Erschienen in: NeuroImage. 2018, 181, pp. 142-148
https://doi.org/10.1016/j.neuroimage.2018.06.077

The EADC-ADNI harmonized protocol for hippocampal segmentation: A validation study

Azar Zandifar [a,b,*], Vladimir S. Fonov [a], Jens C. Pruessner [c,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] *McGill Centre for Studies in Aging, Faculty of Medicine, McGill University, Montreal, Canada*
[d] *University of Constance, Department of Psychology, Constance, Germany*

Keywords:
Hippocampal segmentation
EADC-ADNI harmonized protocol (the HarP)
Pruessner protocol
Alzheimer's disease
Dice's κ
Cohen's d
Area under receiver operating characteristic curve (AUC)

A B S T R A C T

Recently, a group of major international experts have completed a comprehensive effort to efficiently define a harmonized protocol for manual hippocampal segmentation that is optimized for Alzheimer's research (known as the EADC-ADNI Harmonized Protocol (the HarP)). This study compares the HarP with one of the widely used hippocampal segmentation protocols (Pruessner, 2000), based on a single automatic segmentation method trained separately with libraries made from each manual segmentation protocol. The automatic segmentation conformity with the corresponding manual segmentation and the ability to capture Alzheimer's disease related hippocampal atrophy on large datasets are measured to compare the manual protocols. In addition to the possibility of harmonizing different procedures of hippocampal segmentation, our results show that using the HarP, the automatic segmentation conformity with manual segmentation is also preserved (Dice's κ 0.88, κ 0.87 for Pruessner and HarP respectively (p 0.726 for common training library)). Furthermore, the results show that the HarP can capture the Alzheimer's disease related hippocampal volume differences in large datasets. The HarP-derived segmentation shows large effect size (Cohen's d 1.5883) in separating Alzheimer's Disease patients versus normal controls (AD:NC) and medium effect size (Cohen's d 0.5747) in separating stable versus progressive Mild Cognitively Impaired patients (sMCI:pMCI). Furthermore, the area under the ROC curve for a LDA classifier trained based on age, sex and HarP-derived hippocampal volume is 0.8858 for AD:NC, and for 0.6677 sMCI:pMCI. These results show that the harmonized protocol-derived labels can be widely used in clinic and research, as a sensitive and accurate way of delineating the hippocampus.

1. Introduction

Structural Magnetic Resonance Image (MRI)-derived estimates of hippocampal atrophy are considered one of the key supportive imaging markers for diagnosis of Alzheimer's Dementia (AD) (Dubois et al., 2007). Since manual delineation of the structure is a laborious, time consuming task, numerous methods have been developed to automatically and accurately segment the hippocampus (HC) (Chupin et al., 2007; Coupé et al., 2011; Collins and Pruessner, 2010). Manual segmentation is considered the gold standard, and automatic segmentation methods try to get as close as possible to the manual delineation (Dill et al., 2015). However, there are many different protocols for manual segmentation (Boccardi et al., 2011) that are based on different anatomical landmarks, and these can result in up to a two-fold difference in the volume of the hippocampus depending on the chosen protocol. Due to these differences, a fair comparison among different automatic segmentation methods that are based on different gold standards is virtually impossible (Frisoni and Jack, 2011).

* Corresponding author. McConnell Brain Imaging Centre, Montreal Neurological Institute, 3801 University Street, Room WB320, Montreal, QC, H3A 2B4, Canada.
E-mail addresses: azar.zandifar@mail.mcgill.ca (A. Zandifar), vladimir.fonov@mcgill.ca (V.S. Fonov), jens.pruessner@mcgill.ca, jens.pruessner@uni-konstanz.de (J.C. Pruessner), 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-ontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Recently, a group of major international experts on hippocampal segmentation have united to address the difficulties that arise from the lack of a standardized hippocampal segmentation protocol (Frisoni and Jack, 2011). The European Alzheimer's Disease Consortium (EADC), in collaboration with the Alzheimer's Disease Neuroimaging Initiative (ADNI), have completed a comprehensive effort to efficiently define a harmonized protocol for hippocampal segmentation that is optimized for Alzheimer's research, known as the EADC-ADNI Harmonized Hippocampal Protocol (the HarP). To initiate work on HarP, the most cited protocols in the literature were surveyed (Boccardi et al., 2011) and their authors were asked to certify the anatomic landmarks on one patient with Alzheimer's dementia and one normal control subject. These numerous landmark differences were then summarized into a well-defined number of assessable units (so called difference units). As the third step, the difference units are given to a panel of hippocampus experts to carry out an evidence-based Delphi procedure facilitating a consensual definition of the protocol. In addition to the actual segmentation units and their 3D renderings, the Delphi panelist were presented with the unit volume relative to total hippocampal volume, ICC of intra- and inter-rater reliability in the segmentation of each unit, and percent tissue reduction in mild cognitive impairment (MCI) and AD compared with controls. Therefore, the final protocol encompasses the hippocampal sub-units that are more sensitive to the AD-related pathology. Finally, a small group of tracers segmented a set of benchmark images based on the Harmonized protocol (Boccardi et al., 2013).

