Neuroinformatics (2019) 17:43–61
https://doi.org/10.1007/s12021-018-9380-2

ORIGINAL ARTICLE

# Longitudinal Neuroimaging Hippocampal Markers for Diagnosing Alzheimer's Disease

Carlos Platero[1] · Lin Lin[2] · M. Carmen Tobar[1]

Published online: 21 May 2018

**Abstract**
Hippocampal atrophy measures from magnetic resonance imaging (MRI) are powerful tools for monitoring Alzheimer's disease (AD) progression. In this paper, we introduce a longitudinal image analysis framework based on robust registration and simultaneous hippocampal segmentation and longitudinal marker classification of brain MRI of an arbitrary number of time points. The framework comprises two innovative parts: a longitudinal segmentation and a longitudinal classification step. The results show that both steps of the longitudinal pipeline improved the reliability and the accuracy of the discrimination between clinical groups. We introduce a novel approach to the joint segmentation of the hippocampus across multiple time points; this approach is based on graph cuts of longitudinal MRI scans with constraints on hippocampal atrophy and supported by atlases. Furthermore, we use linear mixed effect (LME) modeling for differential diagnosis between clinical groups. The classifiers are trained from the average residue between the longitudinal marker of the subjects and the LME model. In our experiments, we analyzed MRI-derived longitudinal hippocampal markers from two publicly available datasets (Alzheimer's Disease Neuroimaging Initiative, ADNI and Minimal Interval Resonance Imaging in Alzheimer's Disease, MIRIAD). In test/retest reliability experiments, the proposed method yielded lower volume errors and significantly higher dice overlaps than the cross-sectional approach (volume errors: 1.55% vs 0.8%; dice overlaps: 0.945 vs 0.975). To diagnose AD, the discrimination ability of our proposal gave an area under the receiver operating characteristic (ROC) curve (AUC) = 0.947 for the control vs AD, AUC = 0.720 for mild cognitive impairment (MCI) vs AD, and AUC = 0.805 for the control vs MCI.

**Keywords** Alzheimer's disease · MRI · Hippocampal segmentation · Longitudinal analysis

## Introduction

The analysis of the hippocampus is a very active field of research because it is one of the first structures where early Alzheimer's disease (AD) pathology is observed (Jack et al. 2004; Schröder and Pantel 2016). T1-weighted (T1w) magnetic resonance imaging (MRI) measurements of the hippocampus (e.g. volume) have become useful markers in the diagnosis of AD and prodromal AD, such as mild cognitive impairment (MCI) (Dubois et al. 2007; Frisoni et al. 2010). Several published research criteria recommend the use of hippocampal atrophy from MRI to improve the differential diagnosis between AD or MCI in patients (Albert et al. 2011; McKhann et al. 2011). Many automatic approaches for extracting the hippocampal structures from brain MRI have been proposed (Louis Collins and Pruessner 2010; Lotjonen et al. 2010; Leung et al. 2010; Coupé et al. 2011; Wang et al. 2013). Among such approaches, atlas-based methods have been demonstrated to outperform other algorithms (Nestor et al. 2013) that rely on manual segmentations. However, the hippocampus is a complex anatomical structure, and different manual segmentation protocols have been proposed. An international effort to harmonize existing protocols has defined the harmonized hippocampal protocol (HarP) (Boccardi et al. 2011; Frisoni et al. 2015). This protocol has proven to be very reliable and to provide a hippocampal segmentation estimate that can be considered as a standard measure, enabling the use of the hippocampal measures as proper markers for AD and MCI. Nevertheless,

✉ Carlos Platero
carlos.platero@upm.es

[1] Health Science Technology Group, Universidad Politécnica de Madrid, Ronda de Valencia 3, 28012, Madrid, Spain

[2] Universidad Politécnica de Madrid, Ronda de Valencia 3, 28012, Madrid, Spain

the hippocampal markers from MRI are not fully validated, and the level of standardization for their measurements is still too limited for clinical use (Schröder and Pantel 2016).

Many large scale studies are now collecting longitudinal MRI data (Lawrence et al. 2017). Over time, the hippocampal markers have been shown to correlate with the progression of AD (Wolz et al. 2010; Aubert-Broche et al. 2013; Frankó and Joly 2013; Iglesias et al. 2016). Many publications have reported that the hippocampal atrophy rates over time are accelerated in AD patients compared to the normal control (NC) subjects and included MCI subjects (Chételat et al. 2008; Apostolova et al. 2010; Mert et al. 2011). Compared to the cross-sectional approach, the longitudinal data can provide increased statistical power by reducing the confounding effect of between-subject variability (Thompson et al. 2011). However, most of the existing methods address the extraction of hippocampal markers using a single sample. One of the major caveats in longitudinal analysis is the use of tools that were originally designed for the analysis of data collected from the cross-sectional approach.

When subtle hippocampal structure changes caused by atrophy are being measured, a robust procedure for all samples from a subject is crucial. Simultaneous registration, segmentation and classification tasks of subject scan sequences have been shown to increase the accuracy of atrophy measurement (Reuter et al. 2012; Aubert-Broche et al. 2013; Bernal-Rusiel et al. 2013; Chincarini et al. 2016). Reuter et al. (2012) presented a longitudinal analysis framework in which a template for each subject was first created from all time points to estimate average subject anatomy using a symmetric robust registration method. Then, information from the subject template and from the individual runs was used to initialize the segmentation algorithm. Regarding longitudinal segmentation, Wolz et al. (2010) introduced a 4D image segmentation with a graph-cut algorithm. In their model, unary terms included intensity and anatomical priors, whereas pairwise terms were addressed to enforce spatial and temporal smoothness in the segmentation. The segmentation of all time points was then computed simultaneously with graph cuts. Wang et al. (2016) have proposed a longitudinal patch-based label fusion method to propagate atlases to the subject scan sequence by simultaneously labeling a set of temporally corresponding voxels with a temporal consistency constraint on sparse representation.

