NeuroImage 155 (2017) 383–393

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A comparison of accurate automatic hippocampal segmentation methods

Azar Zandifar[a,b], Vladimir Fonov[a], Pierrick Coupé[c], Jens Pruessner[d], D. Louis Collins[a,b,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada
[b] Department of Biomedical Engineering, McGill University, Montreal, Canada
[c] Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400, Talence, France
[d] McGill University Research Centre for Studies in Aging, Canada

ARTICLE INFO

Keywords:
Hippocampal segmentation
Alzheimer's disease
Dice's κ
Cohen's d
Area under receiver operating characteristic curve

ABSTRACT

The hippocampus is one of the first brain structures affected by Alzheimer's disease (AD). While many automatic methods for hippocampal segmentation exist, few studies have compared them on the same data. In this study, we compare four fully automated hippocampal segmentation methods in terms of their conformity with manual segmentation and their ability to be used as an AD biomarker in clinical settings. We also apply error correction to the four automatic segmentation methods, and complete a comprehensive validation to investigate differences between the methods. The effect size and classification performance is measured for AD versus normal control (NC) groups and for stable mild cognitive impairment (sMCI) versus progressive mild cognitive impairment (pMCI) groups. Our study shows that the nonlinear patch-based segmentation method with error correction is the most accurate automatic segmentation method and yields the most conformity with manual segmentation ($\kappa = 0.894$). The largest effect size between AD versus NC and sMCI versus pMCI is produced by FreeSurfer with error correction. We further show that, using only hippocampal volume, age, and sex as features, the area under the receiver operating characteristic curve reaches up to 0.8813 for AD versus NC and 0.6451 for sMCI versus pMCI. However, the automatic segmentation methods are not significantly different in their performance.

Introduction

The hippocampus is part of the limbic system and is located in the temporal lobe, medial to the inferior horn of the lateral ventricle. It plays an important role in spatial navigation and consolidation of information from short-term to long-term memory. Hippocampal neuronal degeneration and dysfunction has been vastly studied in a large number of neurodegenerative diseases and mental health disorders such as Alzheimer's dementia (Braak and Braak, 1991), post-traumatic stress disorder (Bremner et al., 1997), major depressive disorder (Kempton et al., 2011), schizophrenia (Nelson et al., 1998), and epilepsy (Wieshmann et al., 1997).

Alzheimer's disease (AD) is the most common cause of dementia (Sperling et al., 2011). The global burden of AD is approximately 25 million people worldwide and is predicted to quadruple in prevalence by 2050 due to aging of the population (Brookmeyer et al., 2007). As with most neurodegenerative diseases, early detection, before too much degeneration of brain tissue occurs, may be crucial for effective treatment.

A quantitative measure of the atrophy pattern can be obtained by segmenting different anatomical structures in magnetic resonance (MR) images. Many recent studies have focused on hippocampal atrophy as one of the first AD biomarkers, since the hippocampus is easier to segment than other medial temporal lobe structures (Chupin et al., 2009; Cuingnet et al., 2011; Coupé et al., 2012) and is highly sensitive to AD-related atrophic patterns (Jack et al., 1997). Manual segmentation of anatomical structures is often considered the gold standard for volumetric assessment. However, it is a laborious, time-consuming task (taking between 30 and 120 min for hippocampal segmentation) and suffers from both inter- and intra-rater variability. To overcome these limitations, several semi- and fully automated methods have been developed for hippocampal segmentation (Coupé et al., 2011; Fischl et al., 2002; Collins and Pruessner, 2010; Chupin et al., 2009; Duchesne et al., 2002; Hu et al., 2011) (for a recent review, see Dill et al. (2015)).

[*] Corresponding author at: McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada H3A 2B4.
E-mail address: louis.collins@mcgill.ca (D.L. Collins).

[1] 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.

http://dx.doi.org/10.1016/j.neuroimage.2017.04.018
Received 21 June 2016; Accepted 7 April 2017
Available online 09 April 2017

Medial temporal lobe anatomy presents automatic segmentation procedures with many nontrivial problems, the most important of which is the striking similarity in the intensity distributions of different anatomical structures such as the amygdala, hippocampus, and entorhinal cortex. As a result, different brain structures cannot be differentiated based solely on intensity information (Fischl et al., 2002). Furthermore, the high anatomical variability, low contrast, small size, and discontinuous boundaries of the hippocampus render its segmentation extremely difficult without the use of prior information (Fischl et al., 2002).

Semi-automated methods use manual intervention as a first step in defining seed points or bounding boxes (Chupin et al., 2007). Yet, even such a simple intervention can pose a major challenge given the size of the databases used in clinical studies (e.g., the Alzheimer's Disease Neuroimaging Initiative, ADNI, http://www.adni.loni.usc.edu).

To fully automate the segmentation procedure, some methods attempt to integrate shape information using appearance-based strategies (Duchesne et al., 2002; Hu and Collins, 2007). Others combine both level-set shape representation and statistical gray level information (Yang and Duncan, 2004; Hu et al., 2011). Template-based segmentation (or label propagation) is one of the major categories in structural segmentation due to its efficacy and accuracy. In these methods, single (Barnes et al., 2008), probabilistic or multiple templates (Heckemann et al., 2006; Aljabar et al., 2009; Collins and Pruessner, 2010; Coupé et al., 2011) are used to spatially label the desired structure through a label propagation step. FreeSurfer, a well-known segmentation tool (Fischl et al., 2002), uses both probabilistic template priors and intensity features from each anatomical structure to complete the structure identification using a classifier.

Using only a single template can be suboptimal for several reasons: First, the template may not be the best match to represent the target image; this type of error may be more pronounced in a diseased population. In addition, certain types of registration errors may be unavoidable, and any labeling error resulting from imperfect manual segmentation may also affect the segmentation procedure (Aljabar et al., 2009).

Multi-template methods combine multiple templates to decrease these types of errors. These methods define each voxel label by combining the labels derived from multiple templates using a decision-making approach such as majority voting (Aljabar et al., 2009; Collins and Pruessner, 2010; Coupé et al., 2011). They are further improved by the use of similar templates from a template library in segmenting a new subject. Multi-template methods implicitly benefit from shape priors by using fine nonlinear registration. However, this improvement comes at the expense of high computational burden. Furthermore, these methods weight each template equally and are therefore highly susceptible to registration error and template choice. Recently, inspired by nonlocal means image-denoising techniques, a patch-based segmentation method has been shown to reduce computational burden while preserving the segmentation accuracy, accomplished by weighting each expert label based on the intensity similarity with the to-be-segmented patch (Coupé et al., 2011). Patch-based methods have become extremely popular in medical image segmentation and have been employed in brain structure segmentation (Coupé et al., 2011; Rousseau et al., 2011; Xiao et al., 2015), lesion segmentation (Guizard et al., 2015; Roy et al., 2015), and segmentation of cardiac MR images (Bai et al., 2015). Different images can be efficiently represented based on patches of a training library, and patch-based segmentation methods serve as an example of a dictionary learning problem. A recent method approached image segmentation based on discriminative dictionary learning and sparse reconstruction (Tong et al., 2013). However, patch-based methods are not without their shortcomings. For example, the image similarity index over a small patch may not be the optimal descriptor for the overall structure. The two approaches (multi-template and patch-based) can be combined to take advantage of both implicit shape priors of the non-linear registration and the robustness of patch-based methods to segmentation errors (Fonov et al., 2012; Wang et al., 2013).

Despite the progress achieved by these different methods, segmentation errors and mislabeled voxels remain unavoidable. Wang et al. (2011) presented a machine learning approach to find and correct mislabeled voxels, which can be used as a post hoc correction for any automatic segmentation method. For a given image, their classifier investigates the automatic segmentation for the mislabeled voxels and corrects them. In addition, they showed that a patch-based method with nonlinear registration can be further enhanced using this correction technique, producing the most accurate segmentation strategy developed thus far (Wang and Yushkevich, 2013).