It has been shown that using the harmonized protocol significantly increases intra-rater consistency compared to the individual local protocols (Frisoni et al., 2015). Furthermore, the hippocampus volume when segmented with the HarP is found to be highly correlated with Braak and Braak staging in AD, tau, Aβ burden, and neuronal count, which demonstrates that the HarP successfully captured AD-related pathologies (Apostolova et al., 2015). The resulting HarP is expected to become the standard segmentation method for hippocampal volumetry in diagnostic studies, clinical trials, and algorithm validation efforts (Boccardi et al., 2011; Frisoni and Jack, 2011; Jack et al., 2011). However, considering the recency of the HarP, there is a need for validation studies. Importantly, the ability of the HarP labels to capture AD-related atrophy patterns needs to be further validated in a large cohort of subjects.

In this study, our goal is to demonstrate that the manual HarP protocol labels can be used to achieve automatic segmentations with accuracies at least as good as previously published techniques. We therefore compare HarP-based automatic hippocampal segmentations with automatic segmentations based on the Pruessner protocol (Pruessner, 2000) since the latter is widely used, widely cited and is among those protocols surveyed in designing the HarP (Boccardi et al., 2011). It has been previously shown that the automatic segmentation methods based on Pruessner protocol show promising performance when applied in AD populations (Collins and Pruessner, 2010; Pipitone et al., 2014; Zandifar et al, 2014, 2017). Furthermore, we had access to a reliable dataset of manual segmentations of hippocampus from ADNI data using both protocols.

This study consists of two experiments. Given that the HarP label volumes were manually delineated on 2D coronal slices and then stacked together in 3D, it is not clear if slice-to-slice inconsistencies will adversely affect automatic segmentation. Our primary goal in this study is to test if HARP labels can be used to achieve consistent, robust and accurate automatic segmentation of the HC. Therefore, in our first experiment, we evaluate the accuracy of our automatic hippocampus segmentation method (Zandifar et al., 2017; Fonov et al., 2011), using the two different labeling strategies (the HarP and Pruessner) as template libraries, using a leave-one-out cross validation strategy. The automatic segmentation method used in the study is one of the most accurate multi-label segmentation methods in the field (Zandifar et al., 2017). Our previous study shows that the method shows high similarity with the manual training library (κ values of 0.887 and 0.885 for left and right respectively) (Zandifar et al., 2017). In this method, the target image is non-linearly registered to the templates, and the corresponding label is assigned using non-local patch-based label fusion (Fonov et al., 2011). The label of the similar patches from the training library is weighted based on their intensity-wise similarity to the target patch to define the label for the central voxel of a target patch (Fonov et al., 2011). In this experiment, the automatic segmentation will be used as a surrogate to show the accuracy, consistency and similarity in the manual training libraries. Furthermore, since the goal of HarP was to arrive at a segmentation protocol that maximizes the NC:AD difference, our second goal is to demonstrate that automatic segmentations using HarP labels yielded a NC:AD difference that was at least as big as our previously-known accurate segmentation. Manual segmentations based on the two protocols are available only on a relatively small dataset making it difficult to accurately estimate the NC:AD difference. A direct comprehensive comparison of hundreds of manually segmented datasets is not feasible. Therefore, as a proxy of manual segmentation, in the second experiment, all ADNI-1 1.5T MRIs are segmented using our automatic library-based segmentation method twice; once using a HarP-based training library and once using the Pruessner protocol-based training library, using both common and different datasets to train the method. We then compare their effect size (Cohen's d) and area under the receiver operator curve (ROC) to differentiate patients with Alzheimer's from age and sex-matched normal controls. We also compare HarP-based segmentations with Pruessner protocol-based segmentations to differentiate subjects with progressive mild cognitive impairment (pMCI) with stable mild cognitive impairment (sMCI).

2. Methods

2.1. Datasets

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 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 ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) (For up-to-date information visit http://www.adni-info.org/).