Inspired by these previous works, we propose a new method to measure hippocampal atrophy in longitudinal MRI scans for diagnosing AD in clinical groups. First, we use a longitudinal registration based on the creation of linear subject-specific templates. Second, we introduce a novel approach to the simultaneous hippocampal segmentation across multiple time points. This method is based on a 4D graph-cut algorithm, extending our cross-sectional approach to longitudinal data (Platero and Carmen Tobar 2015, 2016). Third, once the longitudinal markers of the segmented hippocampus are calculated, linear mixed effects (LME) modeling is used for the classification tasks. The longitudinal tendency of a hippocampal marker is estimated from the time points of a subject and compared with the mean trajectories of the clinical groups of AD, generating a diagnosis for the subject.

The goal of the present paper is to study how the proposed longitudinal framework improves the detection results from clinical groups during the progression of AD. The proposed longitudinal pipeline is compared with three other implementations: a cross-sectional pipeline, a hybrid pipeline that combines longitudinal registration with cross-sectional segmentation, and the FreeSurfer longitudinal framework. These four pipelines are applied to a scan/rescan database to study the reliability of the hippocampal markers. Then, the competing methods are used to analyze the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for detecting subjects among clinical groups during the progression of AD.

## Method

Given a set of longitudinal MRI scans, the data are firstly pre-processed with the FreeSurfer longitudinal registration framework (Fischl et al. 2002; Reuter et al. 2012). First, all scans of the subjects are processed independently (*Cross*, see Fig. 1). For each subject a template is created from all time points to estimate average subject anatomy (Reuter et al. 2010), which is built by registering scans from each time point of a subject using a robust and inverse consistent registration algorithm (*Base*). Then, each time point is processed longitudinally, where information from the subject-template and from the individual runs is used to initialize several of the algorithms (*Long*). Several steps in the processing of the serial MRI scans, such as Talairach transforms and brainmasks, were then initialized with common information from the subject-specific template. This implementation ensures that all time points are treated uniformly and with a very high degree of reproducibility (Reuter et al. 2012). Then, all longitudinal processed scans are independently registered into the MNI152 template (Evans et al. 2012) by means affine transformations using advanced normalization tools (ANTs) (Klein et al. 2009) and 12 degrees of freedom.

A set of atlases is built using the HarP annotations on 135 images belonging to the ADNI database (Boccardi et al. 2011; Frisoni et al. 2015). The atlases are co-registerd to the MNI-152 space with an affine transformation using ANTs and 12 degrees of freedom. A region of interest (ROI) is defined for each structure studied (left and right

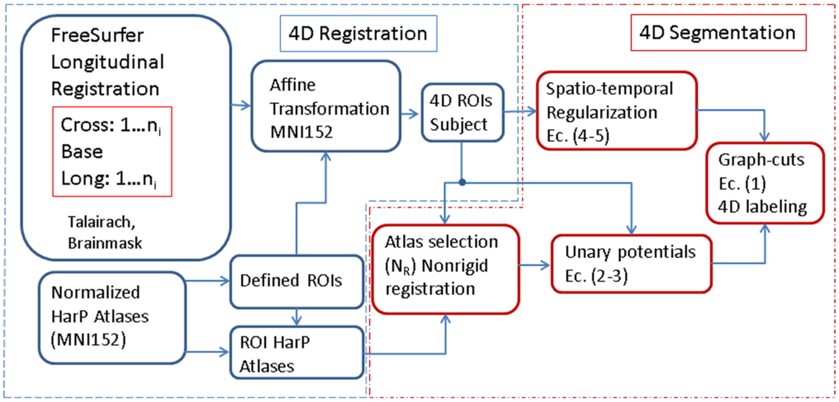
FreeSurfer Longitudinal Registration
Cross: 1...n$_i$
Base
Long: 1...n$_i$
Talairach, Brainmask
4D Registration
Affine Transformation MNI152
4D ROIs Subject
Normalized HarP Atlases (MNI152)
Defined ROIs
ROI HarP Atlases
4D Segmentation
Spatio-temporal Regularization Ec. (4-5)
Graph-cuts Ec. (1) 4D labeling
Atlas selection (N$_R$) Nonrigid registration
Unary potentials Ec. (2-3)

**Fig. 1** Flow chart summarizing the processing of the time-points from a subject

hippocampus) as the minimum bounding box containing the structure for all of the normalized HarP atlases and expanded by three voxels along each dimension. The definitions of the ROIs are used as starting points of the segmentation algorithm (see Fig. 1).

## Longitudinal Segmentations

We introduce a new label fusion method based on the spatio-temporal conditional random field (CRF) (Lafferty et al. 2001) formulation and on ROIs. The method is an extension of our proposed label fusion applied to cross-sectional analysis (Platero and Carmen Tobar 2015, 2016, 2017). The CRF modeling allows us to efficiently incorporate shape, appearance and context information into the segmentation process (Lotjonen et al. 2010; van der Lijn et al. 2008; Song et al. 2006; Wolz et al. 2010; Platero and Carmen Tobar 2015). The CRF model is characterized by a pseudo-Boolean function defined by unary and pairwise potentials. The proposed Boolean function is representable by graphs. A 4D graph is built to segment the longitudinal data: Voxels of a subject scan sequence are the vertices of the graph, and the edges connect spatial-temporal voxel neighborhoods. The edge weights are based on prior data from the HarP atlases and spatial-temporal constraints. The longitudinal hippocampal segmentation is simultaneously found by applying the min-cut/max-flow algorithm of Boykov and Kolmogorov (2004).

