Hippocampal segmentation with MAGeT

Jon Pipitone

November 21, 2012

Abstract

Neuroimaging research often relies on automated anatomical segmentations of MR images of the brain. Current multi-atlas based approaches provide accurate segmentations of brain images by propagating manually derived segmentations of specific neuroanatomical structures to unlabelled data. These approaches often rely on a large number of such manually segmented atlases that take significant time and expertise to produce. We present an algorithm for the automatic segmentation of the hippocampus that minimizes the number of atlases needed while still achieving similar accuracy to multi-atlas approaches.

TODO (MC) Finish this...

1 Introduction

The hippocampus is of particular interest to many researchers because it is implicated in forms of brain dysfunction such as Alzheimer's disease and schizophrenia, and has functional significance in cognitive processes such as learning and memory. For many research questions involving magnetic resonance imaging (MRI) data accurate identification of the hippocampus and its subregions is a necessary first step to better understand the individual neuroanatomy of subjects.

Currently, the gold standard for neuroanatomical segmentation is manual delineation by an expert human rater. This is problematic for segmentation of the hippocampus for several reasons. First, manual segmentation takes a significant investment of time and expertise [**?**] which may not be readily available to researchers or clinicians. Second, the amount of data produced in neuroimaging experiments increasingly exceeds the capacity for identification of specific neuroanatomical structures by an expert manual rater. Third, the true definition of hippocampal anatomy in MR images is disputed [**?**], as evidenced by efforts to create an unified segmentation protocol [**?**].

Compounding each of these problems is the significant neuroanatomical variability in the hippocampus throughout the course of aging, development, and neuropsychiatric disorders [**?**]. Additionally, it may be necessary to use several different hippocampal definitions or, in fact, make specific modifications in the course of research. For example, Poppenk et al. [**?**] found that subdividing the hippocampus into anterior and posterior regions resulted in a predictive relationship between volume difference of those regions and recollection memory performance. Thus, while manual segmentation of the hippocampus is a necessary technique, to researchers or clinicians who do not have access to the needed human expertise its use may be infeasible.

Automated segmentation techniques overcome the need for human expertise by performing segmentations computationally. A popular class of automated methods, *multi-atlas-based segmentation*, rely on a set of expertly labeled neuroanatomical atlases. Each atlas is warped to fit a subject's neuroanatomy using nonlinear registration techniques [? ?]. Atlas labels are then transformed by this warping and a *label fusion* technique, such as voxel-wise voting, is used to merge the competing label definitions from each atlas into a final segmentation for a subject.

Many descriptions of multi-atlas-based segmentation algorithms report relying on an atlas library containing anywhere between 30 and 80 expertly labeled brains [? ? ? ? ?]. As noted, the production of an atlas library requires significant manual effort, and is limited since the choice of atlases or segmentation protocol may not reflect the underlying neuroanatomical variability of the population under study or be suited to answer the research questions at hand.

In this paper we propose an automated segmentation method to address the above issues of existing multi-atlas-based methods. Principally, our method aims to dramatically reduce the number of manually labelled atlases needed (under 10). This is achieved by using the small atlas library to boot-strap a much larger "template library", which is then used to segment the subjects in a similar fashion to basic multi-atlas segmentation. This approach has the additional advantage of using the unique subject population on hand to initialize the segmentation process and improve accuracy.

The essential insight of generating a template library is not new. Heckemann [?] compared generating a template library from a single atlas to standard multi-atlas segmentation and found poor performance and so deemed the approach as inviable. The LEAP algorithm [?] proceeds by iteratively segmenting the unlabelled image most similar to the atlas library images and then incorporating the now-labelled image into the atlas library, but requires 30 starting atlases. The novelty of our method is to demonstrate the possibility of producing comparable segmentation accuracy to these and other multi-atlas-based methods while using significantly fewer manually created atlases.

In our previous work [?], we applied MAGeT brain to segmentation of the human striatum, globus pallidus, and thalamus using a single histologically-derived atlas. The main contribution of this paper is to extend our approach to the human hippocampus and perform a thorough validation over a range of atlas and template library sizes, which was not done in our previous work. Due to the small number of atlases required, our method can easily accommodate different hippocampal definitions. Our aim is not to improve on segmentation accuracy beyond existing methods, but instead to provide a method that trades off manual segmentation expertise for computational processing time while providing sufficient accuracy for clinical and research applications.

2 Materials and Methods

explain two registration methods, definitions, fusion methods. compare to each other, and then compare atlas-based, true multi-atlas, naive methods. trying to make choices as to the best reg/fusion/

evaluations described previously in Chakravarty2012.

