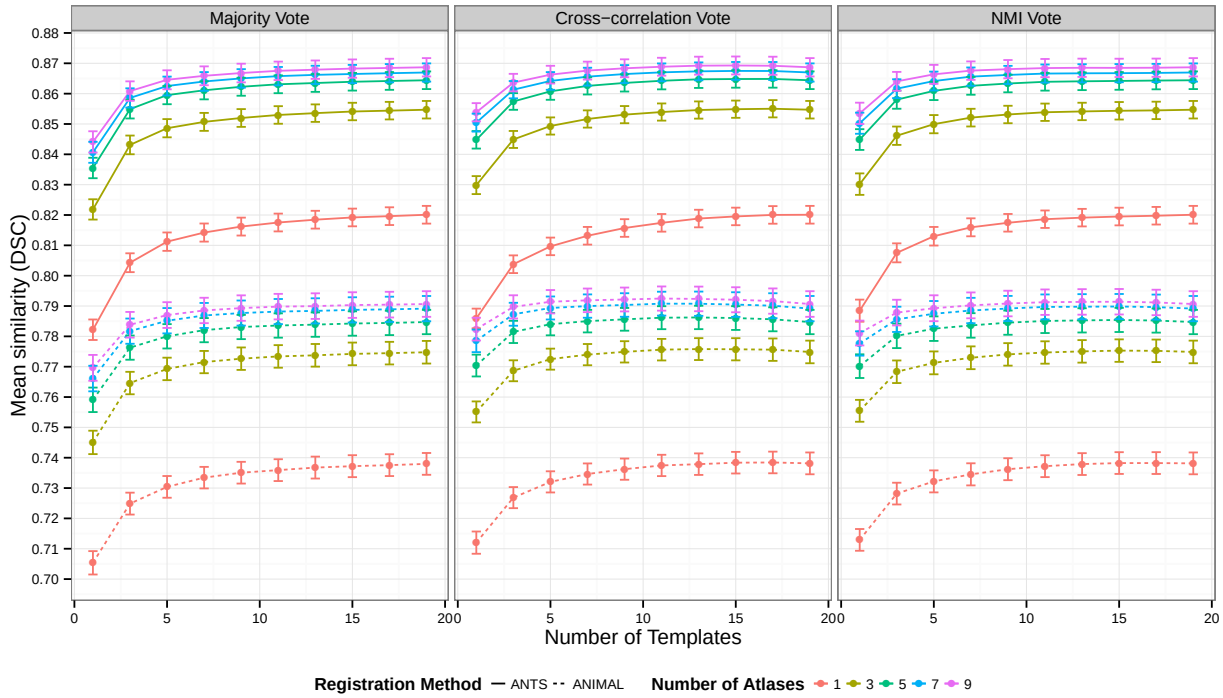
The ANTS registration method consistently outperforms ANIMAL registration over all variable settings we tested (mean increase in DSC is 0.078). Pearson correlations of MAGeT-Brain DSC scores when using weighted voting and when using non-weighted majority vote label fusion (with ANTS registration) for 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.

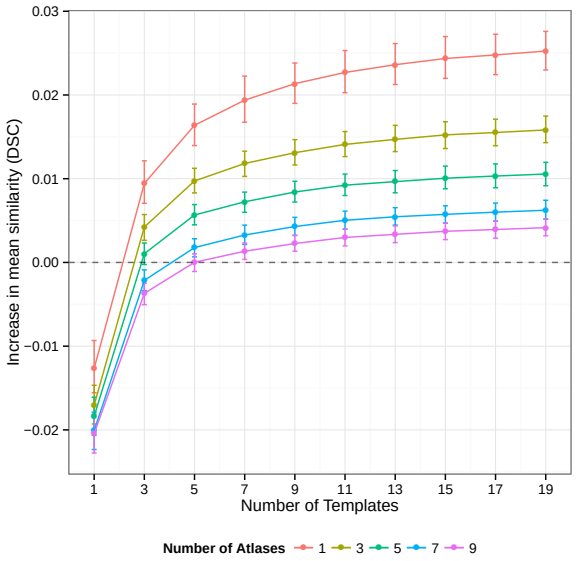
In addition to a mean increase in similarity score over multi-atlas-based segmentation, MAGeT-Brain also shows more consistency in similarity scores across all subjects and validation folds (Figure 2c). A template library of at least 13 images is sufficient to show significant ($p < 0.05$) decrease in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

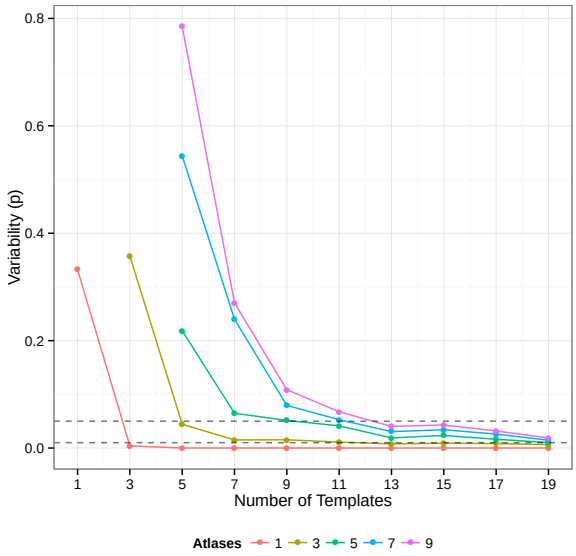
We have omitted results obtained when using an even number of atlases or templates since with these configurations we found significantly decreased performance. We believe this results from an inherent bias in the majority vote fusion method used (see Discussion).




(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).

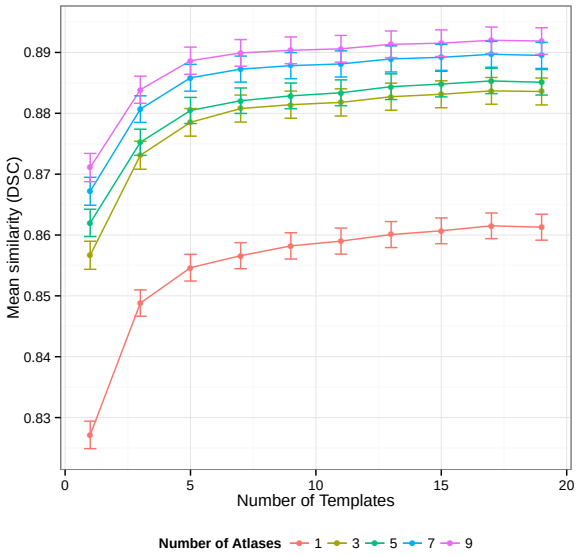
### 3.2.2 Experiment 2: Results

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.