Given a 4D ROI of a subject scan sequence, which is the 3D ROI defined by the overlaps in the HarP atlases and extended with the time points $I : \Omega \subset \mathbb{N}^4 \rightarrow \mathbb{R}$, we seek to minimize an energy function under a discrete random field, $S : \Omega \subset \mathbb{N}^4 \rightarrow \{0, 1\}$, with a neighborhood system $\mathcal{E}$. The neighborhood system $\mathcal{E}$ is the set of edges that connect variables in the random field. A CRF is defined by the following pseudo-Boolean function:

$$E(S) = \sum_{p \in \Omega} \psi_p(S(p)) + \lambda \sum_{p,q \in \mathcal{E}} \psi_{pq}(S(p), S(q)), \qquad (1)$$

where $\psi_p(S(p))$ and $\psi_{pq}(S(p), S(q))$ are the unary and pairwise potentials, respectively, and $p$ and $q$ denote the voxel position in space and time. A $p$ voxel is defined by spatial coordinates, $x$, and also by a time coordinate, $t$. $\lambda$ is a tunable parameter that determines the trade-off between the unary and pairwise potentials.

### Atlases to Target

The unary potentials $\psi_p(S(p))$ use a Bayesian formulation, which allows for prior information about the shape and the appearance of the structures to be segmented from the atlases. The conventional unitary potentials measure only the pairwise similarity between the atlas and target voxels. In contrast, our unary potentials combine an appearance model based on multiple features extracted from each voxel and its neighborhood (Platero and Carmen Tobar 2015) with a label prior using a weighted voting method (Artaechevarria et al. 2009).

Consider a 4D ROI of a subject scan sequence, $I = \{I_t\}_{t=1,\ldots,n}$, where $I_t$ is a 3D ROI scan and $n$ is the number of scans of a subject. From the definition of the HarP atlases, it is only possible to construct 3D appearance-shape models. Therefore, for each $I_t$, the HarP atlas images are first ranked based on their similarity according to $I_t$ using the mutual information (MI) measure (Viola and Wells 1997), and the $N_R$ atlases most similar to the target image are registered non-rigidly into $I_t$, where $N_R$ is the total number of registered atlases (Platero and Carmen Tobar 2015) (see Fig. 1). The non-rigid registrations are computed using Elastix (Klein et al. 2010), a publicly available package for medical image registration. The registered atlas images are convolved with a filter bank. A set of feature extraction kernels are used to produce different feature maps. Discriminative features are necessary due to the hippocampus in T1w MRI having similar intensities as the background. These feature vectors are used to train a $k$-nearest neighbor ($k$-NN) appearance model (Platero and Carmen Tobar 2015, 2017). For computational efficiency, the $k$d-tree algorithm (Mount and Arya 2010) is used to perform the nearest neighbor search. $I_t$ is also convolved, and the image likelihoods of the voxels are calculated (Platero and Carmen Tobar 2015).

The label prior probability models the joint probability of all voxels that belong to the 4D ROI in a particular label configuration. The spatial prior is obtained from the $N_R$-selected registered atlases and from weighting the transferred label maps probabilistically using a local similarity measure between $I_t$ and the registered atlas image (Artaechevarria et al. 2009).

The main assumption in longitudinal segmentation is that given a robust longitudinal registration on the subject scan sequence, the segmentations at different time points can be forced to be consistent in inner hippocampal areas. Conversely, the hippocampal boundaries can be expected to have differences in the segmentations of the serial scans; these differences are caused by the biological atrophy of the subcortical structure. Tissue loss resulting from AD can be observed as a shift in the boundaries of the hippocampus. Therefore, a 3D edge detector is applied to each $I_t$. For each spatial point $x$ of the inner hippocampal area, the probability of belonging to the structure will be the same in all scans of a subject, which will be inferred from the average of the time points. In contrast, the probabilities of the edge voxels will be estimated independently in each scan. The image likelihood and label prior terms are combined, and the negative logarithm of the probability maps $p(S(p)|I(p))$ serves as the unary potentials:

$$\psi_p(S(p)) = \begin{cases} -\log\left(p(S(p(x,t))|I(p(x,t));\mathbb{A})\right) & \text{if } T(I(p(x,t)) > h, \\ -\log\left(\frac{1}{n}\sum_{t=1}^{n} p(S(p(x,t))|I(p(x,t));\mathbb{A})\right) & \text{otherwise.} \end{cases}, \tag{2}$$

with:

$$T(I(p(x,t)) = \max_{t=1,\ldots,n} g(\|\nabla I_t(x)\|), \tag{3}$$

where $\mathbb{A}$ is the set of registered HarP atlases to $I$ as a collection the $N_R$-selected atlases for each $I_t$, $g(\|\nabla I_t(x)\|) = 1 - \exp\left(-\frac{\|\nabla I_t(x)\|}{\sigma_G}\right)$ is a 3D edge detector with a normalization factor $\sigma_G$ and $h$ is a threshold value, which defines if a voxel is a hippocampal edge.