Quantitative comparison of the various automatic segmentation methods poses some difficulties. The reported results may be affected by varying image acquisition parameters, different manual structure segmentation protocols, different gold standards, and different metrics. While manual segmentation is used as the gold standard for automatic segmentation methods, there are diverse ways of segmenting anatomical structures. This inconsistency between expert priors greatly affects the comparison of different segmentation methods. A reliable protocol for manual segmentation is crucial for a valid comparison.

To address the issues identified above, this study compares four different automatic hippocampal segmentation methods, based on a well-defined manual segmentation protocol, using the same dataset. We improve upon the four methods using error correction (Wang et al., 2011) and then compare the resulting eight segmentations. We also evaluate the power of these techniques to be used as an AD biomarker. We characterize some of the factors that lead to more accurate segmentation and further investigate whether this accuracy improves the ability of hippocampal volumetry to distinguish between different AD clinical groups.

Materials and methods

Among the different automatic hippocampal segmentation approaches reported in a recent review from Dill et al. (2015), multi-template segmentation strategies appear to be the most accurate, with Dice's kappa values ranging from 0.72 to 0.91. Three of the four methods compared in this study are multi-template segmentation methods. The first two methods (ANIMAL with template library and label fusion; patch-based segmentation using expert priors) were selected from the most successful and accurate segmentation methods, with reported Dice's kappa values of 0.88 (Collins and Pruessner, 2010; Coupé et al., 2011). The third method (multi-atlas labeling with population-specific template and nonlocal patch-based label fusion) is an enhancement to Coupé et al. (2011) using fine nonlinear registration, with a Dice's kappa of 0.89 (Fonov et al., 2012).

We compared the three multi-template methods (Collins and Pruessner, 2010; Coupé et al., 2011; Fonov et al., 2012) using a common training and testing dataset, manually segmented using the protocol presented in Pruessner et al. (2000). Use of a common dataset avoids the inhomogeneity that might be imposed on comparative studies that evaluate different datasets. A fourth automatic segmentation method, FreeSurfer (a well-known freely available tool) (Fischl et al., 2002), has shown better results in terms of agreement with manual segmentation in comparison with both FSL/FIRST (Morey et al., 2009) and IBASPM (Tae et al., 2008). Therefore, we also compare hippocampal segmentation derived from FreeSurfer with manual segmentation. Finally, since the wrapper-based error correction technique (Wang et al., 2011) can be applied to any host segmentation strategy (Wang et al., 2013; Dill et al., 2015), we applied the error correction technique (Wang et al., 2011) to all four compared methods.

In the second part of the study, we estimated the power of each technique to differentiate between groups based on hippocampal volume. The methods were compared based on their ability to differentiate between AD and normal control (NC) groups (diagnosis ability) and between groups of subjects with stable versus progressive mild cognitive impairment (MCI) (prognosis ability). In this study, progressive MCI (pMCI) refers to MCI subjects who converted to AD during three years of follow-up, and stable MCI (sMCI) refers to subjects who maintained their MCI status. We measured effect size and classification performance, the latter measured by area under the receiver operating characteristic (ROC) curve.

Datasets

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership. The project is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations led by Michael W. Weiner, MD, as principal investigator. The primary goal of the ADNI has been to test whether serial MR imaging, positron emission tomography (PET), other biomarkers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD (for more information, visit http://www.adni-info.org).

Experiment 1

To validate the different approaches to hippocampal automatic segmentation, we selected sixty 1.5T images from the baseline ADNI scans. Selection was based on age, sex, handedness, and years of education, so that the groups in different clinical stages (NC (n=20), MCI (n=20), and AD (n=20)) were comparable in terms of age, sex, and median years of education (see table 1). This dataset was used for training for all the selected methods, except for FreeSurfer, both without and with error correction. In FreeSurfer, the segmentation was performed based on its default training dataset. However, the labels are corrected (in FreeSurfer with error correction) based on the dataset of 60 images described above.

Experiment 2

For the clinical grouping and power analysis study, we selected all the ADNI-1 baseline visit 1.5T T1 MR images. The demographic information can be found in Table 2.

Preprocessing

All selected images went through a preprocessing pipeline, consisting of denoising (Coupé et al., 2008), N3 inhomogeneity correction (Sled et al., 1998), linear intensity normalization based on histogram matching between the image and the average template, and affine registration to ICBM152 template space with $1 \times 1 \times 1$ mm^3 resolution (Collins et al., 1994) using a population-specific template from the ADNI-1 database (Fonov et al., 2011). Lastly, all the images were coarsely aligned, and image intensities were normalized within each image and among the whole database (Coupé et al., 2012; Fonov et al., 2012; Collins and Pruessner, 2010).

Table 1
Experiment 1: dataset information. NC: normal controls; MCI: mild cognitive impairment; AD: Alzheimer's disease; MMSE: Mini-Mental State Examination.

	NC	MCI	AD	combined
Number	20	20	20	60
Median age at baseline (yrs)	75.5	75.6	74.9	75.2
Sex: female (%)	50%	50%	50%	50%
Median education (yrs)	16.0	16.0	15.5	16.0
Median MMSE score	29.5	27.5	23.0	27.0

Table 2
Experiment 2: dataset information. NC: normal controls; sMCI: stable mild cognitive impairment; pMCI: progressive mild cognitive impairment; AD: Alzheimer's Disease; MMSE: Mini-Mental State Examination.

	NC	sMCI	pMCI	AD
Number	231	240	168	199
Median age at baseline (yrs)	75.93	75.68	75.14	76.04
Sex: female (%)	48%	33%	39%	49%
Median MMSE score	29.0	27.5	26.0	23.0

Manual segmentation

The hippocampi were manually segmented on the first experimental dataset (60 subjects from ADNI-1) using the manual segmentation protocol presented in Pruessner et al. (2000). The labels were segmented on linearly transformed stereotaxic volumes. We used the same preprocessed dataset as the direct input for all the automated segmentation methods to facilitate the comparison. The tracings were performed by expert raters. Raters were blinded to the hemisphere (images were randomly flipped left-right) and to the group (the filenames did not identify the subject as NC, MCI, or AD).

Automatic segmentation methods

FreeSurfer

FreeSurfer is a suite of tools for analysis of human brain data obtained from MR images and includes automatic segmentation of many macroscopically visible brain structures (Fischl, 2012). The tool performs subcortical structure segmentation using the method presented in Fischl et al. (2002). First, each image is linearly aligned with an average template. A frequency histogram of possible structures (defined by the spatial template prior) is used to compute the probability of a given anatomical label occurring at a given location. The prior of a given spatial arrangement of different labels is also incorporated into the segmentation.

The segmentation procedure is considered an anisotropic nonstationary Markov random field. The probabilistic labels and corresponding predicted image intensities are priors, and the intensity similarity between the target image and the atlas is a likelihood term (Fischl et al., 2002). In this study, we used version 5.3.0 of FreeSurfer, which is freely available from https://surfer.nmr.mgh.harvard.edu/.