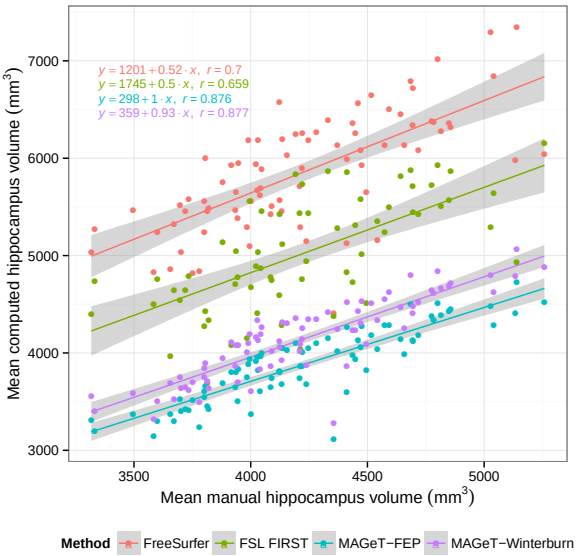
We found a close relationship in average hippocampal volume between the manual label volumes and MAGeT-Brain when using the Winterburn atlases, or manually segmented FEP subjects as atlases (Figure 3b). Both sets of volumes are correlated with Pearson $r > 0.88$. FreeSurfer and FSL FIRST volumes are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

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

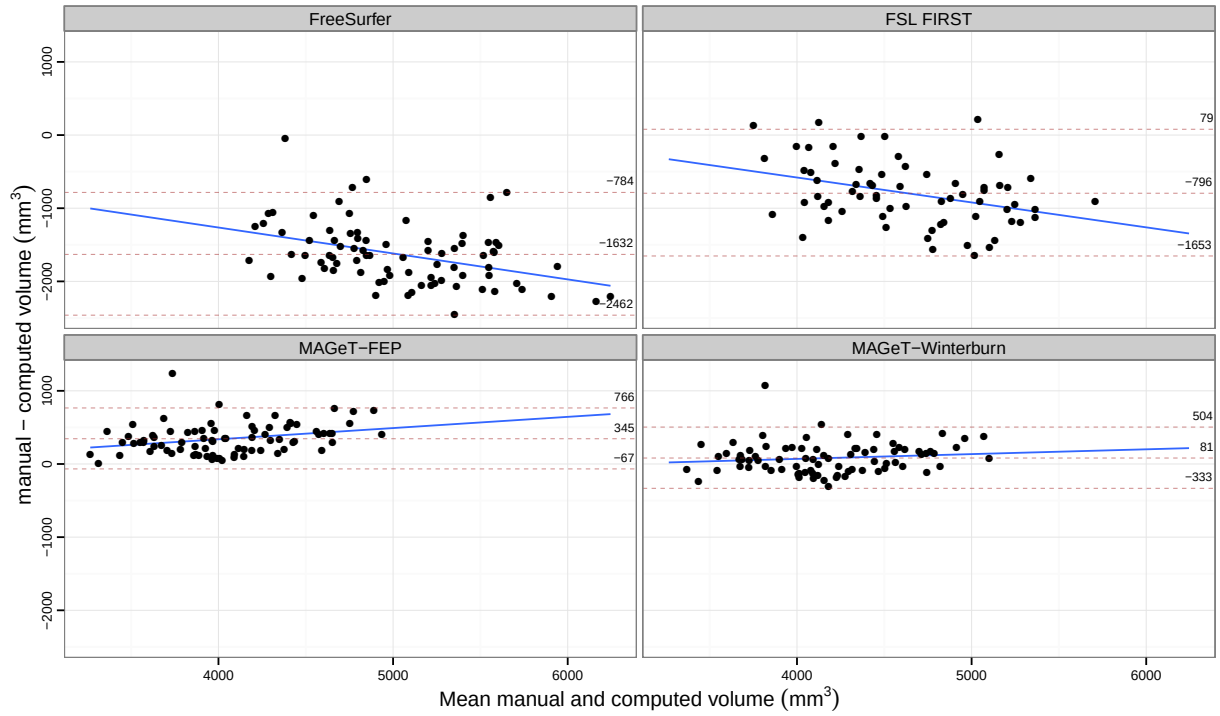
As Bland and Altman (1986) noted, high correlation amongst measures of the same quantity does not necessarily imply agreement (as correlation can be driven by a large range in true values, for instance). Figure 3c shows Bland-Altman plots illustrating the level of agreement of each method with manual 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<br>$N = 584$ | LMCI<br>$N = 931$ | AD<br>$N = 404$ | Combined<br>$N = 1919$ |
|---|---|---|---|---|---|
| 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.**

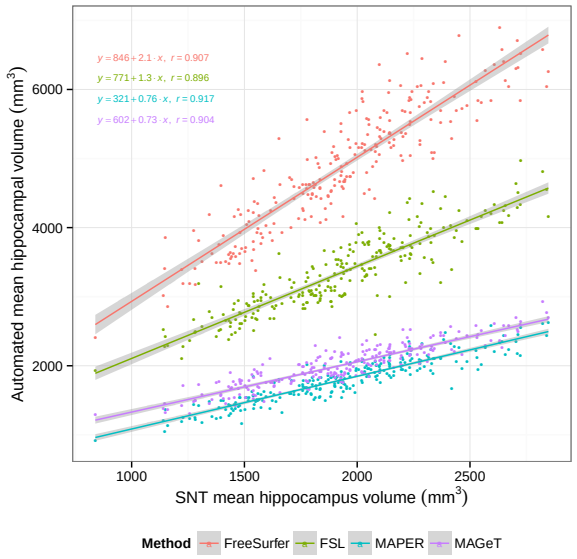
| X | SNT | MAGeT | MAPER | FSL | FS |
|---|---|---|---|---|---|
| Images | 368 | 368 | 368 | 368 | 368 |
| Failures | n/a | 30 | n/a | 20 | 88 |