### Spatio-temporal Regularization

The pairwise potentials $\psi_{pq}(S(p), S(q))$, which describe the spatio-temporal regularization, consider the labeling coherence among the voxels located in a certain spatial and temporal neighborhood of the serial scans (see Fig. 1). According to the work of Song et al. (2006), a smoothness term is added to the energy function. These authors combined intensity and local boundary information into the pairwise potentials. These pairwise potentials take the form of a contrast-sensitive Potts model:

$$\psi_{pq}(S(p), S(q)) = \begin{cases} 0 & \text{if } S(p) = S(q), \\ \gamma(S(p), S(q)) & \text{otherwise.} \end{cases}, \tag{4}$$

where

$$\gamma(S(p), S(q)) = c\left(1 + \ln\left(1 + \frac{\|I(p) - I(q)\|^2}{2\sigma^2}\right)\right)^{-1} + (1-c)\left(1 - \max_{r \in M_{pq}} B(I(r))\right), \tag{5}$$

where $M_{pq}$ is a line that joins $p$ and $q$ and $\sigma$ is the robust scale of image $I$. If $p$ is $(x_1, t_1)$ and $q$ is $(x_2, t_2)$, then $B(I(x,t)) = g(\|\nabla I(x,t)\|)$ with $t_1 = t_2 = t$; i.e. if $p$ and $q$ are voxels in the same scan (i.e. spatial neighborhood) or $B(I(x,t)) = g(\|\partial_t I(x,t)\|)$ when $x_1 = x_2 = x$, then $p$ and $q$ are voxels belonging to different scans (i.e. temporal neighborhood). The parameter $0 \leq c \leq 1$ controls the influence of the boundary-based and intensity-based parts. $\gamma(S(p), S(q))$ is a weighting function that depends on the ”spacing” in each dimension. In spatial edges (within time points), the spatial resolution is taken into account, while temporal edges (between time points) are selected using a spacing based on time intervals. Furthermore, an additional parameter is used for weighting among spatial and temporal edges. Figure 1 shows a flow chart that summarizes the processing of the serial scans.

## Data Analysis

Once the markers have been calculated from the segmented hippocampus, we use LME modeling for longitudinal data analysis to account for the between-subject and within-subject sources of variation. We consider the following LME model:

$$Y_i = X_i\beta + Z_i b_i + e_i,$$

where $Y_i$ is the vector of the hippocampal marker for the time points of subject $i$, $X_i$ is the design matrix for the fixed effects (including variables such as clinical group, age and scan-time), $\beta$ are the fixed effects coefficients and are identical for all subjects. In addition to the fixed effects, a mixed effects model is used for subject-specific random effects, where $Z_i$ is the design matrix for the random effects, $b_i$ is a vector of the random effects, and $e_i$ is a vector of measurement errors. The components of $b_i$ reflect how the subset of regression parameters for the $i$-th subject deviates from those of the population.

In this study, the LME model is built with a intercept and slope as random effects to be included in the longitudinal trajectory:

$$y_{ij} = (\beta_r + b_{ri}) + (\beta_s + b_{si})\, t_{ij} + e_{ij}, \tag{6}$$

where $j = 1, ..., n_i$ indexes the time points with $n_i$ indicating the number of scans for subject $i$, $y_{ij}$ is the $j$ measure of the hippocampal marker from subject $i$, $t_{ij}$ is the scan time from baseline (in years), and $\beta_r$ and $\beta_s$ are the intercept and slope, respectively. The model allows for a linear mean trajectory for each subject and the clinical group mean trajectory. This longitudinal analysis is then used to estimate the subject's hippocampal atrophy rate and consequently improves the classification accuracy in the diagnostic of a subject during the progression of AD.

The atrophy rate for a subject is defined as the slope of the longitudinal marker divided by the intercept. Clinical group and age are found as significant variables using a linear model (Fraser et al. 2015). For the analysis of the atrophy rate, a LME model is built with clinical groups and ages as fixed effects. Considering a clinical group, the atrophy rate for a subject $i$ of this group is calculated as:

$$\Lambda_i = \frac{\beta_s + b_{si}}{\beta_r + b_{ri}}. \tag{7}$$

Regarding the classification of subjects between clinical groups of AD, we propose to compare the longitudinal tendency of the marker using a LME model trained between pairs of clinical groups.

The difference between the longitudinal marker trajectory of an $i$-subject and the LME model will be described by the random vectors $b_i$ and $e_i$, which follow mean zero-Gaussian multivariate distributions, indicating a population-averaged mean of $E(Y_i) = X_i\beta$ (Bernal-Rusiel et al. 2013). Therefore, the longitudinal trajectory residue is defined as:

$$l_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij} - X_{ij}^k\beta\right), \tag{8}$$

where $X_{ij}^k$ is the $k$-column vector of the design matrix and $k$ is a binary variable, which indexes the pair of the clinical groups to be classified. The $l_i$-samples belonging to the group indexed by $k$ will follow a normal distribution of zero mean and of variance determined by $b_i$ and $e_i$. In contrast, $l_i$-samples not belonging to the $k$-group will also follow normal a distribution with the same variance but with a biased mean value of zero, which is related to the fixed effects of the clinical groups. The random variable $l_i$ is used to train and classify the hippocampal markers by discriminant analysis.

## Materials

Two publicly available datasets were used in the experiments: Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) and ADNI.

Scan-rescan reliability was evaluated using the MIRIAD database. The MIRIAD dataset consists of T1w brain MRI acquired at regular interval times. All scans were acquired on the same 1.5 T scanner. A subset of 23 healthy elderly people with three time points acquired 14 days apart was used. No changes between scans were expected. The mean age at baseline of the subjects was 69.7 ± 7.2 years with mini-mental state examination (MMSE) scores of 29.4±0.8 and 52% men (Malone et al. 2013).