2.1 Segmentation Algorithm

In this paper, we use the term *atlas* to mean any manually segmented MR image, and the term *atlas library* to mean a set of such images. We use the term *template* to refer to any MR image, and associated labelling, used to segment another image, and the term *template*

TODO Make sure we explain what MAGeT stands for *library* to refer to a set of such images. An atlas library may be used as a template library but, as we will discuss, a template library may also be composed of images with computer generated labellings.

The segmentation approach we propose is best understood as an extension of basic multi-atlas segmentation [**?**]. In multi-atlas segmentation, an atlas library and unlabelled MR images are given as input. Every atlas image is nonlinearly registered to each unlabelled image, and then each atlas' labels are propagated via the resulting transformations. These labels are then fused to produce a single, definitive segmentation by some label fusion method (e.g. voxel-wise majority vote).

Our extension adds a preliminary stage in which a template library is constructed from input images, and used in place of an atlas library in the standard multi-atlas-based method. To create the template library, labels from each atlas image are propagated to each template library image via the transformation resulting from a non-linear registration between pair of images. As a result, each template library image has a label from each atlas. Basic multi-atlas segmentation is then used to produce segmentations for the entire set of unlabelled images (including those images used in the template library).

Label fusion is performed by cross-correlation weighted voting, a strategy weighted towards an optimal combination of subjects from the template library which has been previously shown to improve segmentation accuracy [**? ?**]. In this method, each template library image is ranked in similarity to each unlabelled image by the normalized cross-correlation of image intensities after linear registration in 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 [**?**].

Source code can be found at `http://github.com/pipitone/MAGeTbrain`.

Algorithm 1 Pseudocode for the MAGeT Brain algorithm

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

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

2.2 Image Processing and Registration Methods

Before images were registered, the N3 algorithm [16] is first used to minimize the intensity nonuniformity in each of the atlases and unlabeled subject images. In this work we investigated the performance of two registration methods.

2.2.1 Automatic Normalization and Image Matching and Anatomical Labeling (ANIMAL)

The ANIMAL algorithm carries out Image registration is in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that maximizes the correlation between blurred MR intensities and gradient magnitude over the whole brain [?]. In the second phase, nonlinear registration is completed using the ANIMAL algorithm [?]: an iterative procedure that estimates a 3D deformation field between two MR images. At first, large deformations are estimated using blurred version of the input data. These larger deformations are then input to subsequent steps where the fit is refined by estimating smaller deformations on data blurred with a Gaussian kernel with a smaller FWHM. The final transformation is a set of local translations defined on a bed of equally spaced nodes that were estimated through the optimization of the correlation coefficient.

For the purposes of this work we used the regularization parameters optimized in Robbins et al. [?]. It should be noted that the MAGeT brain algorithm is not dependent on this, or any, particular choice of registration method [?].

TODO should add the MNI website here; also we should add the parameters we used somewhere

2.2.2 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. 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`

TODO (same here we should add the parameters we used).

2.3 Label Fusion Methods

Label fusion is term given to the process of combining the information from several candidate labellings for an MR image into a single labelling. In this paper we explore the benefits of three different fusion methods.

2.3.1 Voxel-wise Majority Vote

Labels are propagated from all template library images to a subject. Each output voxel is given the most frequent label at that voxel location amongst all candidate labellings. Ties are broken arbitrarily.

2.3.2 Cross-correlation Weighted Majority Vote

Labels are propagated from only those template library images which most similar to the subject. A voxel-wise majority vote is carried out from the labels from the top n ranked template library images, where n is parameter set by the user.

Similarity between a subject and a template image is measured using normalized cross-correlation of intensity values, over a region of interest defined on each template, after the subject has been linearly registered to the template. The heuristic used for defining a meaningful region of interest for each template is to use the extent of all the propagated labellings from each atlas (after linear registration only) and then dilating this region by three voxels.

The utility to calculate the cross-correlation similarity measure is implemented as part of the ANIMAL toolkit.

TODO (ANIMAL or MINC)

2.3.3 Normalised Mutual Information Weighted Majority Vote

This label fusion method is identical to the above process except a normalised mutual information score is used over the region of interest between a template and subject instead of cross-correlation.

The utility to calculate the normalised mutual information measure is implemented as part of the EZMinc package (based on ITK NMI routine).