To validate the segmentation accuracy for the two hippocampal segmentation protocols in the first experiment, we used three different subsets of images from ADNI data. The first subset consists of 60 ADNI-1 baselines 1.5 T images. The PI of the Pruessner protocol provided expert hippocampal labels (Pipitone et al., 2014). The dataset contains equal number of subjects in three clinical group (NC (n = 20), MCI (n = 20), and AD (n = 20)), and the different groups were comparable in terms of age, sex and median years of education (see Table 1). Hereafter, we refer to this dataset (MRI volumes and manual labels) as the Pruessner-60 dataset. The second subset contains 100 MR images selected by the EADC-ADNI group from both ADNI-1 and ADNI-2 study (Boccardi et al., 2015). The images were selected from both 1.5 T and 3T scans, and balanced by magnet field strength, scanner manufacturer, diagnosis, qualitative medial temporal atrophy (MTA) severity, sex and age ranges (Boccardi et al., 2015). Hereafter, we refer to this dataset as the HarP-100 dataset. The third subset consists of 13 MRI volumes that were common to both the Pruessner and HarP datasets, thus enabling a direct comparison between the two segmentation protocols. These will be named the Pruessner-13 and Harp-13. The demographic information for the Pruessner-60 and HarP-100 datasets can be found in Table 1.

In the second experiment, the automatic segmentation methods using (i) the Pruessner-60 training library and (ii) the HarP-100 protocol training library are tested over all baseline 1.5 T datasets from the ADNI-

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

Diagnostic group	Pruessner-60 Dataset				HarP-100 Dataset			
	NC	MCI	AD	Combined	NC	MCI	AD	Combined
Number	20	20	20	60	30	32	38	100
Median Age at Baseline	75.5	75.6	74.9	75.2	73.4	73.9	72.1	73.4
Sex: Female (%)	50	50	50	50	43	41	47	44
Median Education (yrs)	16.0	16.0	15.5	16.0	18.0	16.0	16.0	16.0
Median MMSE score	29.5	27.5	23.0	27.0	29.0	28.0	23.0	27.0

1 study (see Table 2). The ADNI-1 1.5T dataset is labeled based on clinical state of the subjects as identified in 2011. Patients with Alzheimer's dementia were labeled as the AD group. The MCI subjects who maintained their clinical state after 3 years were labeled as stable MCI (sMCI), while the subjects who progressed to dementia within 3y after the baseline scan received the progressive MCI (pMCI) label. Finally, the cognitively healthy subjects were labeled as normal controls (NC). The demographic information of each group can be found in Table 2. Since the second experiment outcome may be affected by the test dataset examples, we used an identical cohort to test both segmentation protocols. In addition, the test dataset for the second experiment is further refined so as not to include subjects from either the HarP-100 or the Pruessner-60 library, nor the preprocessing failures (4 failures). The detailed description of the final test dataset used for the second experiment can be found in the result section (Table 3).

2.2. Preprocessing and manual delineation

The Pruessner-60 and Pruessner-13 datasets went through a preprocessing pipeline before manual delineation (Pipitone et al., 2014); including denoising (Coupe 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 1×1 x 1 mm^3 resolution (Collins et al., 1994). After preprocessing, all images were coarsely aligned, and image intensities were normalized within each image and among the whole database (Fonov et al., 2011; Coupé et al., 2012; Zandifar et al., 2017). The images then went through manual delineation. Inter-rater reliability was 0.94 for right HC and 0.86 for left HC. Intra-rater reliability was 0.91 for right HC and 0.94 for left HC for manual labels (Pruessner, 2000).

On the HarP-100 and HarP-13 datasets, a single preprocessing step of rigid registration was applied to align the images along the AC-PC line before manual delineation (Boccardi et al., 2015). The HarP manual segmentation protocol was applied on 2D coronal slices using MultiTracer 1.0.[2] The Intra-class Correlation Coefficient (ICC) values across five raters were 0.97 for left HC and 0.99 for right HC. A complete overview of the manual segmentation steps can be found in (Boccardi et al., 2015). For sake of harmonization in the comparison process, the HarP dataset images were passed through the same preprocessing pipeline as the Pruessner dataset images after manual delineation, and the resulting transformation is applied on the manual labels with nearest neighbor resampling to keep all processing in the same space.

2.3. Automatic segmentation method

To address the time-requirements and inter- and intra-observer variability associated with manual segmentation, we used our automatic hippocampus segmentation method (Zandifar et al, 2014, 2017; Fonov et al., 2011) to estimate the volume of the hippocampus in all subjects. The multi-atlas segmentation procedure uses a population-specific atlas with patch-based label fusion to automatically label the hippocampus in T1-weighted MRI volumes (Zandifar et al., 2017; Fonov et al., 2011; Coupé et al., 2012). The procedure requires a library of MRI templates and their associated hippocampal labels. Here, four sets of labels are used – one from HarP-100 and one from Pruessner-60, also from both HarP-13 and Pruessner-13. Each new dataset to be segmented is subjected to the preprocessing described above. Afterwards, it is nonlinearly warped to the average template space and the patch-based segmentation algorithm is applied in the template space. This segmentation method has been shown to have Dice's κ as high as 0.887, showing significant overlap with manual segmentation (Zandifar et al., 2017).