### 3.3.2 Experiment 3: Results

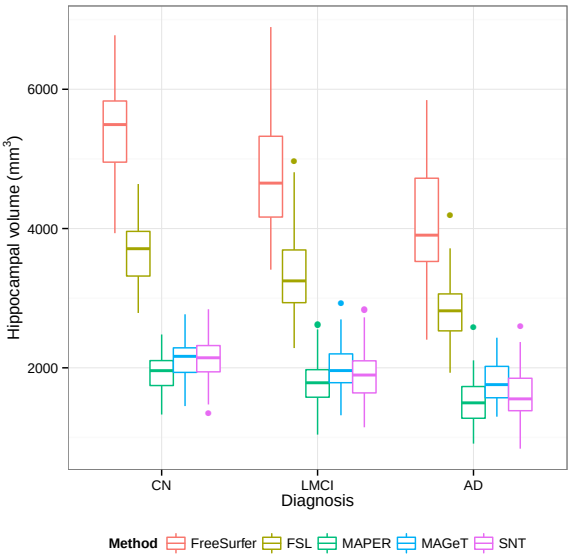
We found a close relationship in total bilateral hippocampal volume between all methods and the SNT semi-automated label volumes (Figure 4a). Volumes are well correlated ($r > 0.78$)c for all methods, and across disease categories. Within disease categories (Figure 4b), MAGeT-Brain is consistently well 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.

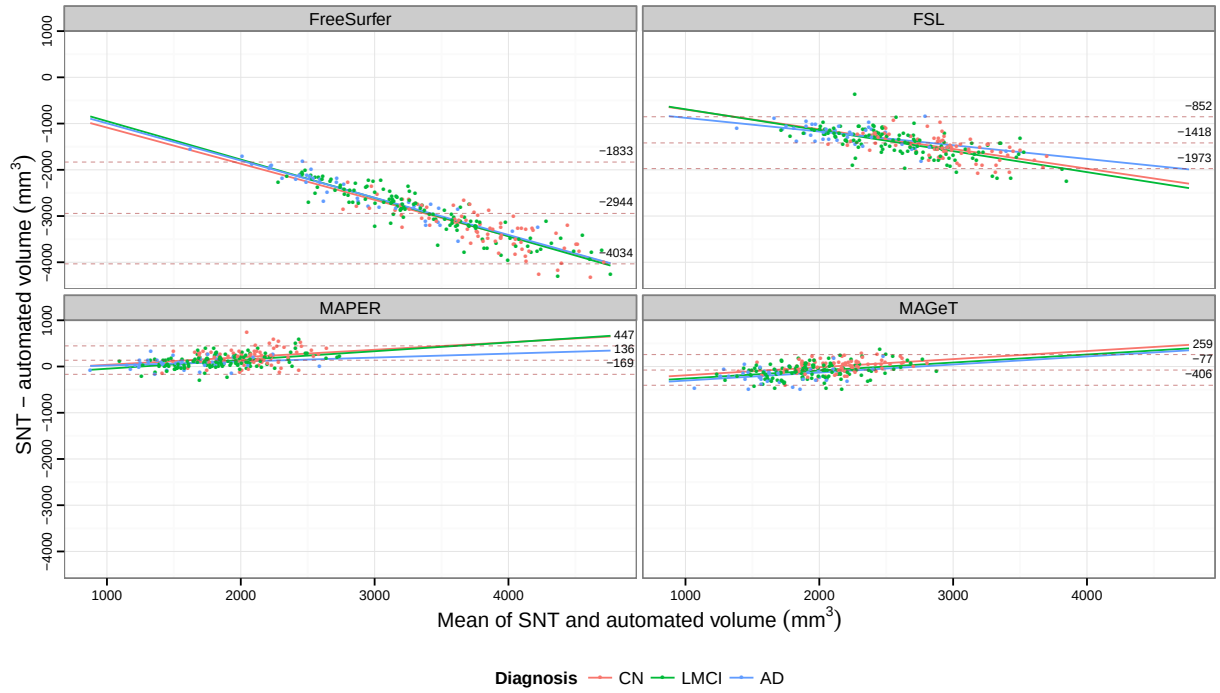
Figure 5 shows a qualitative comparison of MAGeT-Brain and SNT hippocampal segmentations for 10 randomly selected subjects in each disease category, and illustrates some of the common errors found during visual inspection. Mostly frequently, we found MAGeT-Brain improperly includes the vestigial hippocampal 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.