ANIMAL with template library and label fusion

The ANIMAL segmentation method (Collins and Evans, 1997; Collins et al., 1995) is enhanced with a multi-template strategy (Collins and Pruessner, 2010). The enhanced version benefits from a library of templates and label fusion of their corresponding propagated labels, while the original ANIMAL method uses a single nonlinear averaged template. In ANIMAL, the target image is nonlinearly warped to a template, which is manually segmented by an expert. The structural voxel labels are propagated back to the corresponding voxel in the target space using the inverse of the estimated nonlinear transformation. However, the average template might not be representative of the anatomy of all subjects. To take into account anatomical variability, ANIMAL with template library and label fusion uses n most similar templates (using normalized mutual information as the similarity metric) from a template library. Then, an ANIMAL segmentation is performed using each of these templates. The final label is determined based on a majority voting step between the n labels assigned to each voxel. The number of templates used for training is n=11, as suggested in the original paper (Collins and Pruessner, 2010). Throughout the rest of this paper, we refer to this method as *ANIMAL with label fusion*.

Patch-based segmentation using expert priors

The first patch-based approach to segment brain structures was proposed in Coupé et al. (2011). The algorithm attempts to determine a desired label for a voxel in the target image based on similar patches of neighborhood voxels in the template library. This label fusion step uses a nonlocal means estimator that weights each of the template patch labels using an intensity-based similarity metric with the target patch. Finally, the target patch's label is defined by thresholding the weighted average of its similar patches from the template library. Rigid registration is only used in the segmentation procedure. The implementation is described comprehensively in Coupé et al. (2011). Throughout the rest of this paper, we refer to this method as the *patch-based method.*

Multi-atlas labeling with population-specific template and nonlocal patch-based label fusion

This method is a combination of the previously described patch-based segmentation method with a population-specific nonlinear template construction (Fonov et al., 2012; Wang et al., 2013). When creating a population-based average model, all individual templates (and their labeled structures) are nonlinearly aligned (Avants et al., 2008) to the average model. In this segmentation method, each target image is then nonlinearly warped to the average template space, and the patch-based segmentation algorithm is applied. Finally, the inverse of the nonlinear transformation is used to warp the labels back to the subject's native space. Since this method offers an enhancement to the patch-based method by incorporating a nonlinearly aligned template library, throughout the rest of this paper, we refer to this method as the *nonlinear patch-based method.*

A learning-based wrapper method to correct systematic errors in automatic image segmentation

Wang et al. (2011) presented a method that learns the pattern of mismatch between automatic segmentation labels and their corresponding manual segmentations. The method uses a classification technique in which the classifier is trained by a set of automatically labeled voxels and their corresponding manual segmentation. For an unseen image, the method corrects the mislabeled segmentation using the learned pattern. Both intensity and neighborhood information are used as features to train an AdaBoost learner (Freund and Schapire, 1995). We applied error correction to the four previously described methods. For error correction, we used version 1.9 of the freely available online implementation of the method, which can be found here http://www.nitrc.org/projects/segadapter. All hyperparameters were set based on the original paper (Wang et al., 2011). Throughout the rest of this paper, we refer to the corrected version of each method as the method *with error correction.*

Validation schema

Experiment 1

To investigate segmentation accuracy, we used a leave-one-out (LOO) procedure to validate the methods. Specifically, for each subject in the 60-subject dataset, the methods were trained with the other remaining 59 subjects and their corresponding labels. For testing, the automatic segmentation of the one subject is compared with its corresponding manual segmentation.

Experiment 2

As stated above, the entire ADNI-1 1.5T dataset was used to investigate the power of the segmented structure's volume to act as an AD biomarker. We used two classification experiments (NC vs AD, and sMCI vs pMCI) to evaluate biomarker performance using a LOO procedure.

Metrics

Experiment 1

- Dice's kappa metric

  Volumetric overlap between each automatic segmentation method (i.e., FreeSurfer, ANIMAL with label fusion, patch-based, nonlinear patch-based, and nonlinear patch-based with error correction) and manual segmentation is measured using Dice's kappa similarity index (Zijdenbos et al., 1994). The kappa index for two segmented labels is computed as follows:

$$\kappa = 2 \times \frac{V(M) \cap V(A)}{V(M) + V(A)}, \tag{1}$$

  where M and A represent a set of manually and automatically labeled voxels, respectively. $V(.)$ is the volume operator, while $\cap$ represents set intersection. The value of κ varies between 0 and 1, where 1 indicates the complete overlap between the manual segmentation and that of the automatic method. Dice's kappa similarity metric is computed by comparing the segmentation from each automatic method with the manual segmentation for both left and right hippocampi. Since FreeSurfer uses a different anatomical definition of the hippocampus for training, we expect its κ values will not be as high as those for the other methods.
- Intraclass correlation coefficient

  We used the regression coefficient and the intraclass correlation coefficient (ICC) to show the similarities between automatically and manually segmented volumes (Shrout and Fleiss, 1979).

Experiment 2

- Cohen's *d* effect size

  In order to investigate the sensitivity of each method in detecting between-group differences in a clinical setting, we computed Cohen's *d* effect size based on the hippocampal volumes derived by each automatic segmentation method. Cohen's *d* effect size measures the separability of two normal distributions. It is computed as follows:

$$\text{Cohen's } d = \frac{m_1 - m_2}{SD_{\text{Pooled}}} \tag{2}$$

$$SD_{\text{Pooled}} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}, \tag{3}$$

  where m and SD are the mean and standard deviation, respectively.
- Receiver operating characteristic curve

  To assess the performance of each method in detecting the clinical state of single subjects, we trained and tested a linear discriminant analysis (LDA) classifier to assign the subjects to different groups. We used the LDA classifier in two different classification tasks: to assign subjects to either the AD or NC group and to either the sMCI or pMCI group. For this problem, we used a simple linear classifier fed with mean hippocampal volumes along with age and sex as features. We used scikit-learn (Pedregosa et al., 2011), a Python-based implementation of LDA, and all the hyperparameters were set to their default values. The classification task is evaluated based on ROC curves, which are calculated through a LOO cross-validation procedure. Since separate training and test datasets are used in this framework, classification analysis yields a more conservative performance evaluation than Cohen's *d* effect size.

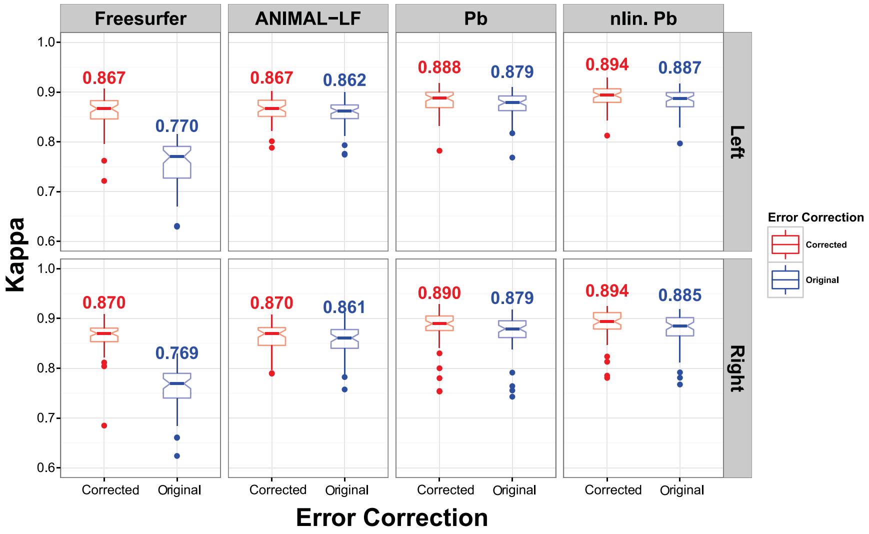

Fig. 1. Method comparison. Kappa index distribution for different methods with and without error correction for both hippocampi (Top: left; Bottom: right). Methods, from left to right: FreeSurfer, ANIMAL with label fusion, patch-based (Pb), nonlinear patch-based (nlin. Pb). Original method in red; error-corrected version in blue.