Table 2
ADNI Dataset information. NC: normal controls; MCI: mild cognitive impairment; AD: Alzheimer's disease; MMSE: Mini-Mental State Examination.

Diagnostic group	NC	sMCI	pMCI	AD	Combined
Number	231	240	168	199	838
Median age at baseline (yrs)	75.93	75.68	75.14	76.04	75.63
Sex: Female (%)	48	33	39	49	42
Median Education (yrs)	16.0	16.0	15.0	16.0	16.0
Median MMSE score	29.0	27.5	26.0	23.0	27.0

Table 3
Dataset information for experiment 2. NC: normal control; sMCI: stable mild cognitive impairment; pMCI: progressive mild cognitive impairment; AD: Alzheimer's disease; MMSE: Mini-Mental State Examination.

Diagnostic group	NC	sMCI	pMCI	AD	Combined
Number	189	213	144	152	698
Median age at baseline	76.05	74.93	74.72	76.05	75.52
Sex: Female (%)	47	32	37	49	41
Median Education (yrs)	16.0	16.0	16.0	14.0	16.0
Median MMSE score	29.0	27.0	26.0	23.0	27.0

3. Metrics

3.1. Dice's κ metric

Volumetric overlap between the automatic segmentation method and manual segmentation is measured using the Dice's Kappa similarity index (Zijdenbos et al., 1994), which is computed as follows:

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

where V(.) is the volume operator, and M and A represent a set of manually and automatically labeled voxels respectively. The value of κ varies between 0 and 1, where 1 indicates the complete overlap labels. The Dice's κ similarity metric is computed comparing each automated segmentation method with the manual segmentation for both left and right hippocampi. For automatic segmentation, we used Leave One Out (LOO) cross-validation technique. That is, to segment one subject of the

[2] http://www.bmap.ucla.edu/portfolio/software/MultiTracer/.

training library, it is removed from the training set and the remaining subjects are used in the patch-based multi-atlas automatic segmentation process. The automatic segmentation is then compared to the manual segmentation for that subject. This process is repeated for all subjects of the training library.

3.2. Intra-class correlation (ICC)

We used the regression coefficient and the Intra-class Correlation Coefficient (ICC) to show the similarities between automatically and manually segmented volumes (Shrout and Fleiss, 1979). The automatic labels are derived using LOO cross validation as explained in the last subsection.

3.3. Cohen's d effect size

To investigate the sensitivity of each method in detecting between-group differences in a clinical setting, we computed the Cohen's d effect size based on the hippocampal volumes derived by automatic segmentation trained with each manual segmentation protocol. The Cohen's d effect size measures the distance between two normal distributions:

$$\text{Cohen's } d = \frac{m_1 \quad m_2}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

where m, and SD are the mean and standard deviation, respectively. Based on a conventional operational definition of Cohen's *d,* small, medium and large effect sizes are defined as $d < 0.5$, $0.5\ d < 0.8$, *and* $d > 0.8$, respectively.

3.4. Receiver Operating Characteristic (ROC) curve

We further trained and tested a Linear Discriminant Analysis (LDA) classifier to classify the subjects to different groups. We used a simple linear classifier in two different classification tasks. The first task is the classification of subjects to either AD or NC group, and the second task is to classify the subjects between sMCI and pMCI. For this problem, we fed the classifier with mean hippocampal volumes (averaged over left and right hippocampi) along with age and sex as features. We used the scikit-learn (Pedregosa et al., 2011) implementation of LDA, and all hyper parameters were set to their default values. The classification task was evaluated on ROC curves. We used a leave-one-out (LOO) strategy during classification validation to evaluate the performance using both the Pruessner-60 and HarP-100 template libraries.

3.5. Statistical analyses

All the statistical analyses are done using the R computing language (RStudio version 1.0.136). For the first experiment, we used two-way repeated ANOVA followed by a Wilcoxon test to compare the Dice distribution for each protocol-based segmentation. When comparing the absolute value of volume differences between manual and automatic segmentation for the protocols, a Wilcoxon test is used as well. In the second experiment, the distributions driven by the boot strapped replicates of the effect sizes are compared using a *t*-test, while classification performances are compared using the χ^2 Mc-Nemar test.

4. Results

4.1. First experiment

Segmentation accuracy was evaluated with a leave-one-out strategy using the Pruessner-13 and HarP-13 template libraries (Fig. 1, left panel). The Pruessner-13 library yields a median Dice coefficient (standard deviation) of 0.877 (0.039), and 0.867 (0.030) for left and right respectively. The HarP-13 library yields 0.862 (0.017) and 0.865 (0.014) for left and right hemispheres. A two-way repeated ANOVA (with the hemisphere and protocol as independent variables) shows that there are no significant differences between hemispheres or protocols ($p \quad value = 0.726, 0.256$ for protocols and hemispheres respectively).