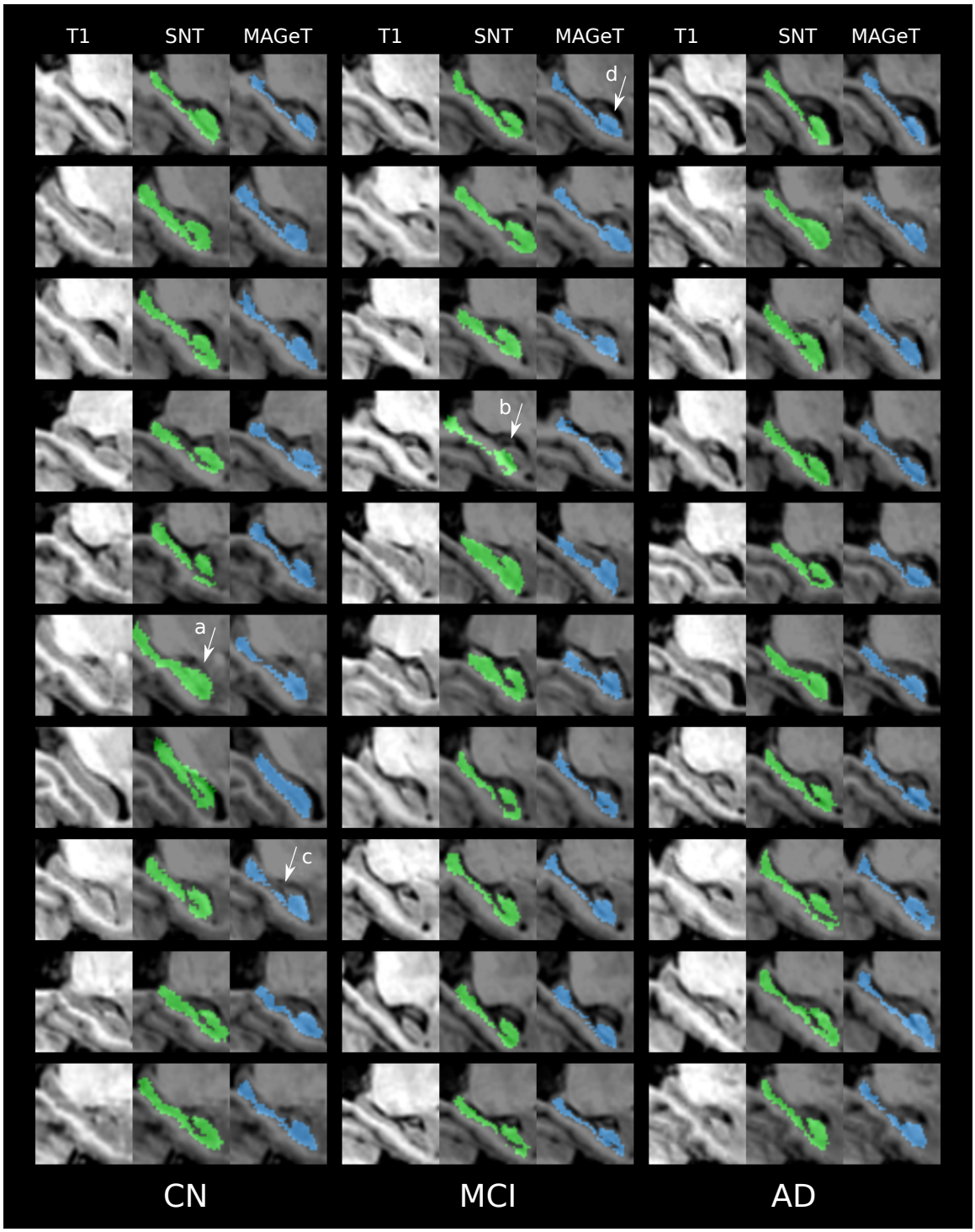



Figure 5: SNT and MAGeT-Brain segmentations for 30 ADNI subjects — 10 subjects randomly selected from each disease category in the subject pool used in Experiment 1 (Section 3.1. Sagittal slices are shown for each unlabelled T1-weighted anatomical image. SNT labels appear in green, and MAGeT-Brain labels 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 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 MAGeT-Brain produces hippocampal subfields segmentations at this resolution that correspond in form to the segmentation protocol used by the given high-resolution atlas library 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, using the resampled Winterburn segmentations as definitive for the $0.9mm^3$ resolution we evaluate reliability of MAGeT-Brain segmentations using DSC overlap scores and evaluate consistency across the range of hippocampal sizes using Bland-Altman plots of subfield volumes.

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.

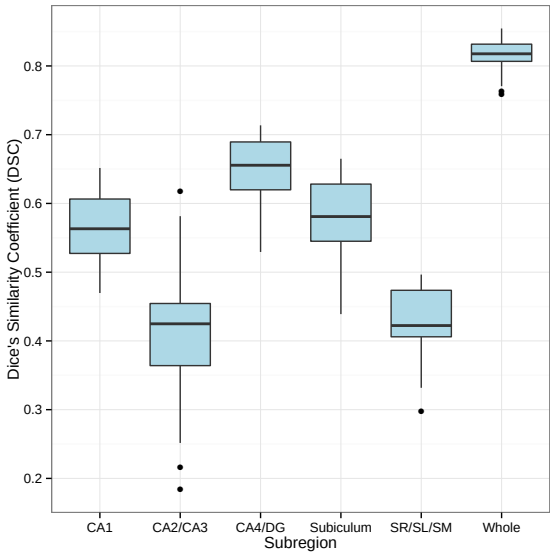
Additionally, by shifting the original manual $0.3mm$-isotropic voxel segmentations by one voxel in the x, y, and z direction and then resampling it to $0.9mm$-isotropic voxels we obtain a simulated manual segmentation having a small amount of error. We can compare the DSC overlap score of the shifted 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

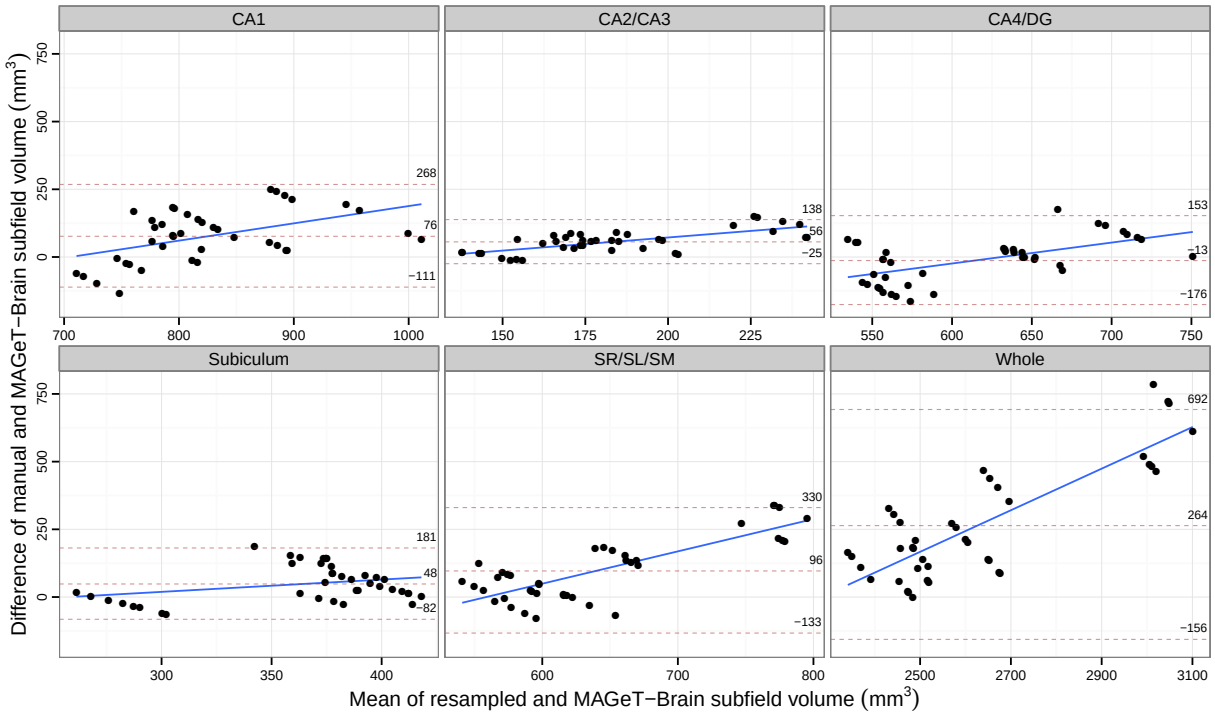
Figure 6a shows the overlap similarity scores between the MAGeT-Brain segmentations and the resampled Winterburn atlases for each hippocampal subfield across all subjects and folds of the validation. Mean and standard deviation DSC scores of the subfields are shown in Table 7, along with DSC scores for 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.