Experiment 1: results

Dice's kappa metric

The κ distribution is shown for the left and right hippocampus respectively in Fig. 1 and is summarized numerically in Table 3. The nonlinear patch-based segmentation method with error correction yields the highest median κ ($\kappa = 0.894$ and 0.894 for the left and right hippocampus, respectively). These values are among the highest Dice's kappas reported in the literature (Dill et al., 2015). A two-way repeated measures ANOVA, with segmentation method and the hemisphere in which the structure is located as independent variables, shows that the differences between the κ values for different methods are significant, while no significant difference was observed when considering either left or right hemisphere. A post hoc paired Wilcoxon test with Bonferroni correction for multiple comparisons shows that the nonlinear patch-based method with error correction has the most overlap with manual segmentation, and the difference is statistically significant. The pairwise comparison further shows that all the differences are significant ($p < .01$), except between FreeSurfer with error correction and ANIMAL with label fusion both with and without error correction, as well as between the patch-based method with error correction and the nonlinear patch-based method. The same pattern was observed for both left and right hippocampi. The test shows that the difference between each method and its corresponding error corrected version is statistically significant. Furthermore, the large bias of FreeSurfer was expected due to a different anatomical definition of the hippocampus in its training phase.

Table 3
κ statistical results for segmentation methods.

Method	Left Hippocampus	Right Hippocampus
FreeSurfer	0.770 (0.049)	0.769 (0.048)
FreeSurfer with EC	0.867 (0.033)	0.870 (0.031)
ANIMAL with label fusion	0.862 (0.028)	0.861 (0.032)
ANIMAL with label fusion with EC	0.867 (0.024)	0.870 (0.030)
Patch-based	0.879 (0.025)	0.879 (0.036)
Patch-based with EC	0.888 (0.025)	0.890 (0.036)
Nonlinear patch-based	0.887 (0.022)	0.885 (0.032)
Nonlinear patch-based with EC	0.894 (0.021)	0.894 (0.031)

Values are median κ, with standard deviation in parentheses.

Volumetric correlation

Fig. 2 shows the volumetric correlations between manual segmentation and the different automatic segmentation methods. In this study, we consider the raters (automatic and manual segmentation methods) to be fixed effects. The results show that the nonlinear patch-based method, with and without error correction, shows the highest correlation with manually segmented volumes for both the left and right hippocampi. The ICC values can be found in Table 4 for both left and right hippocampi for each method.

A two-way repeated measures ANOVA of manually and automatically segmented hippocampal volume difference, with segmentation method and hemisphere as independent variables, shows that volumetric differences are significant for between-methods, but not between the left and right hemispheres. A post hoc paired Wilcoxon test with Bonferroni correction for multiple comparisons shows that the difference between the volumes measured by automatic and manual segmentation is insignificant for all methods, except for FreeSurfer, for which there was a significant difference when compared with manual segmentation and with all other methods ($p < .001$).

Bland-Altman plots (Fig. 3, Table 5) show that FreeSurfer oversegments the structure and furthermore has a tendency to oversegment larger hippocampi more so than smaller ones. An opposite tendency was observed for ANIMAL with label fusion. That method has a slight bias toward undersegmenting, and undersegmentation is more dominant for larger hippocampi, while smaller ones tend to be oversegmented. No significant pattern of difference between small and large hippocampi was observed in either of the other two methods. The plots show that error correction can be considered a bias correction technique, since the magnitude of the mean bias is reduced for each method. Error correction is more successful when applied to methods with large bias, such as FreeSurfer, whereas in the case of the methods with small bias, such as the patch-based techniques, error correction showed little improvement to segmentation performance. However, the variance of the pattern of differences between automatic and manual segmentations is still observable after the correction step has been applied.

Experiment 2: results

Dataset selection

Each method has a different number of failures; hence, we limited

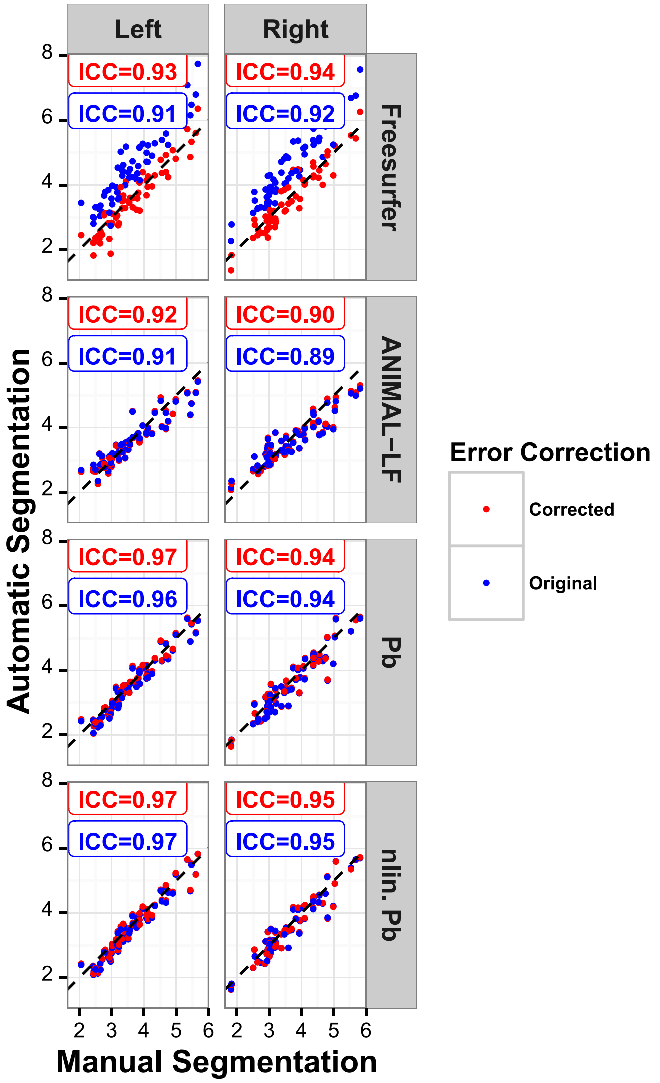

Fig. 2. Volumetric comparison of hippocampal segmentations by automatic methods and manual labeling. Dashed black line represents the unity line. From top to bottom: FreeSurfer, ANIMAL with label fusion, patch-based method (Pb), and nonlinear patch-based method (nlin. Pb). All values are reported in cubic centimeters.

our analysis to the datasets which have successfully passed all the segmentation methods. The largest number of failures belongs to FreeSurfer (see Table 6). We further excluded the 60 subjects (20 AD, 8 pMCI, 12 sMCI, 20 NC) of the first dataset. The final dataset consists of 135 patients with Alzheimer's disease, 152 pMCI, 215 sMCI, and 178 normal controls, for which the demographic information can be found in Table 7.

Table 4
ICC between different automatic segmentation methods and manual segmentation.