2.4 Goodness-of-fit

Each segmentation was evaluated against the 'gold-standard' manual segmentation from the dataset using the Dice Kappa (κ) overlap metric:

$$\kappa = \frac{2a}{2a + b + c}$$

where a is the number of voxels common to the candidate segmentation and the gold standard and $b + c$ is the sum of the voxels uniquely identified by either the automatically generated candidate segmentation or the gold-standard.

TODO I would do this after you define the experiments

2.5 Data and Experiments

Three different experiments were conducted to evaluate the influence of registration method, atlas set size and resolution, template set size, and label fusion method and parameters.

2.5.1 ADNI Validation Experiment

In this experiment we performed repeated random sub-sampling cross-validation of the MAGeT algorithm with a pool of 69 images and labels from the ADNI dataset. The images are T1-weighted . The hippocampal labels are provided as part of the ADNI dataset, and are produced . The parameters we varied, and the ranges over which we varied are listed in . We performed 10 validations per parameter combination. In each validation trial, a random sample of subjects were chosen from the pool without replacement as the atlas set, whereas the template set was chosen randomly with replacement.

TODO ¡insert description of scan/machine types¿

TODO ¡insert description of SNT labellings¿

TODO table

We used the following parameters for ANTS, during registration:

Parameter	Values tested
Number of Templates	3 to 20
Number of Atlases	3 to 9
Registration Method	ANTS or ANIMAL
Label Fusion Method	majority vote, cross-correlation weighted vote, NMI weighted vote.

```
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.

TODO ¡rationale?¿

For the purposes of this work we used the regularization parameters optimized in Robbins et al.[15].

TODO right? (but added a second layer).

TODO multi-atlas comparison, "naive" comparison

2.5.2 Winterburn High-resolution Hippocampal Atlas - ADNI Validation

In this experiment we explored using the MAGeT brain algorithm with the Winterburn high-resolution atlases to segment 21 randomly chosen MR images from the ADNI dataset (seven each of healthy, MCI and AD subjects). The Winterburn atlases are digital segmentations of the hippocampus in five in-vivo 300u isotropic T1-weighted MR scans, and include subfield segmentations for the cornus ammonis (CA) 1, CA4, dentate gyrus, subiculum, and CA 2 and 3 combined. Subjects in the Winterburn atlases range in age from 29-57 years (mean age of 37), and include two males and three females.