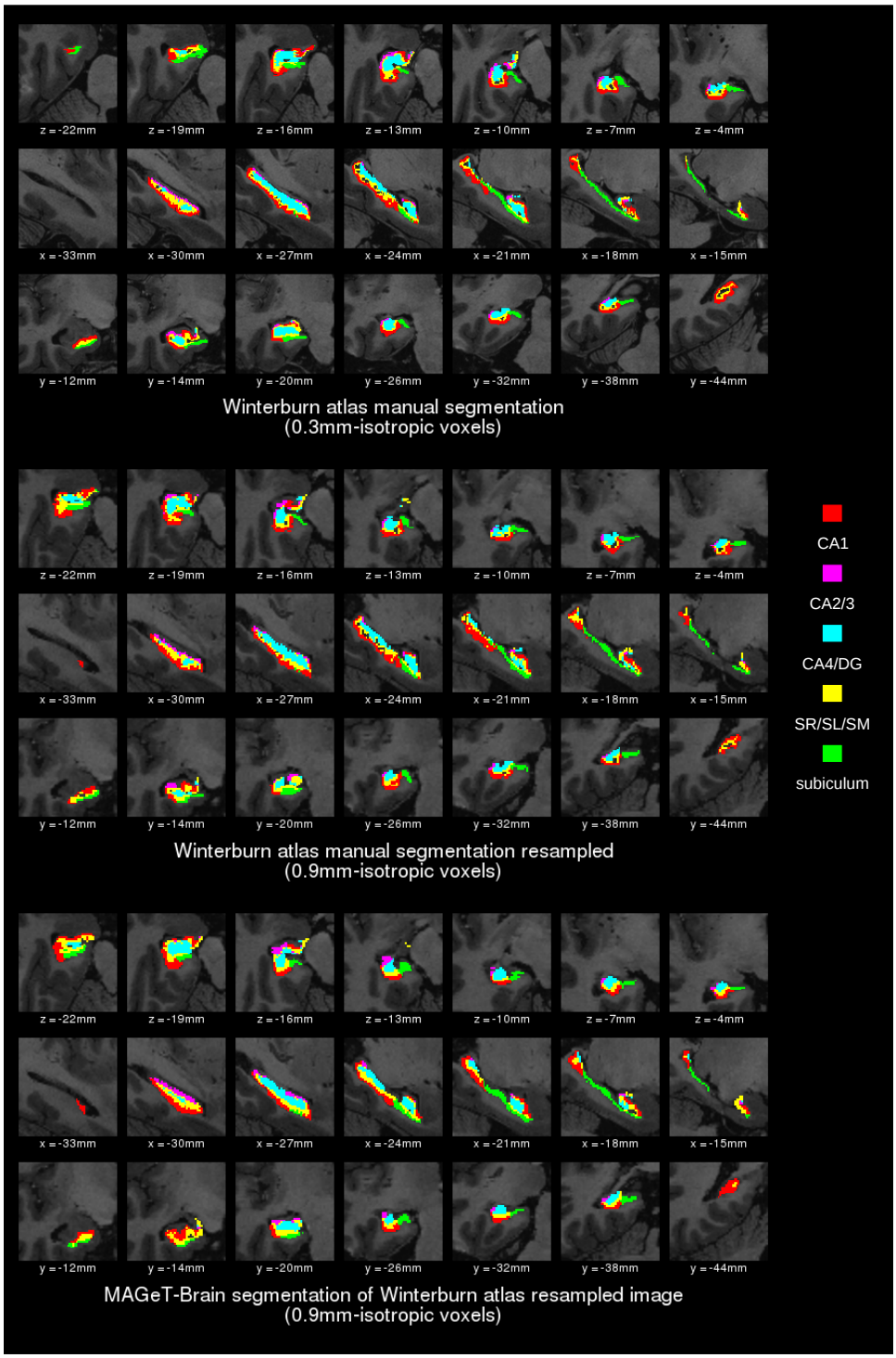



Figure 7: **Detailed subfield segmentation results for a single subject.** In the upper left corner is the original high-resolution Winterburn atlas manual subfield segmentation; in the upper right corner is the Winterburn atlas segmentation subsampled from 0.3mm- to 0.9mm-isotropic voxels; in the lower 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 | $0.56 \pm 0.05$ | $0.27 \pm 0.03$ |
| CA2/CA3 | $0.41 \pm 0.10$ | $0.12 \pm 0.05$ |
| CA4/DG | $0.65 \pm 0.05$ | $0.42 \pm 0.05$ |
| SR/SL/SM | $0.43 \pm 0.05$ | $0.19 \pm 0.04$ |
| Subiculum | $0.58 \pm 0.06$ | $0.14 \pm 0.04$ |

# 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. 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 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 from a small set of labelled atlas images – is validated in the cross validation conducted in Experiment 1 (and the replication in Experiment 2 and Experiment 5, Supplementary Materials). We find that both increasing the number of atlases and the number of templates used improves MAGeT-Brain segmentation over and above basic-multi-atlas segmentations using the same number of atlas images. That is, by taking the extra step of generating a template library 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). 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 over simple majority vote label fusion, and yet we did not see a significant indication of this effect in the MAGeT-Brain segmentations. Selectively filtering out atlases

Table 8: **Automated segmentation accuracy of the Hippocampus** This table shows the best published Dice's overlap measure between automated and ground truth segmentations. 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<br>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 |

with lower image similarity is believed to reduce sources of error from estimating deformations via non-linear 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 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.