Method	Left Hippocampus	Right Hippocampus
FreeSurfer	0.91	0.92
FreeSurfer with EC	0.93	0.94
ANIMAL with label fusion	0.91	0.89
ANIMAL with label fusion with EC	0.92	0.90
Patch-based	0.96	0.94
Patch-based with EC	0.97	0.94
Nonlinear patch-based	0.97	0.95
Nonlinear patch-based with EC	0.97	0.95

Hippocampal volume as an AD biomarker

The effect size between the AD and NC groups for each method is given in Table 8. A prognosis study was also completed, where the effect size was measured between the sMCI and pMCI groups (see Table 9). Effect size was measured as the mean of 200 bootstrapped copies of the dataset. Hippocampal volumes are corrected for age and sex. A paired t-test with Bonferroni correction for multiple comparisons was performed to show significant differences. Based on a conventional operational definition of Cohen's *d*, small, medium, and large effect sizes were defined as $d < 0.5$, $0.5 < d < 0.8$, and $d > 0.8$, respectively. All investigated methods show large effect sizes between the AD and NC groups and medium effect sizes between the sMCI and pMCI groups (see Table 9). Hippocampal volumes are all reported in the population-specific average template stereotaxic space to correct for brain volume. For the AD versus NC experiment, the methods from worst to best (in terms of Cohen's *d*) are ANIMAL with label fusion (d=1.3228), ANIMAL with label fusion and error correction (d=1.3765), FreeSurfer (d=1.4886), the nonlinear patch-based method (d=1.5076), the patch-based method with error correction (d=1.5104), the nonlinear patch-based method with error correction (d=1.5134), the patch-based method (d=1.5793), and FreeSurfer with error correction (d=1.6921). ANIMAL with label fusion (d=1.2356), ANIMAL with label fusion with error correction (d=1.2492), FreeSurfer (d=1.3917), the patch-based method (d=1.4222), the nonlinear patch-based method (d=1.4273), the patch-based method with error correction (d=1.4326), the nonlinear patch-based method with error correction (d=1.4337), and FreeSurfer with error correction (d=1.4880). A pairwise t-test with Bonferroni correction for multiple comparisons shows that FreeSurfer with error correction works significantly better than the other methods to distinguish between AD and NC ($p < .005$). For the sMCI versus pMCI experiment, the methods from worst to best are ANIMAL with label fusion (d=0.4565), ANIMAL with label fusion and error correction (d=0.5040), the patch-based method with error correction (d=0.5450), the nonlinear patch-based method with error correction (d=0.5481), the nonlinear patch-based method (d=0.5551), FreeSurfer (d=0.5919), the patch-based method (0.5988), and FreeSurfer with error correction(d=0.6683). In both experiments a pairwise t-test with Bonferroni correction for multiple comparisons shows that FreeSurfer with error correction works significantly better than the other methods to distinguish in both experiments ($p < .001$). For the sMCI versus pMCI experiment, the methods from worst to best are ANIMAL with label fusion (d=0.4600), ANIMAL with label fusion and error correction (d=0.4620), the nonlinear patch-based method (d=0.5532), the patch-based method with error correction (d=0.5502), the patch-based method (d=0.5545), the nonlinear patch-based method with error correction (0.5605), FreeSurfer (d=0.6006), and FreeSurfer with error correction(d=0.6117). No pairwise significant difference was observed between the nonlinear patch-based method with error correction and FreeSurfer, nor between FreeSurfer with and without error correction, though these methods were significantly better than the others ($p < .005$). Although some of the differences between methods are significant, those differences are very small.

Hippocampal volume as an AD biomarker: ROC curve

We used the LDA classifier with LOO cross-validation to classify all the baseline data. Two similar experiments were run on AD versus NC subjects and stable MCI versus progressive MCI groups (See Fig. 4). The area under the ROC curve (AUC) values are summarized in Table 10. A Cochran's Q test, followed by a post hoc pairwise McNemar test with Bonferroni correction for multiple comparisons, shows there is no significant difference between the methods in terms of classification performance for AD versus NC, nor for sMCI versus pMCI. Furthermore, we observed no significant difference using the volume of left and right hippocampi as separate features.

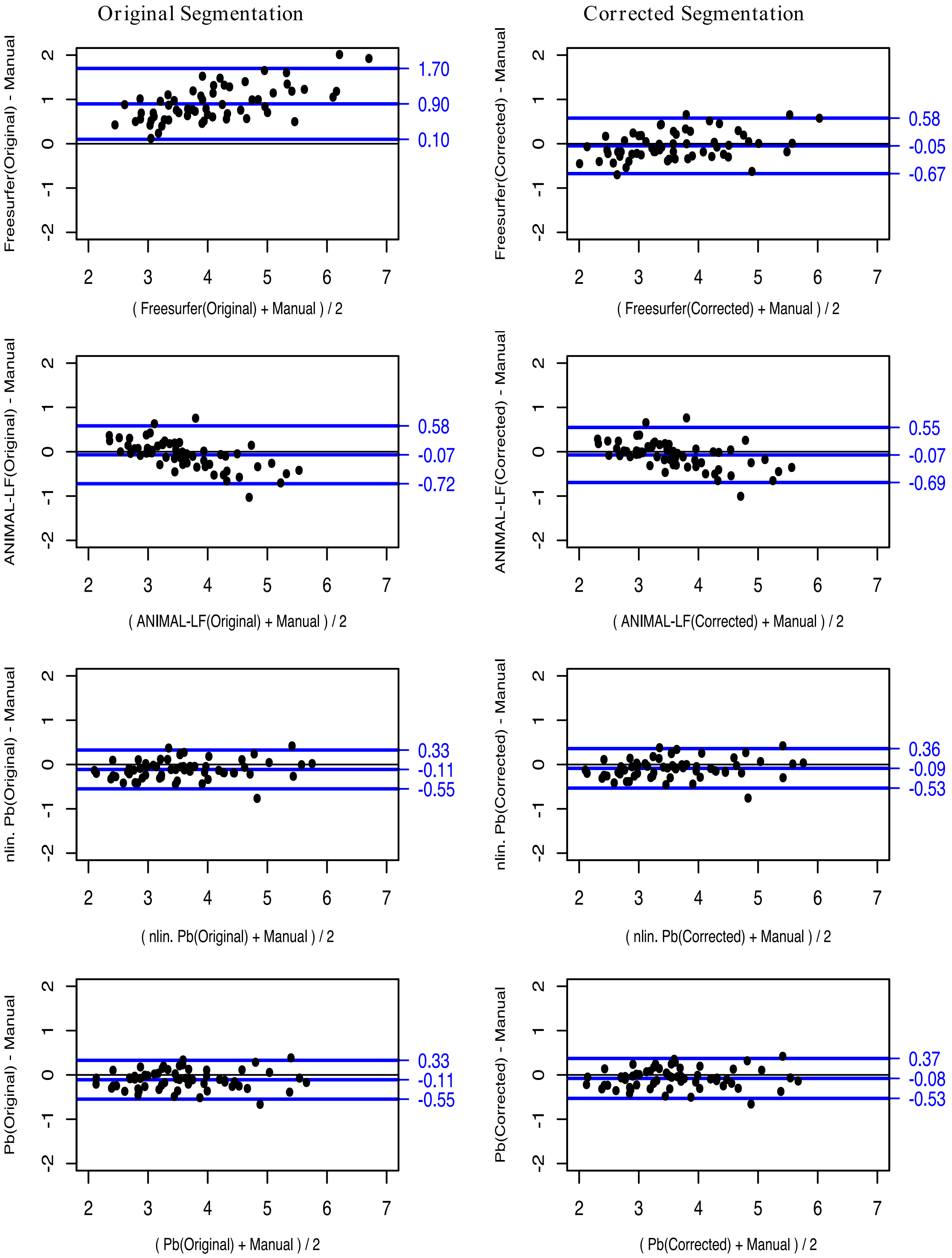

Fig. 3. Bland-Altman plots for volumetric comparison between mean hippocampal volume resulting from automatic segmentation methods and manual labeling. From top to bottom: FreeSurfer, ANIMAL with label fusion, nonlinear patch-based (nlin. Pb) and patch-based (Pb) segmentation. The left column shows the original method, and the right column represents the corrected version. Blue lines represent the 2.5% (bottom) and 97.5% (top) limit lines. All values are reported in cubic centimeters.