The ADNI dataset was selected to analyze the behavior of the longitudinal framework in the AD study, where subjects have different numbers of visits (Bradley et al. 2013). In the ADNI study, brain T1w MRI were acquired at baseline and regular intervals. MRI acquisition was performed according to the ADNI acquisition protocol (Clifford et al. 2008). In our study, all the images downloaded from ADNI are pre-processed using the N3 method (Sled et al. 1998) or grad warped, followed by B1 bias field correction and N3 intensity non-uniformity correction (Clifford et al. 2008).

According to the ADNI evaluators, subjects were grouped into three clinical groups consisting of 114 normal controls (NCs), 183 mild cognitive impairment (MCI) and 77 Alzheimer's disease (AD). MCI subjects were further divided into 101 MCI progressing to AD (pMCI) and 82 stable MCI (sMCI) according to the clinical follow-up. An overview of the subject groups is given in Table 1. For each group, the total number of subjects, the number of males, and the MMSE values (Folstein et al. 1975) are shown. No significant differences in age and gender were observed among the clinical groups. There were significant differences in MMSE scores among the clinical groups for all time points. Table 2 shows the time points (Baseline, Month 6, Month 12, Month 24) that were available for the selected subjects.

## Results

To study the impact of the longitudinal frameworks, three methods were implemented and compared: (a) A standard cross-sectional framework (CC method: cross-sectional registration and cross-sectional segmentation). Each scan was treated independently (Platero and Carmen Tobar 2015, 2016). (b) A hybrid framework (LC method: longitudinal registration and cross-sectional segmentation). This implementation merged these two processes by generating an unbiased within-subject template space while segmentations follow on each 3D ROI independently. (c)

**Table 1** Demographic and clinical details of the subset of ADNI database used in this study

| Type (N. subjects) | NC (114) | sMCI (82) | pMCI (101) | AD (77) | F | p |
|---|---|---|---|---|---|---|
| Gender male (%) | 61 (54%) | 57 (69%) | 58 (57%) | 41 (53%) | | 0.105 |
| Baseline age | 75.5 (5.2) | 74.3 (6.9) | 75.2 (7.0) | 73.9 (6.9) | 1.25 | 0.292 |
| MMSE (Baseline) | 29.1 (1.0) | 27.6 (1.6)[a] | 26.8 (1.9)[a,b] | 23.4 (1.8)[a,b,c] | 210.02 | $< 0.001$ |
| MMSE (Year 0.5) | 29.1 (1.0) | 27.4 (2.0)[a] | 26.0 (3.2)[a,b] | 22.3 (3.2)[a,b,c] | 121.30 | $< 0.001$ |
| MMSE (Year 1) | 29.1 (1.2) | 27.3 (2.3)[a] | 25.7 (3.7)[a,b] | 20.6 (4.5)[a,b,c] | 119.34 | $< 0.001$ |
| MMSE (Year 2) | 29.1 (1.0) | 26.8 (4.8)[a] | 23.4 (5.0)[a,b] | 20.8 (6.7)[a,b,c] | 83.85 | $< 0.001$ |

Data as represented as mean and standard deviation (SD) unless specific otherwise. ANOVA with Bonferroni post hoc test is used for baseline age and neuropsychological score, except for gender where the chi-square test is used. Statistical significance is considered with $p - value < 0.01$. [a] Significant compared to normal control (NC). [b] Significant compared to sMCI. [c] Significant compared to pMCI. NC= Normal control; sMCI = Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease; MMSE= Mini-Mental State Examination

The proposal (LL method, which combines the longitudinal registration and proposed segmentation steps). Additionally, the results of the FreeSurfer longitudinal framework were considered in the performance comparison of the implemented algorithms.

The performance of the above methods was evaluated using four experiments. The first one was quality control to assess the robustness of the compared methods on a large number of images. Second, the test/retest reproducibility of the methods was evaluated using the MIRIAD dataset. Third, the atrophy rates of the clinical groups in the AD study were statistically analyzed. Finally, the longitudinal data were used for training the classifiers and verifying that the accuracy was improved when longitudinal trends of the hippocampal markers were integrated.

## Quality

A total of 1386 scans in ADNI and 69 in MIRIAD were processed by each of the methods used. A quality control test checked whether the applied longitudinal pipeline had failed on the studied data. For each hippocampal region and each time point, the test computed the best Pearson correlation coefficient between the 3D ROI intensities and each HarP atlas image, as well as among the hippocampal segmentation and each HarP atlas labeling (Chincarini et al. 2016). These coefficients were exclusively calculated within the volume defined by the specific ROI, not on the entire scan. The assumption of the test is that the HarP atlases represent all relevant anatomical variability from each clinical group. Then, the lowest pair of correlation coefficients among those calculated values from each subject scan sequence was saved. If the intensity and labeling correlation coefficients of a subject are in the low value regions of the distribution of the population, then the scans of this subject were analyzed manually. To validate the consistency of the applied longitudinal pipeline, another test was also used between the scans and hippocampal segmentations of a subject. In this case, the correlation coefficients were computed between pairs of the ROIs and hippocampal segmentations of the same subject, and the lowest values among these pairs were selected. This quality control was applied to all four implementations. As an example, Fig. 2 shows the distributions of the selected correlation coefficients for all 346 subjects of ADNI from the CC method. Points with low values are accompanied by the digital identifier of the subject. The two tests show results that are consistent with each other. Both scatters indicate a common group of potential outliers. The source of these errors is the spatial normalization into the MNI152 template by affine transformation using FLIRT (Jenkinson et al. 2002), which failed to orient the brain correctly. This bug was fixed using ANTs. However, the test based on correlation analysis between two scans of the same subject shows another subject in a different low value region. In this case, the source of the error is an improper alignment of one of the scans in the registration. This error was also corrected using ANTs.