On that note, one author (JW), an expert manual rater (Winterburn et al., 2013), identified regular in-

consistencies in the SNT segmentations: occurrences of over- and under-estimation, as well as misalignments of the entire segmentation volume (Figure 5). Although the SNT segmentations are used as benchmarks for validation in many other studies (Table 8), these segmentation inconsistencies present the possibility that a more accurate and consistent benchmark segmentation protocol ought to be used in order to truly understand the results of such validations. Indeed, our replication of the 10-fold cross-validation using SNT segmentations (Experiment 5, Supplementary Materials) shows noticeably poorer mean similarity scores for both MAGeT-Brain and multi-atlas.

Thus, in comparison to other methodologies in the field MAGeT-Brain performs favourably. Table 8 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 agreement 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 agreement that follows increasing the number of atlases and templates comes at an increased computational cost ($O(log(n))$), as previously mentioned in other work (Heckemann et al., 2006a).

Among the automated segmentation methods we compared in this paper (FreeSurfer, MAPER, FSL FIRST), we find extremely variable performance of all methods. With the exception of FSL FIRST all methods correlate well with the 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

Table 9: **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/dentate gyrus, subiculum, entorhinal cortex |
| Van Leemput et al. (2009) | CA1, CA2/CA3, CA4/dentate gyrus, 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 |
| Mueller et al. (2007) | CA1, CA2, CA3/CA4 & dentate gyrus, Sibiculum, entorhinal cortex |

so more work is needed to understand the full extent to which these methods differ.

Finally, we have provided evidence that using the Winterburn high-resolution hippocampal subfield atlases (Winterburn et al., 2013) our algorithmic framework is appropriate for the segmentation of hippocampal subfields in standard T1-weighted data. Subfield segmentation is a burgeoning topic in the literature although very few automated methods are available for the segmentation of 3T data (Yushkevich et al., 2009, 2010; Van Leemput et al., 2009). Table 10 compares segmentation agreement from some of these methods and MAGeT-Brain. The overlap DSC scores for MAGeT-Brain subfields are notably lower but a direct comparison of overlap values must be done cautiously. In the present work, our overlap scores are computed on $0.9mm$-isotropic voxel resolution images, whereas Yushkevich et al. (2010) uses focal $0.4 \times 0.5 \times 2.0mm$ voxel resolution images, and Van Leemput et al. (2009) use supersampled $0.380mm$-isotropic voxel resolution images. The larger voxel images we use necessarily entail a greater change in DSC for each incorrectly labelled voxel. In addition, our automated segmentations are compared to manual segmentations resampled from $0.3mm$-isotropic voxel labels; the resampling process inevitably introduces noise which may lower overlap scores. Lastly, as our method is aimed specifically at situations when manually produced atlases are scarce, in our cross validation we are forced to use three rather than all five of the Winterburn atlases (which, based on our findings with whole hippocampal segmentation, would have resulted in improved overlap similarity). Although having more atlases would be ideal in this context, these atlases are very time consuming to generate (Winterburn et al., 2013). Nevertheless, the advantage of evaluating MAGeT-Brain on standard 3T T1-weighted resolution MR images with a publically available atlas library is that our results reflect typical usage scenarios of researchers and clinicians.

Experiments 1, 2, and 5 have demonstrated that our algorithm flexibly accommodates different whole hippocampus manual segmentation methodologies. We have not explicitly evaluated a subfield definition other than the Winterburn protocol, and therefore it is possible that using an alternate subfield definition could improve the reliability of our automated subfield definitions. For example, established definitions such as those from Mueller et al. (2007) could be a prime candidate for further exploration. In addition, the conservative nature of the Mueller definition (labelling of the 5 slices in the hippocampus body only) would likely further aid in reliability measurement. However, there are two main logistical problems that we would have to overcome prior to implementation. The first is that these definitions were developed for data that is highly anisotropic ($0.4mm \times 0.5mm \times 2mm$), and it is unclear how our algorithms would deal with such atlases used as input. The second is that, since these atlases are not publicly available, we would be have to re-implement the protocol using our atlases. At the present time it is unclear how we would adapt these protocol to data that we used, where subfield segmentations are defined on $0.3mm^3$ voxels. However, the impact of subfield definitions in the context of our work is an important one and should be considered in

Table 10: **A comparison of subfield segmentation overlap similarity with manual raters.**

| Subfield | MAGeT-Brain | Van Leemput et al. (2009) | Yushkevich et al. (2010) |
|---|---|---|---|
| CA1 | 0.563 | 0.62 | 0.875 |
| CA2/3 | 0.412 | 0.74 | $CA2 = 0.538, CA3 = 0.618$ |
| CA4/DG | 0.647 | 0.68 | $DG = 0.873$ |
| presubiculum | — | 0.68 | — |
| subiculum | 0.58 | 0.74 | 0.770 |
| hippocampal fissure | — | 0.53 | — |
| SR/SL/SM | 0.428 | — | — |
| fimbria | — | 0.51 | — |
| head | — | — | 0.902 |
| tail | — | — | 0.863 |

subsequent studies.

One further complication common to all subfield segmentation evaluation is that, by its nature, the Dice's Similarity Coefficient score penalizes structures with high surface area-to-volume ratios. Therefore subfield DSC scores will generally be lower than whole hippocampal segmentations. We attempted to put this effect into perspective by comparing MAGeT-Brain subfield segmentation agreement with the agreement of voxel-shifted manual segmentations (Table 7). The results of this exercise show conclusively, despite the very limited number of atlases we had to work with, that MAGeT-Brain subfield segmentations are well within the bounds of error of a $0.3mm^3$ voxel shift.

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

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

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

# 5 Acknowledgements

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

Computations were performed on the gpc supercomputer at the SciNet HPC Consortium (Loken et al., 2010). SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

In addition, computations were performed on the CAMH Specialized Computing Cluster. The SCC is funded by: The Canada Foundation for Innovation, Research Hospital Fund.

ADNI Acknowledgements: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development, LLC.; Johnson & Johnson Pharmaceutical Research Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is Rev March 26, 2012 coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

We would also like to thank G. Clinton, E. Hazel, and B. Worrell for inspiring this work.