We also compared the segmentation accuracy using LOO cross-validation technique using the full manual tracing libraries available (HarP-100 and Pruessner-60). The results show that both the methods show better performance with larger manual training library.

Fig. 1 (right panel) shows that the Pruessner protocol shows median Dice's coefficient (standard deviation) of 0.8885 (0.0218), and 0.8878 (0.0221) for left and right respectively, and these values for the HarP are 0.8748 (0.0201) and 0.8744 (0.0232). The differences between automatic and manual segmentation for each protocol qualitatively is shown for the best and the worst kappa values in the supplementary materials.

4.2. Volumetric correlation

The volumetric correlation between manual and automatic labels is measured using the ICC metric. We consider the raters (i.e. automatic and manual segmentations) as a fixed effect. The ICC values for left and right for HarP-100 protocol are the same at 0.96, while these values are 0.97 and 0.96 for Pruessner-60, respectively (Fig. 2). A Wilcoxon test shows that the difference in HC volume measured by automatic and manual segmentation is not significantly different between the methods ($W = 2031, p \quad value = 0.226, W = 2140, p \quad value = 0.251$ for left and right respectively).

4.3. Second experiment

In this experiment, we applied our automatic segmentation method on a large dataset of subjects (n = 838, Table 2) to compare the HarP-100 and Pruessner-60 template libraries in terms of their ability to capture the hippocampal atrophic pattern due to AD. We measured the distance between different AD clinical stages (i.e. AD:NC and sMCI:pMCI) using Cohen's *d* to measure the effect size, and the area under the ROC curve (AUC) derived from a classification task for the same problem.

4.4. Test dataset selection

We limited our analysis to ADNI-1 1.5T datasets for which both segmentation procedures ran successfully and excluded the datasets from the segmentation libraries (i.e., from the HarP-100 or Pruessner-60 datasets). The final number of 698 subjects was arrived at by using 838 ADNI-1 datasets and subtracting the 89 ADNI-1 subjects that were found in the HarP-100 template library, as well as the 60 ADNI-1 subjects that were in the Pruessner-60 template library and finally 4 pipeline failures. The dataset with 698 subjects consisted of 152 AD patients with Alzheimer's disease, 144 pMCI, 213 sMCI and 189 normal controls. Demographic information is shown in Table 3.

4.5. Hippocampal volume as an AD biomarker: effect size

To compute effect sizes, hippocampal volumes were corrected for age and sex using a method specifically designed for dementia and neurodegenerative studies (Dukart et al., 2011). The method trains a linear model only based on the healthy controls to regress out the effect of confounding factors such as age while keeping degeneration-induced atrophy. We followed the same strategy; with age as the linear term and sex as the offset coefficient. 200 bootstrapped replicates were used to obtain a more robust estimation of the effect size. The mean effect sizes (standard deviation) for AD:NC were 1.5883 (0.1450), and 1.5685 (0.1479) for the labels segmentated based on HarP-100 and Pruessner-60