Discussion

Achieving a fair comparison of the accuracy of different automatic segmentation methods is not a trivial task. The quality of the MR images segmented, the segmentation protocol itself, inter- and intra-rater variability, and specificity of the participant groups can all affect the reported accuracy (Collins and Pruessner, 2010). For example, mean hippocampal volume decreases with age (Pruessner et al., 2001), which can affect overlap-based statistics like Dice's κ. In this study, we compared three different automatic hippocampal segmentation methods using the same testing dataset and manual segmentation protocol (Pruessner et al., 2000). We then compared the three methods to the

Table 5
Bland-Altman estimations for segmentation methods.

Method	mean	2.5% limit	97.5% limit	SD
FreeSurfer	0.8996	0.0999	1.6992	0.3998
FreeSurfer with EC	−0.0483	−0.6745	0.5778	0.3130
ANIMAL with label fusion	−0.0727	−0.7249	0.5795	0.3261
ANIMAL with label fusion with EC	−0.0689	−0.6903	0.5525	0.3107
Patch-based	−0.1087	−0.5468	0.3293	0.2190
Patch-based with EC	−0.0864	−0.5324	0.3595	0.2230
Nonlinear patch-based	−0.1112	−0.5491	0.3268	0.2190
Nonlinear patch-based with EC	−0.0789	−0.5295	0.3717	0.2253

Mean column shows mean value of differences; SD column shows standard deviation of differences.

Table 6
Experiment 2: quality control information. NC: normal controls; sMCI: stable mild cognitive impairment; pMCI: progressive mild cognitive impairment; AD: Alzheimer's disease.

	NC	sMCI	pMCI	AD
Total number of subjects	231	240	168	199
Number of failures for FreeSurfer	33	13	8	44
Number of failures ANIMAL with label fusion	1	2	1	0
Number of failures Patch-based	0	0	0	0
Number of failures nlin. Patch-based	3	2	2	1

Table 7
Experiment 2: dataset information. NC: normal controls; sMCI: stable mild cognitive impairment; pMCI: progressive mild cognitive impairment; AD: Alzheimer's Disease; MMSE: Mini-Mental State Examination.

	NC	sMCI	pMCI	AD
Number	178	215	152	135
Median age at baseline (yrs)	76.14	75.17	75.69	75.44
Sex: female (%)	44%	32%	39%	47%
Median MMSE score	29	28	26	23

Table 8
Cohen's *d* effect size mean and (standard deviation) for NC/AD group difference on 200 bootstrapped replicates.

Method	Mean between group effect size
Freesurfer	1.4886 (0.1544)
Freesurfer with EC	1.6921 (0.1630)
ANIMAL with label fusion	1.3228 (0.1400)
ANIMAL with label fusion with EC	1.3765 (0.1502)
Patch-based	1.5793 (0.1696)
Patch-based with EC	1.5104 (0.1472)
Nonlinear patch-based	1.5076 (0.1522)
Nonlinear patch-based with EC	1.5134 (0.1468)

Table 9
Cohen's *d* effect size mean and (standard deviation) for sMCI/pMCI group difference on 200 bootstrapped replicates.

Method	Mean between group effect size
Freesurfer	0.5919 (0.1044)
Freesurfer with EC	0.6683 (0.1049)
ANIMAL with label fusion	0.4565 (0.1025)
ANIMAL with label fusion with EC	0.5040 (0.1076)
Patch-based	0.5988 (0.1046)
Patch-based with EC	0.5450 (0.1147)
Nonlinear patch-based	0.5551 (0.1066)
Nonlinear patch-based with EC	0.5481 (0.0973)

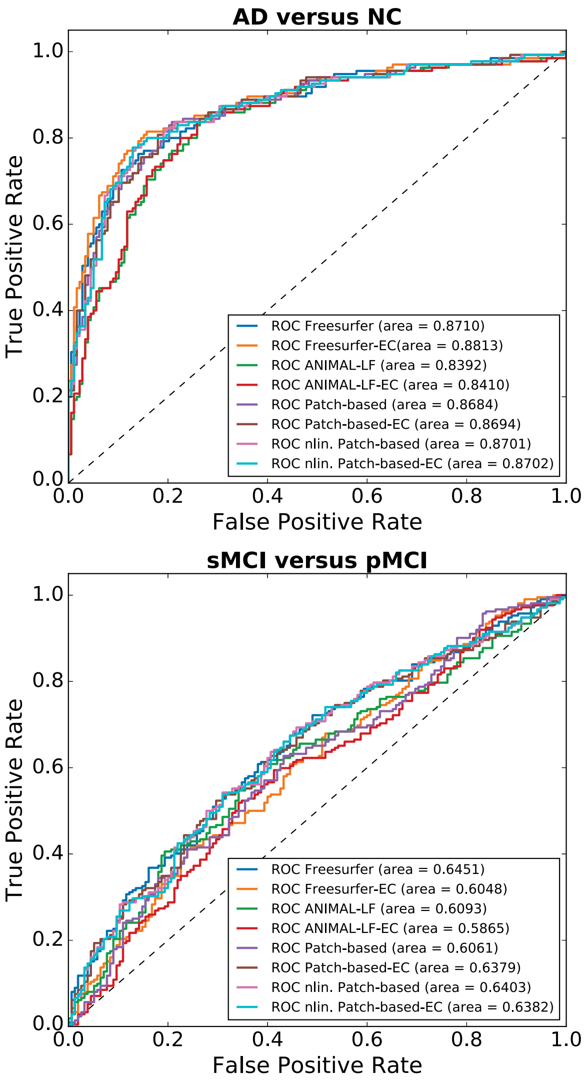

Fig. 4. Receiver operating characteristic (ROC) curve shows the performance of the LDA classifier with LOO cross-validation using volumes derived by each method, age, and sex as features. Left panel shows ROC curves for AD versus NC; right panel shows ROC curves for sMCI versus pMCI.

Table 10
Area under the ROC curve values using LDA with LOO cross-validation based on mean hippocampal volume, age, and sex as features.

Method	AUC (AD/NC)	AUC (sMCI/pMCI)
FreeSurfer	0.8710	0.6451
FreeSurfer with EC	0.8813	0.6048
ANIMAL with label fusion	0.8392	0.6093
ANIMAL with label fusion with EC	0.8410	0.5865
Patch-based	0.8684	0.6061
Patch-based with EC	0.8694	0.6379
Nonlinear patch-based	0.8701	0.6403
Nonlinear patch-based with EC	0.8702	0.6382

hippocampal segmentation of the well-known FreeSurfer tool (Fischl et al., 2002). We also applied an error correction technique to each automatic segmentation method using the same library of manual labels for correction (Wang et al., 2011).

Dice's κ metric is a measure of volumetric overlap, which means that it can be affected by both registration and segmentation errors. In our comparison, the nonlinear patch-based method with error correction showed the best median κ value. The use of nonlinear registration enables better alignment than that accomplished by the linear patch-based method. The intensity-based patch-based method shows promising results in its ability to capture structural variability (Coupé et al., 2011), and applying the wrapper error correction method (Wang et al., 2011) further improves the segmentation. The other methods show proper alignment with manual segmentation as well. The results showed that the key factor affecting higher κ values is registration. However, the linear patch-based method, despite its use of simple linear alignment of scans to the template, shows acceptable results in terms of the Dice's κ metric. The FreeSurfer method does not use the population-based specific template. Its probabilistic template is based on a young, healthy population and thus is perhaps not a fair representation of the test database under study here. This distinction partially explains the oversegmentation pattern in the FreeSurfer segmentation: We observed that FreeSurfer oversegments larger hippocampi more than it does smaller ones, whereas ANIMAL with label fusion undersegments the larger hippocampi while oversegmenting the smaller ones. FreeSurfer's pattern of oversegmentation has been observed in Pipitone et al. (2014) as well. As FreeSurfer uses its own atlas to define the hippocampus, it is not surprising that it differs from the segmentation derived from Pruessner's protocol. The error correction method could successfully correct the bias in FreeSurfer results; yet, the variance pattern persists. Considering the bias toward oversegmentation observed in FreeSurfer, a wrapper error correction technique such as that used in this paper (Wang et al., 2011) is suggested to render the segmentation compatible with the population under study. Specifically, this addition becomes critically important when the population has anatomical characteristics that differ from the FreeSurfer template, which is derived from young, healthy adult population.