# 6 Supplementary Materials

## 6.1 SNT Hippocampal Labels

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

Table 11: **ADNI1:Complete 1Yr 1.5T 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 $N = 23$ | LMCI $N = 23$ | AD $N = 23$ | Combined $N = 69$ |
|---|---|---|---|---|
| Age at baseline Years | 72.2 75.5 78.5 | 71.0 77.1 81.4 | 71.7 77.8 81.8 | 71.5 76.6 81.3 |
| Sex : Female | 43% (10) | 43% (10) | 43% (10) | 43% (30) |
| Education | 16.0 16.0 18.0 | 15.0 16.0 18.0 | 12.0 16.0 16.5 | 14.0 16.0 18.0 |
| CDR-SB | 0.00 0.00 0.00 | 0.75 1.50 1.50 | 4.00 4.50 5.00 | 0.00 1.50 4.00 |
| ADAS 13 | 4.67 5.67 12.34 | 14.34 16.00 20.50 | 23.83 29.00 31.66 | 10.00 16.00 25.33 |
| MMSE | 28.5 29.0 30.0 | 25.0 27.0 28.0 | 21.0 23.0 24.0 | 24.0 27.0 29.0 |

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

of an intraclass coefficient better than .94 (Hsu et al., 2002).

## 6.2 Experiment 5: Whole Hippocampus Segmentation Cross-Validation — Alzheimer's Disease, SNT Segmentations

This experiment is a replication of Experiment 1 using a pool of 69 images and SNT semi-automated segmentations from the ADNI dataset (Hsu et al., 2002). See Experiment 1 for full details on the ADNI dataset, and validation process.

### 6.2.1 Experiment 5: Materials and Methods

**Dataset** 69 1.5T images were arbitrarily selected from the baseline scans in the *ADNI1:Complete 1Yr 1.5T* standardized dataset. 23 subjects were chosen from each disease category: cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD). Demographics for this subset are shown in Table 1. Each image has a corresponding semi-automated segmentation of the left and right whole hippocampus made available with ADNI images (SNT; see Supplementary Materials).

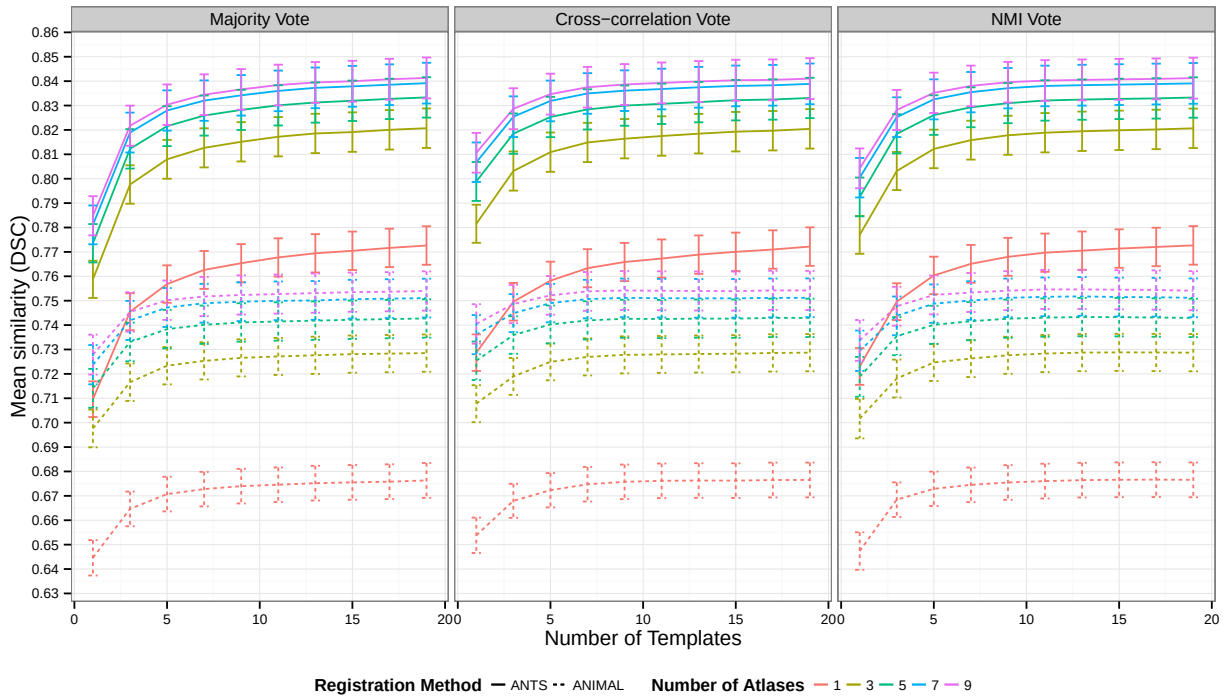
**Experiment details** A total of ten validation rounds were performed on each subject in the dataset, for each combination of parameter settings: atlas library size (1-9), template library size (1-20), registration method (ANTS or ANIMAL), and label fusion method (majority vote, cross-correlation weighted majority vote, and normalized mutual information weighted majority vote). A total of $10 \times 69 \times 9 \times 20 \times 2 \times 3 = 7.452 \times 10^5$ validation rounds are conducted. The computed segmentations for a subject are compared to the SNT labels provided by ADNI using Dice's Similarity Coefficient and the score is averaged over the validation rounds.

### 6.2.2 Experiment 5: Results

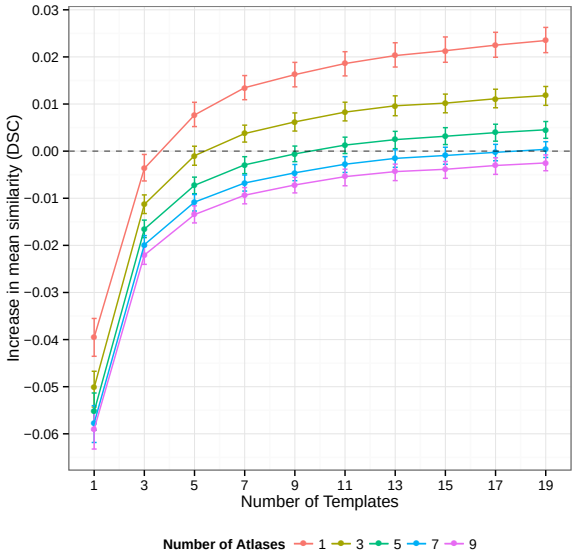
As when comparing against manual labels in Experiment 1, we find similar behaviour when comparing MAGeT-Brain segmentations to SNT labels: similarity scores increase with increasing numbers of atlases and templates, with diminishing increases in improvement trending towards a plateau (Figure 2a). As in Experiment 1, using ANTS registration leads to significantly increased similarity scores, and there is no significant difference in scores from any of the label fusion methods. Mean DSC score peaks at 0.841 when using 9 atlases, 19 templates, ANTS registration, and majority vote label fusion. Compared to multi-atlas

segmentations, we find MAGeT-Brain segmentations show increasing improvement with larger atlas and template libraries when using more than 9 templates and 5 or fewer atlases (Figure 8b). Peak improvement (+0.023 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 8c) with a large enough template library.