**Table 2** Number and timing of scans per time point by clinical group. NC= Normal control; sMCI= Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease

| Type (N. subjects) | NC | sMCI | pMCI | AD | Time from baseline |
|---|---|---|---|---|---|
| Baseline | 114 | 82 | 101 | 77 | 0 |
| Year 0.5 | 112 | 82 | 101 | 77 | 0.57(0.05) |
| Year 1 | 110 | 82 | 101 | 77 | 1.08(0.07) |
| Year 2 | 92 | 60 | 65 | 51 | 2.08(0.10) |
| Total | 428 | 306 | 368 | 282 | |

The scatter plots also display that there is no evidence of bias in longitudinal processing due to the nature of subjects belonging to different clinical groups.

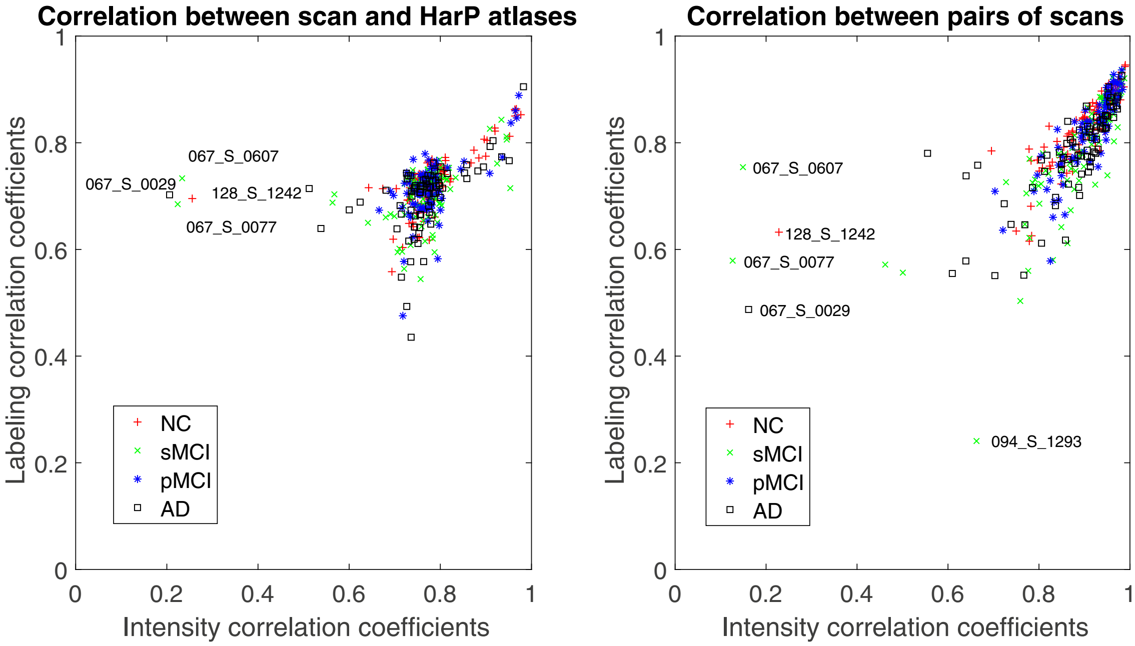

**Fig. 2** Scatter plots between intensity and labeling correlation coefficients from the CC method: **a** Measured values between scan and HarP atlases (left), **b** Measured values between two scans of the same subject (right). Each point represents a subject. Outliers indicate errors in longitudinal processing. NC= Normal control; sMCI= Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease

## Test-retest

We evaluated the statistics of the segmented hippocampus between repeat scans from the same subject, which were acquired in a short period of time and in absence of biological variability. Test/retest reproducibility is a critical measure for reliable markers. To evaluate the test/retest reproducibility of the methods, the scan/rescan data of the MIRIAD dataset were used. For each method and each subject, we compared the hippocampal segmentations from the subject scan sequence. Two measures were calculated for each segmented left and right hippocampus: a) The absolute volume difference (AVD) and b) the dice coefficient between two binary images corresponding to the hippocampal regions. Given a method and a pair of scans from the same subject, the considered values were defined as:

$$\mathrm{AVD}(\%) = 200\frac{|V_k - V_l|}{V_k + V_l}, \quad \mathrm{Dice} = 2\frac{|S_k| \cap |S_l|}{|S_k| + |S_l|}$$

where $V_k$ and $V_l$ are the hippocampal volumes at time points $k$ and $l$ of a subject and $S_k$ and $S_l$ are their binary segmentations, respectively. Figure 3 displays the distributions of these measures depending on the method and hippocampal region of the MIRIAD dataset. Table 3 provides the mean and standard deviation of the AVD and dice coefficient of the hippocampal segmentations. ANOVA with the Bonferroni post hoc test is used for the AVD and dice coefficient in each hippocampal region. The proposed longitudinal method produces the lowest AVDs in both hippocampal regions. The LL method provides the highest dice coefficient and yields a statistically significant improvement with respect to the other three methods.