The ICC ratio, as a measure of volume similarity between the gold standard (manual segmentation) and each automatic segmentation, shows the benefit of the patch-based intensity-driven strategy in segmentation. Considering both the ICC and κ values, the results demonstrate that fine nonlinear registration combined with the patch-based strategy of segmentation is highly successful. A similar strategy won the MICCAI 2012 segmentation challenge (Wang and Yushkevich, 2013). Fig. 2 shows the improvement for the patch-based methods from linear to nonlinear registration; the algorithm is further enhanced when the error correction strategy is applied.

The error correction strategy improves segmentation accuracy (in terms of agreement with manual labeling) for all four methods. Since FreeSurfer has the largest bias in comparison with the other automatic methods, it benefits the most from the bias correction technique; the improvement for all the other methods is slight. The methods already work well in accordance with manual segmentation and show only a very small bias before applying the correction. Furthermore, the variance of the methods does not improve even after applying the correction (see Fig. 3). A similar pattern was observed in the original paper (Wang et al., 2011). Error correction performance is highly sensitive to its hyperparameters, such as the search boundary or the patch size around each voxel. For the sake of harmonization, in this study, we used the default parameters from the original paper (Wang et al., 2011). It is possible to further tune the error correction method for each individual method separately, but such adjustments are out of the scope of this paper.

We computed Cohen's d effect size as a measure of how well each method differentiates between different clinical groups. Our study shows that Cohen's d is large for all the automated segmentation methods for AD/NC between-group analysis. That is, all eight methods can detect the hippocampal volume reduction for the AD group compared with the NC group. This observation confirms the idea that the hippocampus is affected by the disease neurodegeneration pathway.

With the prognosis experiment, we attempted to differentiate between sMCI and pMCI clinical groups. All methods showed medium effect sizes in this experimental setting; that is, the between-group mean difference is much smaller compared with the difference between the AD and NC groups. Considering the huge impact early detection of AD can have for treatment, this study has important clinical implications. The results of this study aligns with previous work (Sperling et al., 2011) showing that hippocampal atrophy on volumetric MR images is informative in separating the different clinical groups.

We further used a linear classifier to classify the subjects to AD or NC groups (diagnosis study) and to sMCI or pMCI groups (prognosis study). The methods are compared based on the area under the ROC curve (AUC), and results show that the methods do not differ in terms of classification performance in either the diagnosis or prognosis experiments.

FreeSurfer has been shown to be fairly successful in previously published clinical studies. Considering the quality of FreeSurfer segmentations, we hypothesize that this difference can be explained by one or a combination of the following points. First, as observed in Fig. 2, FreeSurfer has the tendency to bias the segmentation by oversegmenting the volume in general, with larger volumes showing a greater degree of oversegmentation in relation to smaller volumes. Second, oversegmentation in FreeSurfer is partially explained by the use of a different training population. We expect larger hippocampal volumes for a young, healthy population, which is exactly the population from which FreeSurfer's template is derived. While this may explain the oversegmentation tendency due to differences in MR image intensities, it does not explain the bias toward oversegmenting larger hippocampi. FreeSurfer tends to oversegment larger hippocampi more so than smaller ones, which has led to stronger group differences for disease studies. The pattern is still observable after error correction is applied.

Recently, an effort has been made to harmonize different protocols for manual hippocampal segmentation (Boccardi et al., 2011, 2013). The harmonized protocol (HarP) is highly similar to the manual segmentation protocol by Pruessner, which is used in this manuscript (Pruessner et al., 2000). We have already run a comparison between the HarP and Pruessner's protocol (Zandifar et al., 2015). The result showed no significant difference in terms of the effect size between clinical groups using these two protocols. However, segmentation accuracy measures (e.g., Dice's kappa) are higher for Pruessner's protocol. We hypothesized that the higher values could be in part due to the protocol's use of three-dimensional tracing versus two-dimensional tracing for HarP labels, where boundary smoothness may not be ensured (Zandifar et al., 2015). A detailed comparison between the two manual segmentation protocols would be worthwhile and should be considered in the future.

Limitations

We faced some limitations in conducting this study. First, FreeSurfer was used with its default settings, which may bias the results against it. We applied the error correction method to partially correct for the bias due to different training datasets. Although we included only the subjects that passed quality control for all processing pipelines, it is important to note that the experiments (AD:NC and sMCI:pMCI) are biased towards FreeSurfer. Indeed, for the 101 (=31+24+45) subjects that failed the FreeSurfer pipeline (see Table 6), it was not possible to estimate hippocampal volume. That is, greater than 12% of the test set should receive a random classification (or be considered missing data) in a forced-choice design, which would greatly reduce the advantage when using FreeSurfer hippocampal volumes for both AD:NC and sMCI:pMCI experiments.

For the second experiment, it would have been worthwhile to compare the performance of the automatic methods with manual segmentations. However, because manually segmenting the hippocampus on the entire ADNI dataset would be such a lengthy and laborious undertaking, we limited the comparison to the automatic methods.

Conclusion

We conclude that the four automatic methods tested all show acceptable conformity with manual segmentation. Nonlinear registration is the key factor to obtaining good spatial alignment with the manual segmentation. The patch-based strategies perform well, with good correlation with manual segmentation. Our second experiment demonstrated that all four automatic methods can be used as a proper substitute for costly manual segmentation in a clinical setting. All the effect sizes for diagnosis setting were found to be large, while all methods compared showed a medium effect size in prognosis settings. Overall, the patch-based method enhanced with nonlinear registration and the error correction technique (Wang et al., 2011) shows the most promising results among the different methods conducted in our experiments.

Acknowledgments

This work was supported by grants from the Canadian Institutes of Health Research (MOP-111169), les Fonds de Research Santé Québec Pfizer Innovation fund (25262), and an NSERC CREATE Grant (4140438 - 2012). We would like to acknowledge funding from the Famille Louise & André Charron.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. 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 (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California.

Part of the computations were conducted on the supercomputer Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), ministére de l'Économie, de la Science et de l'Innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46, 726–738.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12, 26–41.

Bai, W., Shi, W., Ledig, C., Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac MR images. Med. Image Anal. 19, 98–109.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T.M.E.K., Schott, J.M., Frost, C., Schahill, R., Fox, N., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage 40, 1655–1671.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J., de Leone, M.J., deToledo Morrell, L., Killiany, R.J., L. S., Pantel, J., Pruessner, J., Soininen, H., Watson, C., Duchesne, S., Jack, C.R., Frisoni, G.B., 2011. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheimer's Dis., 26, 61–75.

Boccardi, M., Bocchetta, M., Ganzola, R., Robitaille, N., Redolfi, A., Duchesne, S., Jack, C. J., Frisoni, G.B., T.E.W.G., 2013. On the harmonized protocol for hippocampal volumetry, the Alzheimer's disease neuroimaging initiative, operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimer's Dement., 11, 184–194.

Braak, H., Braak, E., 1991. Neuropathological staging of Alzheimer's related changes. Acta Neuropathol. 82, 239–259.

Bremner, J.D., Randall, P., Vermetten, E., Staib, L., Bronen, R.A., Mazure, C., Capelli, S., McCarthy, G., Innis, R.B., Charney, D.S., 1997. Magnetic resonance imaging-based measurement of hippo campal volume in post traumatic stress disorder related to childhood physical and sexual abusea preliminary report. Biol. Psychiatry 41, 23–32.