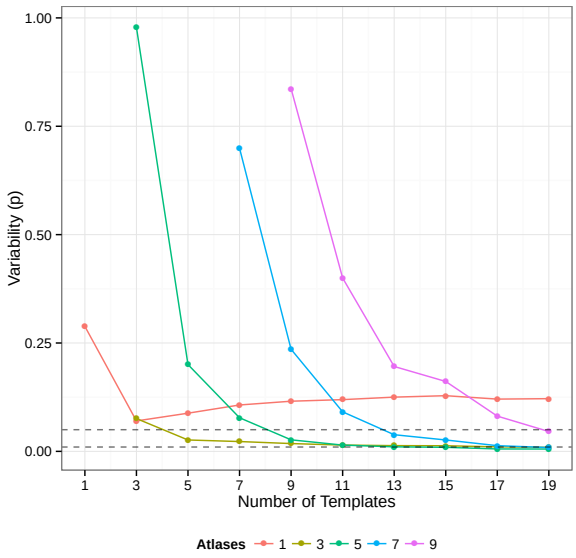
S




(a) DSC vs. atlas and template library size




(b) Increase in similarity score over multi-atlas




(c) Difference in variability with multi-atlas

Figure 8: **Whole hippocampus segmentation cross-validation on ADNI subjects with SNT segmentations.** (8a) Average DSC score of MAGeT-Brain with SNT segmentations for 69 ADNI subjects taken over 10 folds of cross-validation at each parameter setting. Error bars indicate standard error. (8b) Increase in DSC of MAGeT-Brain over multi-atlas segmentations. (8c) 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.

# References


D. H. Adler, J. Pluta, S. Kadivar, C. Craige, J. C. Gee, B. B. Avants, and P. a. Yushkevich. Histology-derived volumetric annotation of the human hippocampal subfields in postmortem MRI. *NeuroImage*, 84:505–23, Jan. 2014. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.08.067.

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. Barnes, J. Foster, R. G. Boyes, T. Pepple, E. K. Moore, J. M. Schott, C. Frost, R. I. Scahill, and N. C. Fox. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *NeuroImage*, 40(4):1655–71, May 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.01.012.

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.

829 G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformation of brain anatomy. *IEEE*
830 *transactions on medical imaging*, 16(6):864–77, Dec. 1997. ISSN 0278-0062. doi: 10.1109/42.650882.

831 M. Chupin, E. Gérardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehéricy, H. Benali, L. Garnero, and
832 O. Colliot. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild
833 cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579–87, June 2009. ISSN 1098-1063.
834 doi: 10.1002/hipo.20626.

835 D. L. Collins and J. C. Pruessner. Towards accurate, automatic segmentation of the hippocampus and
836 amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52
837 (4):1355–66, Oct. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.04.193.

838 D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR
839 volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, 18(2):192–205,
840 1994. ISSN 0363-8715.

841 D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3-D model-based neuroanatomical
842 segmentation. *Human Brain Mapping*, 3(3):190–208, Oct. 1995. ISSN 10659471. doi: 10.1002/hbm.
843 460030304.

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

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

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

853 T. den Heijer, F. V. der Lijn, M. W. Vernooij, M. de Groot, P. J. Koudstaal, a. V. der Lugt, G. P.
854 Krestin, a. Hofman, W. J. Niessen, and M. M. B. Breteler. Structural and diffusion MRI measures of
855 the hippocampus and memory performance. *NeuroImage*, 63(4):1782–9, Dec. 2012. ISSN 1095-9572. doi:
856 10.1016/j.neuroimage.2012.08.067.

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

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

863 I. S. Gousias, D. Rueckert, R. a. Heckemann, L. E. Dyet, J. P. Boardman, a. D. Edwards, and A. Hammers.
864 Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2):
865 672–84, Apr. 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.11.034.

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

M. Hartig, D. Truran-sacrey, S. Raptentsetsang, N. Schuff, and M. Weiner. USCF FreeSurfer Overview and 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.

A. Jeneson and L. Squire. Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory*, 19(1):15–25, 2012. doi: 10.1101/lm.024018.111.

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.

C. Loken, D. Gruner, L. Groer, R. Peltier, N. Bunn, M. Craig, T. Henriques, J. Dempsey, C.-H. Yu, J. Chen, L. J. Dursi, J. Chong, S. Northrup, J. Pinto, N. Knecht, and R. V. Zon. SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. *Journal of Physics: Conference Series*, 256: 012026, Nov. 2010. ISSN 1742-6596. doi: 10.1088/1742-6596/256/1/012026.

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.

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

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

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

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

S. Mueller, L. Stables, A. Du, and N. Schuff. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4T. *Neurobiology of . . .*, l(5):719–726, 2007. doi: 10.1016/j.neurobiolaging.2006.03.007.

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

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

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

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

K. M. Pohl, S. Bouix, M. Nakamura, T. Rohlfing, R. W. McCarley, R. Kikinis, W. E. L. Grimson, M. E. Shenton, and W. M. Wells. A hierarchical algorithm for MR brain image parcellation. *IEEE transactions on medical imaging*, 26(9):1201–12, Sept. 2007. ISSN 0278-0062. doi: 10.1109/TMI.2007.901433.

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

S. Powell, V. A. Magnotta, H. Johnson, V. K. Jammalamadaka, R. Pierson, and N. C. Andreasen. Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage*, 39(1):238–47, Jan. 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2007.05.063.

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.

F. van der Lijn, T. den Heijer, M. M. B. Breteler, and W. J. Niessen. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43(4):708–20, Dec. 2008. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2008.07.058.

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.

J. Wixted and L. Squire. The medial temporal lobe and the attributes of memory. *Trends in cognitive sciences*, 15(5):210–217, 2011. doi: 10.1016/j.tics.2011.03.005.

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.

P. A. Yushkevich, H. Wang, J. Pluta, S. R. Das, C. Craige, B. B. Avants, M. W. Weiner, and S. Mueller. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *NeuroImage*, 53(4):1208–24, Dec. 2010. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.06.040.