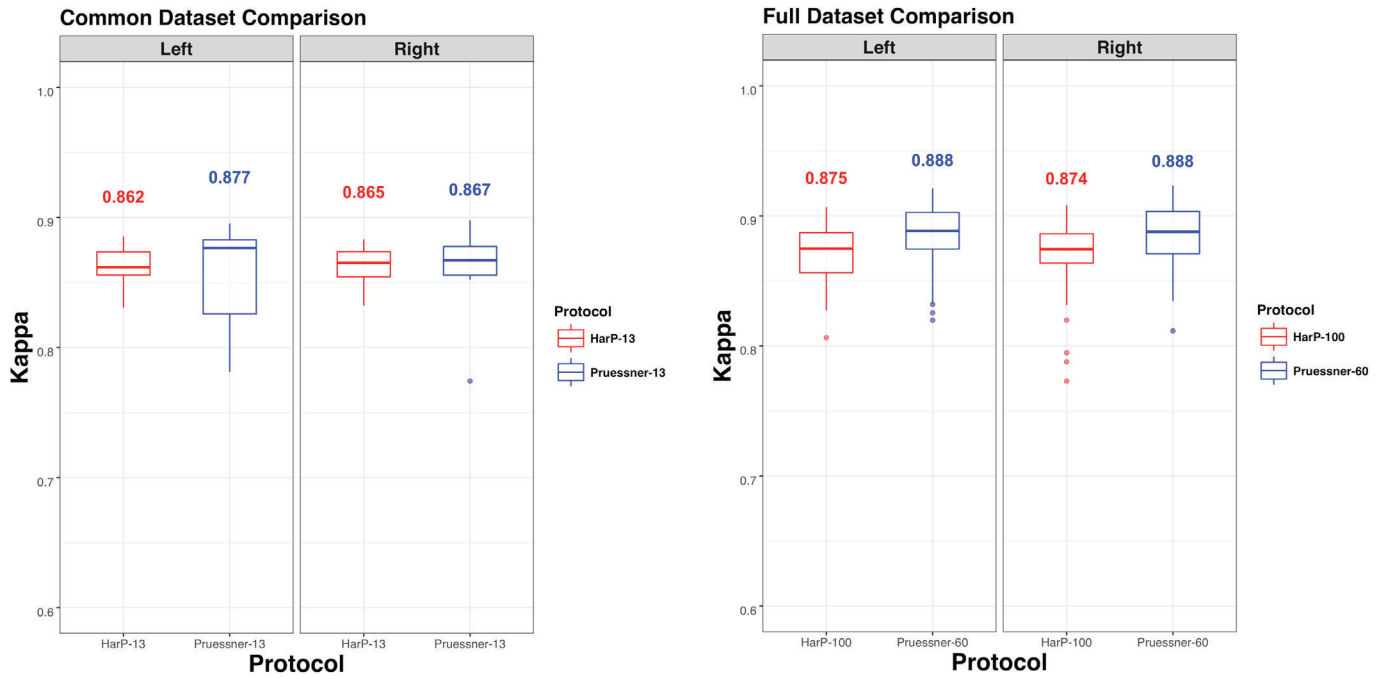

Fig. 1. Kappa distribution for left and right hippocampi. Left panel - segmentations based on HarP-13 (red) and Pruessner-13 (blue). Right panel - segmentations based on HarP-100 (red) and Pruessner-60 (blue).

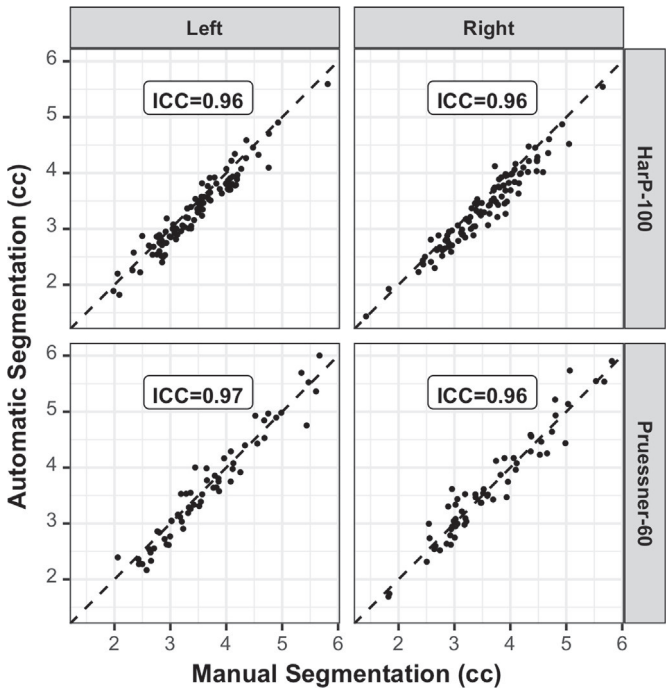

Fig. 2. Volumetric correlation of hippocampal segmentation and corresponding manual labeling. Dashed lines represent the unity line. From top to bottom, HarP-100 and Pruessner-60 template library segmentations. Volumes are reported in cubic centimeters.

template libraries respectively, while these values for sMCI:pMCI were 0.5747 (0.1026), and 0.5572 (0.1126). A pairwise *t*-test showed that there is no significant difference in effect size between the two hippocampal segmentation protocols ($p \quad value = 0.849$, 0.810 for AD:NC and sMCI:pMCI respectively). Therefore, both Pruessner-60 and HarP-100 libraries show large effect size for AD:NC, and medium effect size for sMCI:pMCI.

4.6. Hippocampal volume as an AD biomarker: ROC curve

Fig. 3 shows the ROC curve for AD:NC and sMCI:pMCI experiments. The area under the curve (AUC) for AD:NC experiment is 0.8858 and 0.8846, for HarP-100 and Pruessner-60 template libraries, respectively, while for sMCI:pMCI these values are 0.6677 and 0.6662. A McNemar test shows no significant difference between Pruessner-60 and HarP-100 libraries ($[\chi^2 = 0, \ p \quad value = 1]$, $[\chi^2 = 0, \ p \quad value = 1$ for AD:NC and sMCI:pMCI respectively). The experiment was also performed using left and right hippocampal volume as separate features, where no significant differences in the classification performance observed.