## Analysis Data and Classification

The test-retest experiment described above evaluated only the reliability of the hippocampal segmentations from the longitudinal algorithms. Next, we proceeded to validate the performance of the methods by measuring the hippocampal atrophy over time from the AD study. Two experiments were carried out to analyze the dependency of the longitudinal trends on clinical groups: a) Hippocampal atrophy rates and b) discriminating subjects between clinical groups of AD.

In this study, two markers were used: hippocampal volume (HV) and hippocampal surface roughness (SR). Recent studies have shown that a decrease in HV and an increase in its SR are good markers for studying AD (Kim et al. 2015). The SR was validated as an efficient marker since it presented similar capabilities of discrimination as the traditional measure of HV but had the advantage of providing a local analysis, which may produce atrophy maps of the progression of AD (Platero and Carmen Tobar 2016). Moreover, these markers are scalar features, showing robustness in generalization and avoiding overfitting when the size of the samples is limited.

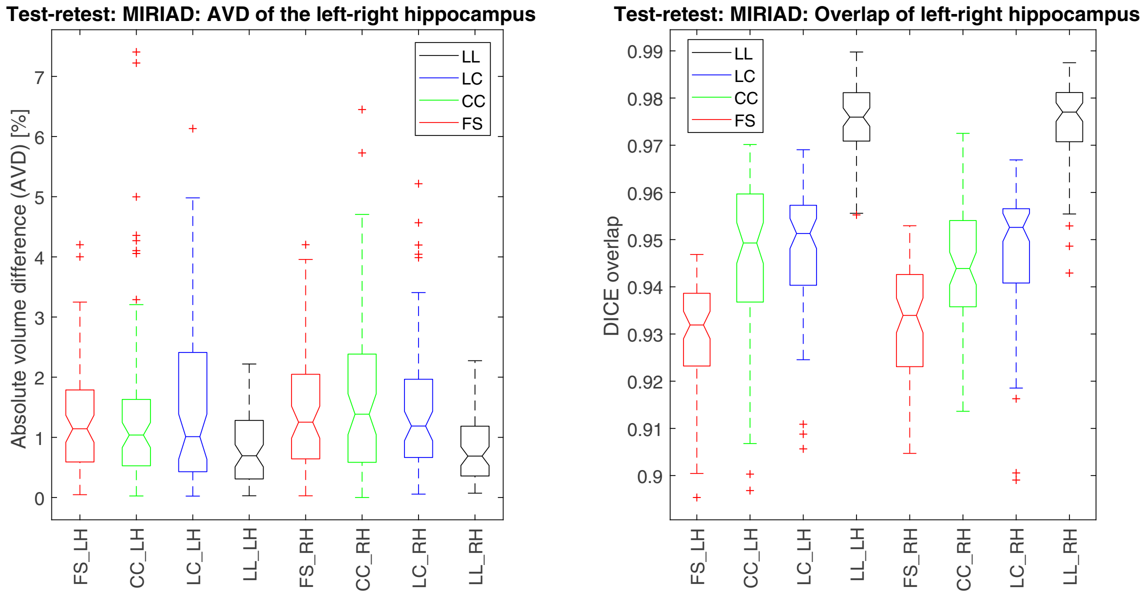

**Fig. 3** The absolute volume differences (in %) and dice distributions for the left and right hippocampus in the test-retest of the MIRIAD dataset using FreeSurfer v5.3 (FS) and the three implemented methods (CC, LC and LL). LH= Left hippocampus; RH= Right hippocampus; CC= Cross-sectional; LC= Hybrid framework; LL=proposed framework

For each scan, we computed the volume of the hippocampus. For more robustness with respect to segmentation errors, the left and right volumes were added. The SR of a hippocampus was given by

$$SR = \sqrt{\frac{1}{m}\sum_{i}^{m} K^2(x_i)},$$

where $m$ is the number of the voxels belonging to the hippocampal surface and $K(x_i)$ is the mean curvature at each voxel $x_i$. These voxels were extracted from the automated hippocampal segmentation in the normalized spatial, i.e., $K(x_i)$ is calculated with the isotropic spacing ($1 \times 1 \times 1mm^3$ from the MNI 152 space). The left and right hippocampal segmentations were embedded in a level set formulation, and the mean curvature was estimated using:

$$K(x_i) = -div\left(\frac{\nabla\Gamma(x_i)}{\|\nabla\Gamma(x_i)\|}\right),$$

where $\Gamma = \{x|\varphi(x) = 0\}$ is the hippocampal surface and $\varphi(x)$ is a signed distance function, which assigns positive distances to the inside of the object and negative distances to the outside (Osher et al. 2004). The estimation of the mean curvature on each $x_i$ is controlled using the Gaussian derivatives. In our experiments, a Gaussian kernel of 5 mm FWHM was applied to the images for calculating the mean curvatures. The left and right SRs were also added to determine the shape marker.

Given that all scans were co-registered to the common template space, instead of calculating the hippocampal makers from the scans in their native space, the measurements were extracted in the MNI152 domain. This method of calculating the markers has two advantages: (i) The hippocampal volume does not need to be normalized by the intracranial volume since all the brains are normalized, and (ii) the spacing of MNI152 is isotropic, which facilitates the calculation of the SR.