Since hippocampal segmentation protocols differ between the ADNI labels and Winterburn atlases, this poses a problem for direct similarity comparisons between labels produced by MAGeT brain and the ADNI labels. To evaluate the performance of MAGeT brain, we compared classification accuracy of subjects by diagnosis based on hippocampal volume using both the SMT labels and our produced labels.

TODO explain why we did(n't) resegment the ADNI images with a the low-res protocol.

TODO description of validation – rms `validate` or t-test: contrast with QDA or LDA (Coupe 2011) used in LOOCV

TODO For minctracc – we registered to TAL

2.6 Results

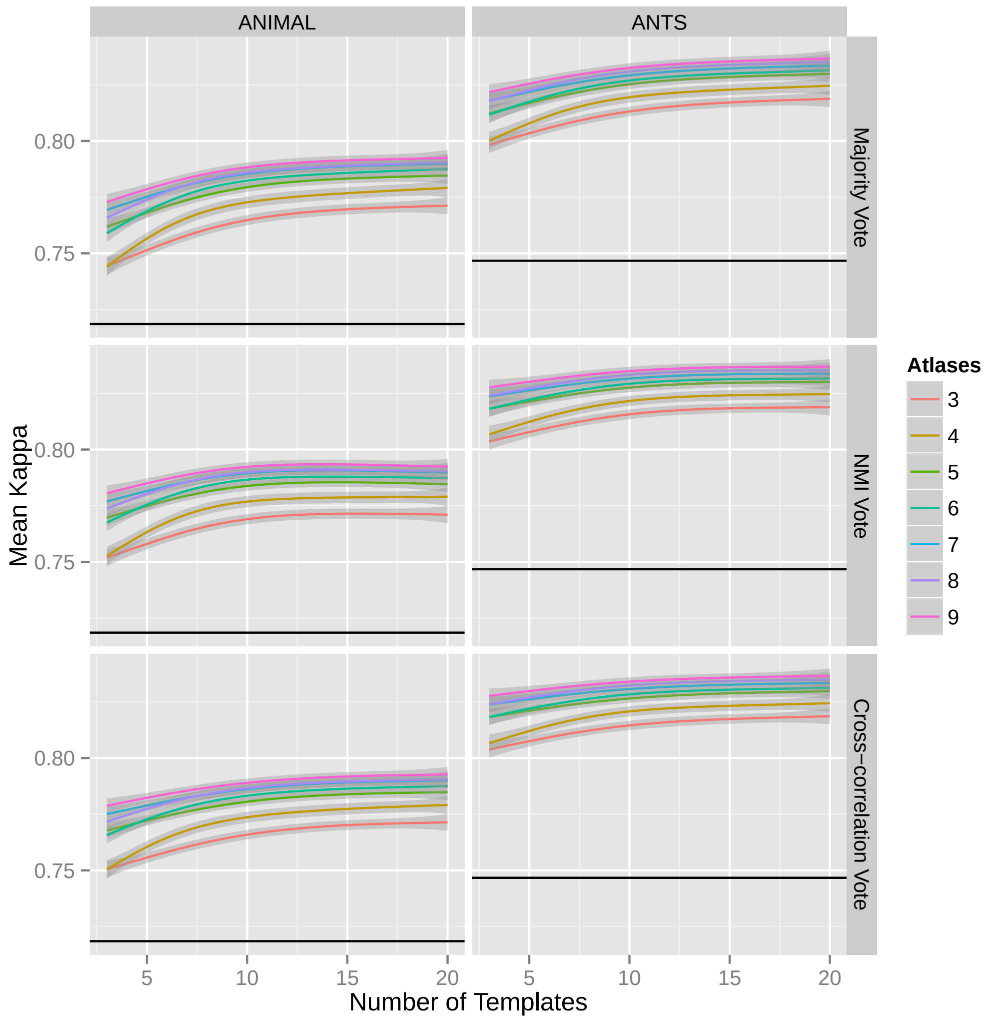
2.6.1 A2A

Kappa vs. Number of templates: Smoothing line fitted using GAM (generalised additive model) from R with defaults from ggplot2 (formula: y   s(x, bs = "cs"))

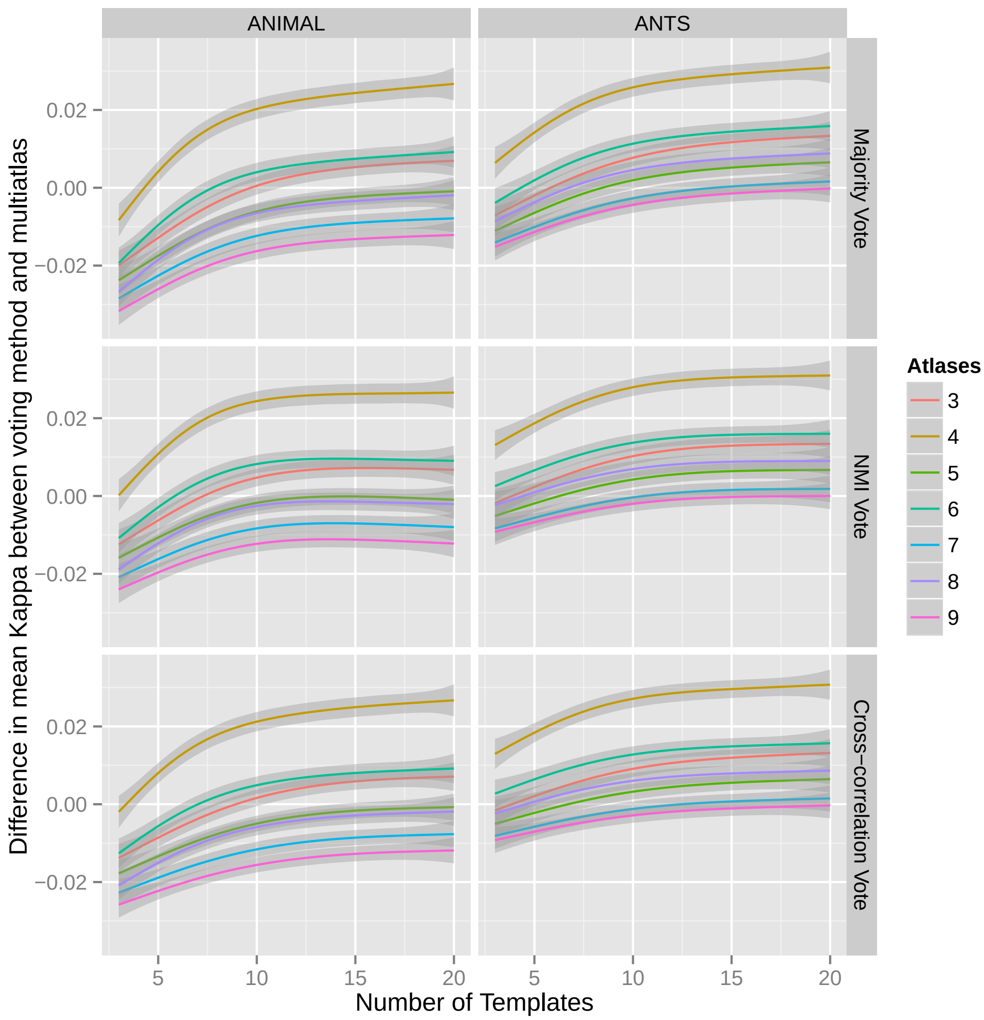
Table 1: Descriptive Statistics by Baseline Dx

	CN $N = 23$	LMCI $N = 23$	AD $N = 23$	Combined $N = 69$	Test Statistic
Age at baseline Years	72.20 75.50 78.45	71.00 77.10 81.40	71.70 77.80 81.75	71.50 76.60 81.30	$F_{2,66} = 0.15,\ P = 0.862$[1]
Sex : Female	43% (10)	43% (10)	43% (10)	43% (30)	$\chi^2_2 = 0,\ P = 1$[2]
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	$F_{2,66} = 3.44,\ P = 0.038$[1]