5. Discussion

Dice's κ metric is a measure of volumetric overlap, which demonstrates how well the automatic segmentation aligns with its corresponding manual segmentation. Our nonlinear patch-based segmentation is highly accurate in terms of volumetric overlap with the training library (Zandifar et al., 2017). In this study, the methods trained by both Pruessner and HarP training libraries showed high concordance with corresponding manual segmentations (see Figs. 1 and 2). The κ values are among the highest reported in the field (Dill et al., 2015; Zandifar et al., 2017). Since the automatic segmentation for each subject is done based on the rest of the training library, high volumetric overlap between each subject's automatic segmentation and its corresponding manual label shows the consistency of the training labels across the population. As expected, the κ values using the full training libraries are higher than those using the Pruessner-13 or Harp-13, presumably because there are more templates to choose from during template selection, thus better representing the anatomical variability of the population. The segmentation procedure using the Pruessner-60 protocol shows higher κ values in comparison with HarP-100 segmentations, however it is difficult to interpret this difference because the two training libraries have different subjects (other than the 13 in common). We note, however, that the Pruessner manual library is traced in 3D, while HarP training labels are delineated by the experts in 2D coronal images, and these were stacked together to form a 3D volume (Boccardi et al., 2015). This last step may impose some noise or jitter to the resulting reconstructed 3D labels which

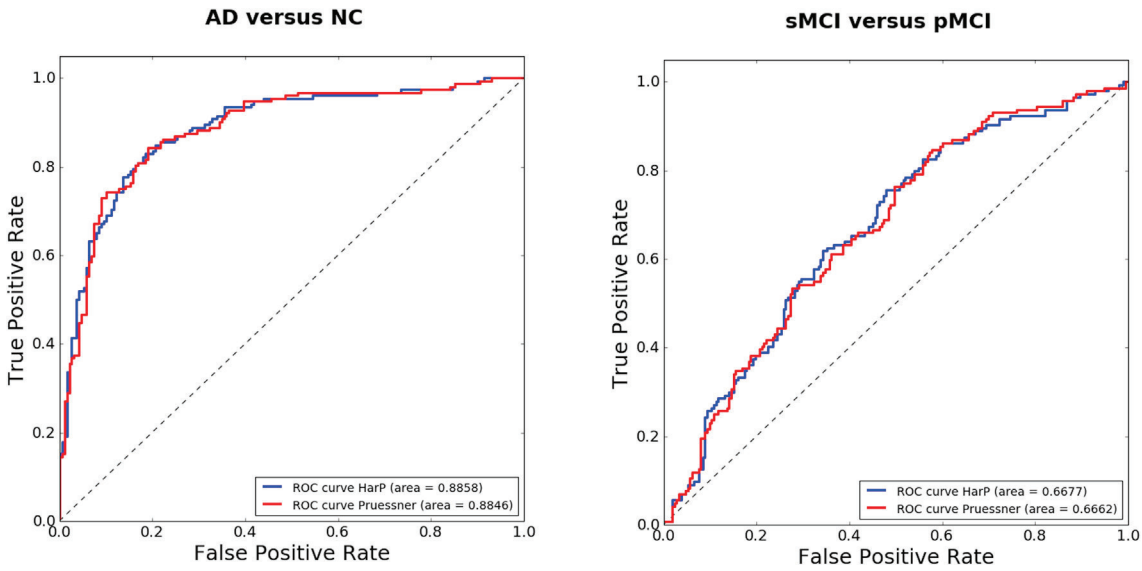

Fig. 3. Receiver Operating Characteristic (ROC) curve shows the performance of the LDA classifier with LOO cross-validation using mean hippocampal volumes (averaged over left and right hemispheres) derived by each method, age and sex as features. Left and right panel show performance in AD:NC and sMCI:pMCI task respectively.

were used in our experiment and may be a source of inconsistency in the HarP manual labels. This said, the correlation between volumes derived by automatic segmentation with either protocol is high ($ICC \cong 0.96$), and there is no significant difference between either method.

We compared Cohen's *d* effect size as a measure of how well each protocol can capture the AD-related pathologies. The larger the effect size between patient and normal groups, the better the protocol is in terms of capturing the AD-related hippocampal atrophy. We ran experiments on a large set of subjects from ADNI data two estimate the effect size between AD and normal controls, and stable MCI versus progressive MCI group. The results show that HarP-100 and Pruessner-60 template libraries are not statistically different from each other in their capacity to capture AD related pathology, and both methods reached a high effect size for AD:NC and medium effect size for sMCI:pMCI experiment.

We further ran a similar experiment to investigate the different patterns on a single-subject basis, rather than between group differences. We trained a classifier for both diagnosis (AD:NC) and prognosis (sMCI:pMCI) studies with performance measured by ROC curves. A post-hoc McNemar test shows that there is no significant difference between the methods in either experiment.