Brookmeyer, R., Jahnson, E., Ziegler-Graham, K., Arrighi, H., 2007. Forecasting the global burden of Alzheimer's disease. Alzheimer's Dement. 3, 186–191.

Chupin, M., Mukuna Bantumbakulu, A.R., Hasboun, D., Bardinet, E., Baillet, S., Kinkingnhun, S., Lemieux, L., Dubois, B., Garnero, L., 2007. Automated segmentation of the hippocampus and the amygdala driven by competition and anatomical priors: method and validation on healthy subjects and patients with Alzheimer's disease. Neuroimage 34, 996–1019.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. The Alzheimer's disease neuroimaging initiative, fully automatic hippocampus segmentation and classification in alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579–587.

Collins, D.L., Evans, A.C., 1997. Animal: validation and applications of nonlinear registration-based segmentation. Int. J. Pattern Recognit. Artif. Intell. 11, 1271–1294.

Collins, D.L., Pruessner, J., 2010. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting animal with a template library and label fusion. NeuroImage 52, 1355–1366.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3d intersubject registration of MR volumetric data in standardized talairach space. J. Comput. Assist. Tomogr. 18, 192–205.

Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-d model-based neuroanatomical segmentation. Hum. Brain Mapp. 3, 190–208.

Coupé, P., Eskildsen, S.F., Manjn, J.V., Fonov, V.S., Collins, D.L., 2012. The Alzheimer's disease neuroimaging initiative, simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. NeuroImage 59, 3736–3747.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. IEEE Trans. Med. Imaging 27, 425–441.

Coupé, J.V., Manjn, P., Fonov, V., Pruessner, J., Robles, D.L., Collins, M., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54, 940–954.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M., Chupin, M., Benali, H., Colliot, O., T., 2011. The Alzheimer's disease neuroimaging initiative, automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage, 56, 766–781.

Dill, V., Franco, A.R., Pinho, M.S., 2015. Automated methods for hippocampus segmentation: the evolution and a review of the state of the art. Neuroinformatics 13, 133–150.

Duchesne, S., Pruessner, J., Collins, D.L., 2002. Appearance-based segmentation of medial temporal lobe structures. NeuroImage 17, 515–531.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Fischl, B., 2012. Freesurfer. NeuroImage 62, 774–781.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327.

Fonov, V., Coupé, P., Eskildsen, S.F., Manjon Herrera, J.M., Collins, D.L., 2012. Multi-atlas labeling with population-specific template and non-local patch-based label fusion. In: MICCAI 2012 Workshop on Multi-Atlas Labeling, MICCAI, pp. 63–66.

Freund, Y., Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. Comput. Learn. theory, 23–37.

Guizard, N., Coupé, P., Fonov, V., Manjn, J.V., Arnold, D.L., Collins, D.L., 2015. Rotation-invariant multi-contrast non-local means for ms lesion segmentation. NeuroImage: Clin. 31, 376–389.

Heckemann, R.A., Hajnal, J., Aljabar, P., Ruckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33, 115–126.

Hu, S., Collins, D.L., 2007. Joint level-set shape modeling and appearance modeling for brain structure segmentation. NeuroImage 36, 672–683.

Hu, S., Coupé, P., Pruessner, J.C., Collins, D.L., 2011. Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. NeuroImage 58, 549–559.

Jack, C.R., Petersen, R.C., Xu, Y.C., Waring, S.C., O'Brien, P.C., Tangalos, E.G., Smith, G.E., Ivnik, R.J., Kokmen, E., 1997. Medial temporal atrophy on mri in normal aging and very mild alzheimer's disease. Neurology 49, 786–794.

Kempton, M.J., Salvador, Z., Munafo, M.R., Geddes, J.R., Simmons, A., Frangou, S., Williams, S.C.R., 2011. Structural neuroimaging studies in major depressive disorder: meta-analysis and comparison with bipolar disorder. Arch. Gen. Psychiatry 68, 675–690.

Morey, R.A., Petty, C.M., Xu, J.P., Hayes, Y., Wagner, H.R., Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, M.G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. NeuroImage 45, 855–866.

Nelson, M., Saykin, A., Flashman, L., Riordan, H., 1998. Hippocampal volume reduction in schizophrenia as assessed by magnetic resonance imaging: a meta-analytic study. Arch. Gen. Psychiatry 55, 433–440.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pipitone, J., Parka, M.M., Winterburna, J., Lett, T.A., Lerch, J.P., Pruessner, J.C., Lepage, M., Voineskosa, A.N., Chakravarty, M.M., 2014. The Alzheimer's disease neuroimaging initiative, multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. NeuroImage 101, 494–512.

Pruessner, J.C., Li, L., Serles, W., Pruessner, M., Collins, D.L., Kabani, N., Lupien, S., Evans, A., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10, 433–442.

Pruessner, J.C., Collins, D.L., Pruessner, M., Evans, A.C., 2001. Age and gender predict volume decline in the anterior and posterior hippocampus in early adulthood. J. Neurosci. 21, 194–200.

Rousseau, F., Habad, P., Studholm, C., 2011. A supervised patch-based approach for human brain labeling. IEEE Trans. Med. imaging 30, 1852–1862.

Roy, S., Carass, A., Prince, J.L., Pham, D.L., 2015. Longitudinal patch-based segmentation of multiple sclerosis white matter lesions. In: Zhou, L., Wang, L., Wang, Q., Shi, Y., (Eds.), Machine Learning in Medical Imaging: Proceedings of the 6th International Workshop, MLMI 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 5, 2015, Proceedings, Springer International Publishing, pp. 194–202.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity non uniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H., 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on Aging-Alzheimer's association workgroups on diagnostic. Alzheimer's Dement. 7, 280–292.

Tae, W.S., Kim, S.S., Lee, K.U., Nam, E., Kim, K.W., 2008. Validation of hippocampal volumes measured using a manual method and two automated methods (Freesurfer and IBASPM) in chronic major depressive disorder. Neuroradiology 50, 569–581.

Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D., Inatiative, T.A.D.N., 2013. Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage 76, 11–23.

Wang, H., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. Front. Neuroinform. 7, 27.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. NeuroImage 55, 968–985.

Wang, H., Poch, A., Takabe, M., Jackson, B., Gorman, J., Gorman, R., Yushkevich, P.A., 2013. Multi-atlas segmentation with robust label transfer and label fusion. In: Information Processing in Medical Imaging?: Proceedings of the Conference, IPMI, pp. 548–559.

Wieshmann, U.C., Woermann, F.G., Lemieux, L., Free, S.L., Bartlett, P.A., Smith, S.J.M., Duncan, J.S., Stevens, J.M., Shorvon, F.D., 1997. Development of hippocampal atrophy: a serial magnetic resonance imaging study in a patient who developed epilepsy after generalized status epilepticus. Epilepsia 38, 1238–1241.

Xiao, Y., Fonov, V.S., Beriault, S., Gerard, I., Sadikot, A.F., Pike, G.B., Collins, D.L., 2015. Patch-based label fusion segmentation of brainstem structures with dual-contrast MRI for Parkinson's disease. Int. J. Comput. Assist. Radiol. Surg. 10, 1029–1041.

Yang, J., Duncan, J.S., 2004. 3d Image segmentation of deformable objects with joint shape-intensity prior models using level sets. NeuroImage 8, 285–294.

Zandifar, A., Fonov, V., , P., Pruessner, J.C., C.D.L., 2015. A quantitative comparison between two manual hippocampal segmentation protocols. Alzheimer's Dement. 11, 67–68.

Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging 13, 716–724.