Marital status at baseline : Married	74% (17)	87% (20)	87% (20)	83% (57)	$\chi^2_4 = 3.04,\ P = 0.551$[2]
Widowed	22% (5)	13% (3)	13% (3)	16% (11)	
Divorced	4% (1)	0% (0)	0% (0)	1% (1)	
Never married	0% (0)	0% (0)	0% (0)	0% (0)	
Unknown	0% (0)	0% (0)	0% (0)	0% (0)	
Ethnicity : Unknown	0% (0)	0% (0)	0% (0)	0% (0)	[2]
Not Hisp/Latino	100% (23)	100% (23)	100% (23)	100% (69)	
Hisp/Latino	0% (0)	0% (0)	0% (0)	0% (0)	
Race : Am Indian/Alaskan	0% (0)	0% (0)	0% (0)	0% (0)	$\chi^2_2 = 5.61,\ P = 0.061$[2]
Asian	0% (0)	0% (0)	0% (0)	0% (0)	
Hawaiian/Other PI	0% (0)	0% (0)	0% (0)	0% (0)	
Black	17% (4)	4% (1)	0% (0)	7% (5)	
White	83% (19)	96% (22)	100% (23)	93% (64)	
More than one	0% (0)	0% (0)	0% (0)	0% (0)	
Unknown	0% (0)	0% (0)	0% (0)	0% (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	$F_{2,66} = 294.2,\ P < 0.001$[1]
ADAS 13	4.665 5.670 12.335	14.335 16.000 20.500	23.830 29.000 31.665	10.000 16.000 25.330	$F_{2,66} = 109.8,\ P < 0.001$[1]
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	$F_{2,66} = 94.17,\ P < 0.001$[1]

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Numbers after percents are frequencies.
Tests used:
[1]Kruskal-Wallis test; [2]Pearson test


ANIMAL
ANTS
Majority Vote
NMI Vote
Cross−correlation Vote
Mean Kappa
0.80
0.75
Number of Templates
5
10
15
20
Atlases
3
4
5
6
7
8
9

merge	Atlases	ANTS	ANIMAL
1	3	0.805414417498623	0.764266832660852