Eventually, considering the power of the HarP in capturing AD-related pathologies and harmonizing hippocampal segmentation procedure, it is suggested to approach hippocampal segmentation using a protocol similar to the HarP. The HarP is certified by the experts in the field, making the protocol a reliable alternative for other similar approaches such as the Pruessner protocol. Moreover, since our previous work (Zandifar et al., 2017) showed that the error correction technique presented in (Wang et al., 2011) could improve the results of automatic segmentation, we suggest that different automatic segmentation procedures may benefit from a similar post-hoc correction procedure to make the segmentations closer to the HarP and thus more comparable with all the different groups who accepted to use the HarP as a reliable hippocampal segmentation protocol.

Our study is not without limitations. We used pre-existing labels for both Harp and Pruessner template libraries. Although all datasets were taken from the ADNI study and have the same age range and followed the same acquisition protocols, only 13 subjects were common to both the HarP and Pruessner template libraries. While this enabled us to make a direct comparison between protocols, the relatively small number of subjects somewhat limits our power to detect any difference and increases our chances of a Type 2 error. In addition, different raters traced the two sets, and the tracing interface differed. These factors might have partially affected the results of the study. However, considering that manual segmentation is a laborious time-consuming task, we could not justify re-segmenting a new set of templates from the ADNI study using both protocols when we had access to labels already created by experts in hippocampal manual segmentation using the HarP and Pruessner protocols.

6. Conclusion

We conclude that the HarP shows promising results when used on a large multi-site multi-scanner dataset. Automatic segmentation conformity with manual segmentation is preserved, while offering the possibility to harmonize different procedures of hippocampal segmentation and compare them based on a unique set of labeling. Furthermore, HarP-derived segmentations show high effect size in separating AD:NC and medium effect-size in separating sMCI:pMCI. These results show that the harmonized protocol-derived labels can be widely used in clinic and research, as a sensitive and accurate way of delineating the hippocampus. Furthermore, considering the experts' consensual support of the HarP, we suggest using the HarP for hippocampal segmentation in future studies.

Acknowledgment

This work was supported by grants from the Canadian Institutes of Health Research (MOP-111169), les Fonds de Research Santé Quebec Pfizer Innovation fund, 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 (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 (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

Apostolova, L.G., et al., 2015. Relationship between hippocampal atrophy and neuropathology markers: a 7T MRI validation study of the EADC-ADNI Harmonized Hippocampal Segmentation Protocol. Alzheim. Dement 11 (2), 139 150.

Boccardi, M., et al., 2011. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. J. Alzheim. Dis. : JAD 26 (3), 61 75.

Boccardi, M., et al., 2013. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimer's Dement; J Alzheimer's Assoc.

Boccardi, M., et al., 2015. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheim. Dement 11 (2), 175 183.

Chupin, M., et al., 2007. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with Alzheimer's disease. Neuroimage 34 (3), 996 1019.

Collins, D.L., Pruessner, J.C., 2010. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 52 (4), 1355 1366.

Collins, D.L., et al., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 18 (2), 192 205.

Coupé, P., et al., 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imag. 27 (4), 425 441.

Coupé, P., et al., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54 (2), 940 954.

Coupé, P., et al., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. Neuroimage: Clin 1 (1), 141 152.

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 (2), 133 150.

Dubois, B., et al., 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS ADRDA criteria. Lancet Neurol. 6 (8), 734 746.

Dukart, J., et al., 2011. Age correction in dementia matching to a healthy brain. PLoS One 6 (7), e22193.

Fonov, V., et al., 2011. Multi-atlas labeling with population-specific template and non-local patch-based label fusion. In: MICCAI 2012 Workshop on Multi-atlas Labeling, pp. 163 166.

Frisoni, G.B., Jack, C.R., 2011. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimer's Dement; J Alzheimer's Assoc. 7 (2), 171 174.

Frisoni, G.B., et al., 2015. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheim. Dement 11 (2), 111 125.

Jack Jr., C.R., et al., 2011. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheim. Dement 7 (4), 474 485 e4.

Pedregosa, F., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (Oct), 2825 2830.

Pipitone, J., et al., 2014. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. Neuroimage.

Pruessner, J.C., 2000. Volumetry of Hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cerebr. Cortex 10 (4), 433 442.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420 428.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17 (1), 87 97.

Wang, H., et al., 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 (3), 968 985.

Zandifar, A., et al., 2014. A unified assessment of fully automated Hippocampus segmentation methods. Alzheimer's Dementia 10 (4), P86.

Zandifar, A., et al., 2017. A comparison of accurate automatic hippocampal segmentation methods. Neuroimage 155, 383 393.

Zijdenbos, A.P., et al., 1994. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imag. 13 (4), 716 724.