2	4	0.793646605043623	0.752420649183594
3	5	0.823292141139203	0.785546924392891
4	6	0.815578947302174	0.778345910844141
5	7	0.831936342756014	0.797761650124922
6	8	0.826298462726884	0.792443538092187
7	9	0.836877632304638	0.804568318173359



Kappa minus multi-atlas mean vs. Number of Templates

Multi-atlas means:

- more atlases -¿ better performance - larger template library -¿ better performance, but tails off around 10-15 templates - no significant difference between majority or weighted vote

Number of Atlases	ANIMAL Kappa (Jaccard)	ANTS Kappa (Jaccard)
3	0.76 (0.63)	0.80 (0.69)
4	0.75 (0.62)	0.79 (0.67)
5	0.79 (0.66)	0.82 (0.71)
6	0.78 (0.65)	0.82 (0.70)
7	0.80 (0.67)	0.83 (0.72)
8	0.79 (0.67)	0.83 (0.72)
9	0.80 (0.68)	0.84 (0.73)

methods (haven???t tested this statistically though). - consistently performs better than average naive performance by XXX - using ANTS, with a large enough template library (¿12) MAGeT brain performs better than the average multi-atlas approach with the same number of atlases. using ANIMAL, 5 or more atlases needed before boost seen. - more atlases -¿ smaller template library required to improve on average multi-atlas performance - discuss variance? best/worst case? –how often do we expect random template library selection to work decently

TODO cost (in registrations) / benefit trade off graph: show number of registrations per Kappa? or hours of manual labour per Kappa?)

2.6.2 Winterburn Atlas Segmentation of ADNI Baseline Images

- A2A shows that if atlas population strongly(?) represents subject set variability, then free choice from atlas population will produce improvements (we know this b/c of extensive validation trials). - what about in the case where atlas population doesn???t strongly represent subject set variability (e.g. a priori atlas set)? then, we can use atlas selection to refine atlas set?

TODO Kappa against our manual rater is low

2.6.3 First Episode Schizophrenic patients

Table 2: Descriptive Statistics by DX

	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.

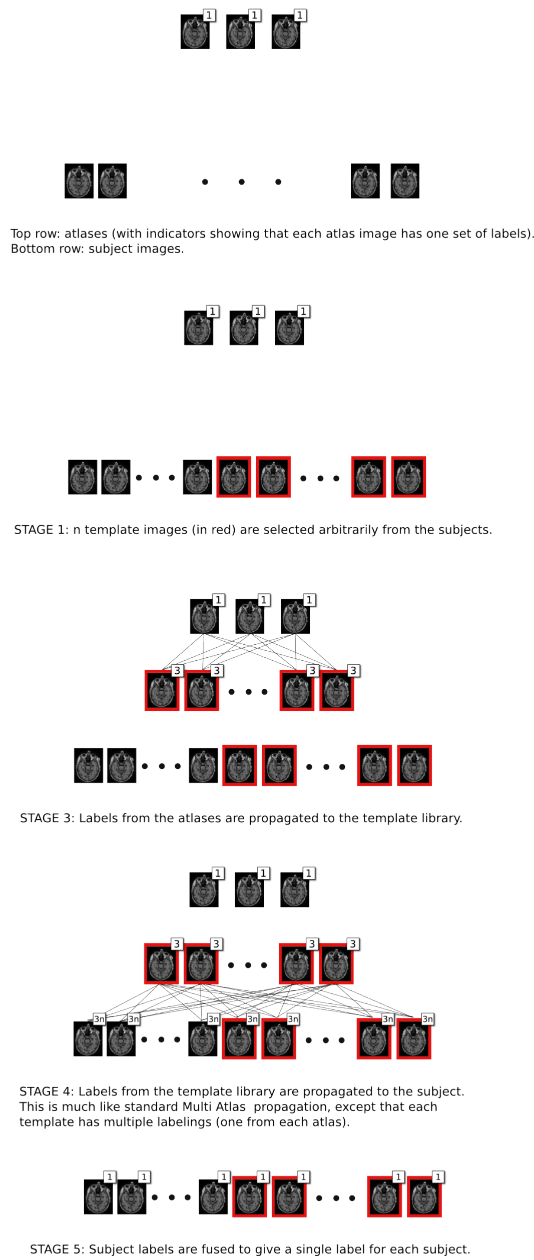


Figure 1: Diagram of the MAGeT Brain algorithm

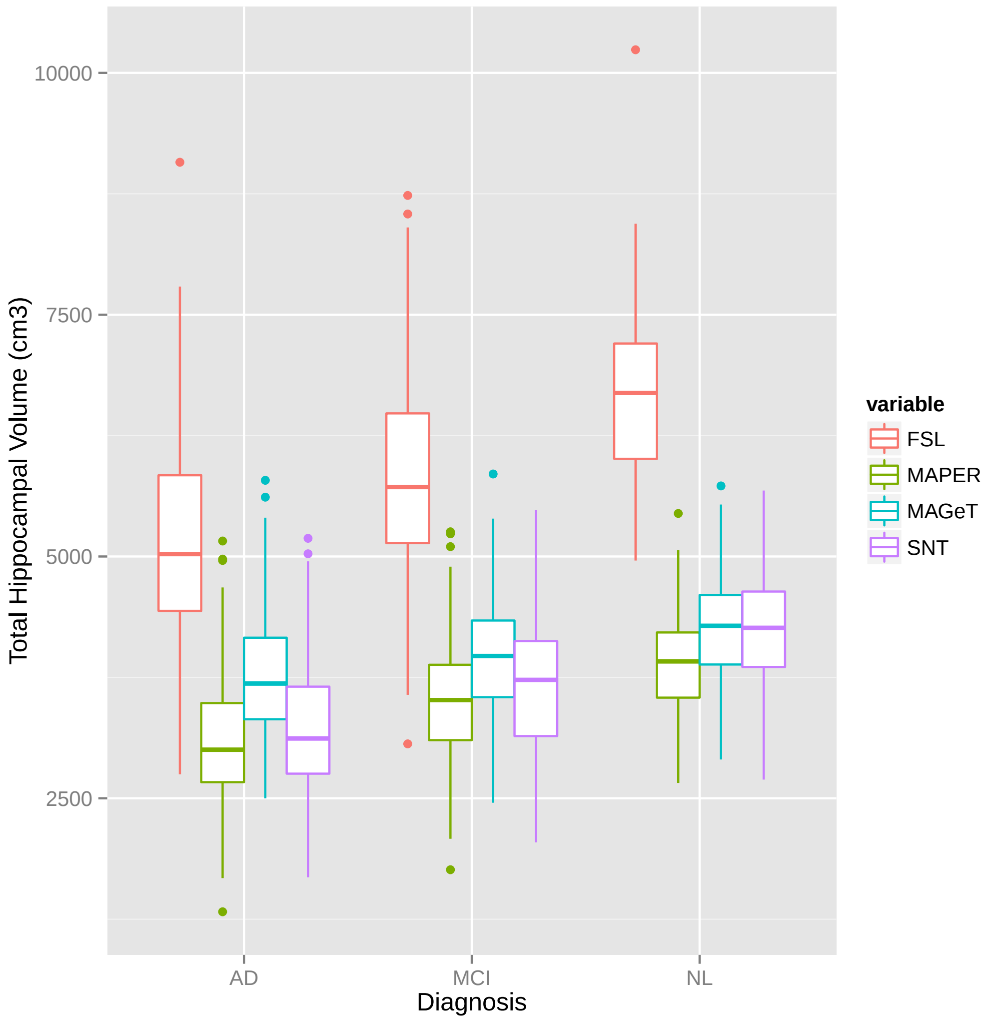


Figure 2: Comparison of HC volumes by FreeSurfer (FSF), MAGeT brain (MAGeT), MAPER, and manual (SNT).

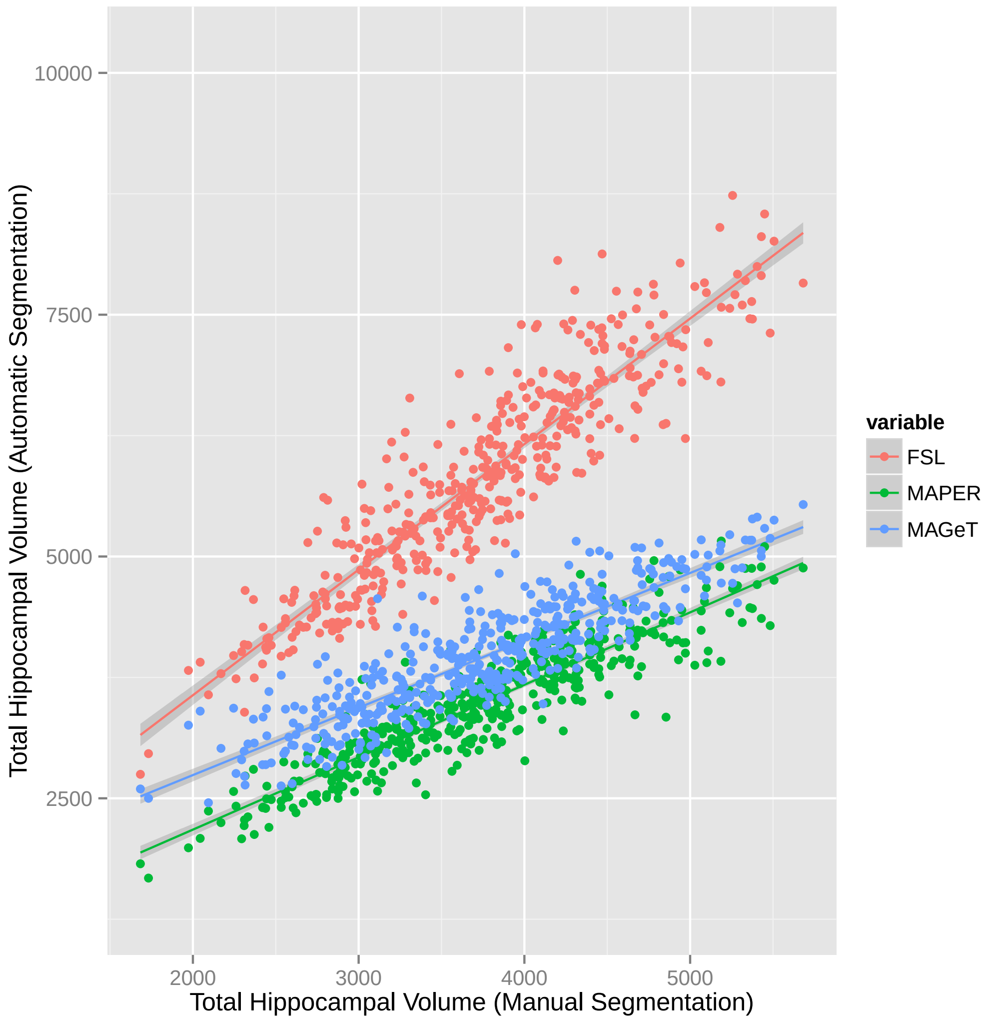


Figure 3: *ADNI Baseline cohort* Comparison of HC volumes by FreeSurfer (FSF), MAGeT brain (MAGeT), MAPER, and manual (SNT).

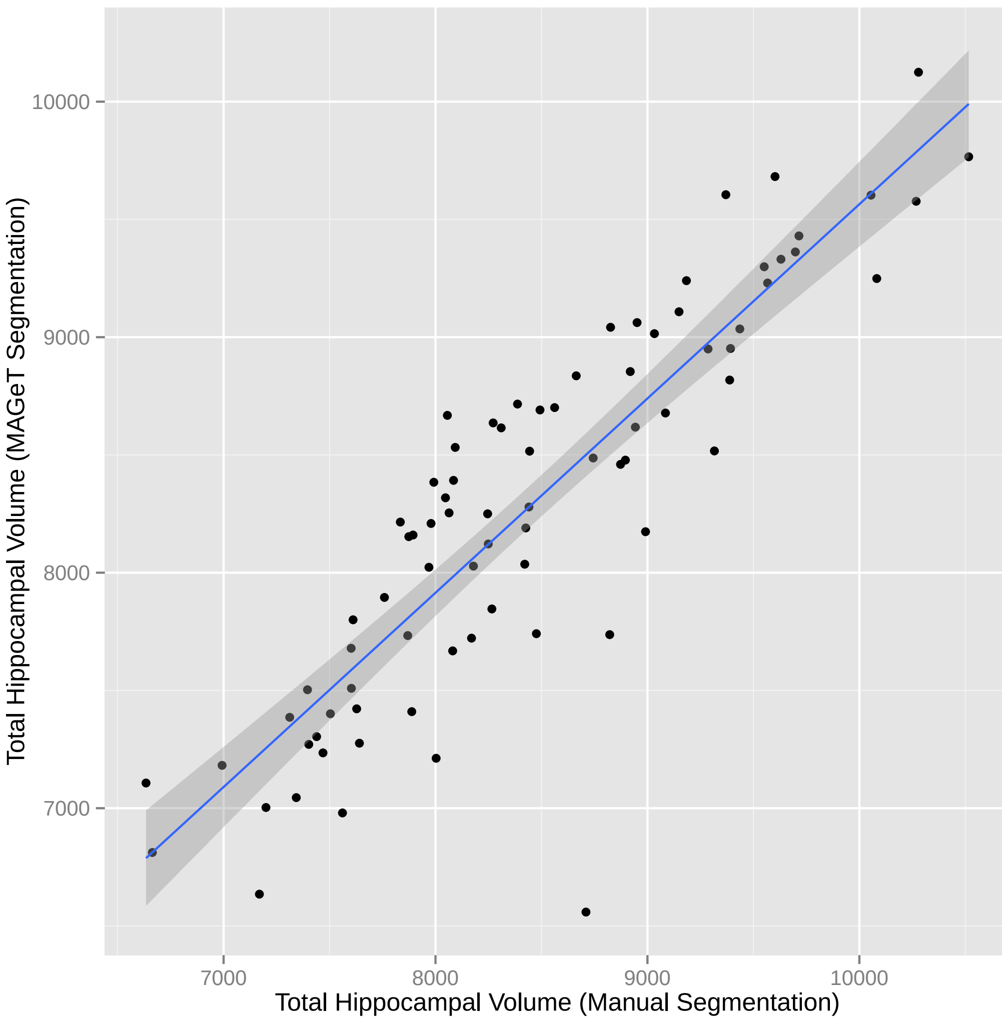


Figure 4: *First Episode Schizophrenic Patients* Comparison of total HC volumes for MAGeT against manually rated volumes of