**Table 3** The absolute volume difference (AVD) and dice overlap for the hippocampal regions in the test-retest scans of the MIRIAD dataset

| Measure (Method) | FS | CC | LC | LL | F | p |
|---|---|---|---|---|---|---|
| AVD (%) LH | 1.33 (0.95) | 1.52 (1.53) | 1.49 (1.34) | 0.79 (0.58) | 0.51 | 0.673 |
| AVD (%) RH | 1.50 (1.11) | 1.62 (1.36) | 1.48 (1.15) | 0.81 (0.57) | 1.57 | 0.196 |
| Dice LH | 0.930 (0.012) | 0.946 (0.017)$^a$ | 0.948 (0.014)$^a$ | 0.975 (0.008)$^{a,b,c}$ | 59.47 | $< 0.001$ |
| Dice RH | 0.932 (0.012) | 0.943 (0.014)$^a$ | 0.948 (0.014)$^a$ | 0.975 (0.009)$^{a,b,c}$ | 48.19 | $< 0.001$ |

Data as represented as mean and standard deviation. ANOVA with Bonferroni post hoc test is used for AVD and Dice coefficient in each hippocampal region. Statistical significance is considered with $p - value < 0.01$. [a] Significant compared to FS . [b] Significant compared to CC. [c] Significant compared to LC. LH= Left hippocampus; RH= Right hippocampus; FS=FreeSurfer; CC=Cross-sectional; LC=Hybrid framework; LL= proposed framework

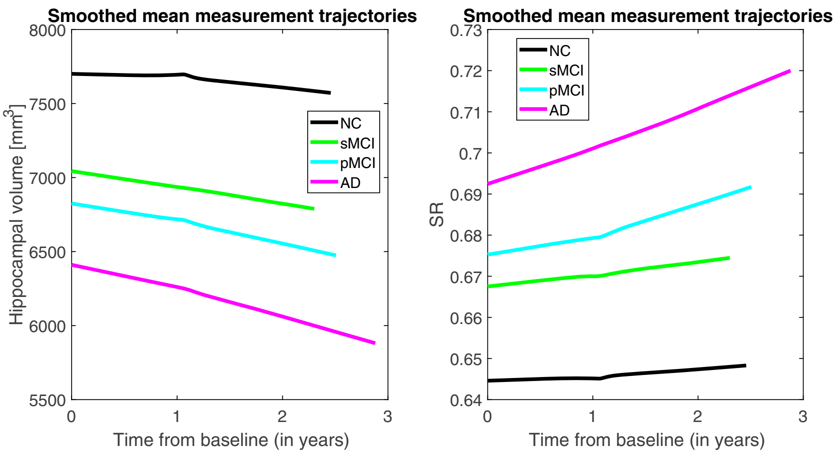

**Fig. 4** The smoothed mean trajectories for each of the four clinical groups in the markers from the proposed longitudinal framework: **a** Hippocampal volume (HV) and **b** Surface roughness (SR). NC= Normal control; sMCI= Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease

For each subject and each time point of the ADNI dataset, hippocampal markers were obtained using the competing methods. Figure 4 show the lowess plots of the two hippocampal markers using the LL method. These plots reveal that a linear model is sufficient for capturing follow-up trends in the four clinical groups.

From these markers, LME models were built using the FreeSurfer LME toolbox (Bernal-Rusiel et al. 2013). Variables considered significant enough to include in the fixed effects were age, clinical group and scan time. According to the LME model suggested in Eq. (6), two variables were used for the random effect model. Therefore, the LME model with fixed and random effects is defined as:

$$y_{ij} = (\beta_0 + \beta_1 \cdot sMCI + \beta_2 \cdot pMCI + \beta_3 \cdot AD + \beta_4 \cdot age + b_{ri}) + (\beta_5 + \beta_6 \cdot sMCI + \beta_7 \cdot pMCI + \beta_8 \cdot AD + b_{si})\, t_{ij} + e_{ij}, \tag{9}$$

where age is the years at baseline and the set of binary variables $\{sMCI, pMCI, AD\}$ represents the flags

**Table 4** Method comparison in terms of HV and SR markers at baseline and their atrophy rates for the four longitudinal methods $(HV_0, \Lambda_{HV}, SR_0, \Lambda_{SR})$

| Feature | Method | NC | sMCI | pMCI | AD | F | p |
|---|---|---|---|---|---|---|---|
| | FS | 8992(1212) | 7816(1464)$^a$ | 7244(1342)$^a$ | 6769(1348)$^{a,b}$ | 51.48 | $< 0.001$ |
| $HV_0$ | CC | 7718(593) | 7052(843)$^a$ | 6798(819)$^a$ | 6487(859)$^{a,b}$ | 42.94 | $< 0.001$ |
| $[mm^3]$ | LC | 7804(622) | 7116(856)$^a$ | 6847(842)$^a$ | 6524(863)$^{a,b}$ | 47.00 | $< 0.001$ |
| | LL | 7636(615) | 6943(855)$^a$ | 6671(835)$^{a,b}$ | 6358(848)$^{a,b}$ | 48.01 | $< 0.001$ |
| | FS | −1.01(0.90) | −2.44(1.46)$^a$ | −3.44(2.68)$^{a,b}$ | −4.64(1.71)$^{a,b,c}$ | 68.91 | $< 0.001$ |
| $\Lambda_{HV}$ | CC | −0.63(1.11) | −1.20(1.01)$^a$ | −1.95(1.30)$^{a,b}$ | −2.47(0.97)$^{a,b}$ | 50.48 | $< 0.001$ |
| [%] | LC | −0.65(0.67) | −1.48(0.88)$^a$ | −2.08(1.33)$^{a,b}$ | −2.86(1.05)$^{a,b,c}$ | 81.80 | $< 0.001$ |
| | LL | −0.86(0.71) | −1.62(1.04)$^a$ | −2.20(1.32)$^{a,b}$ | −3.32(1.21)$^{a,b,c}$ | 84.38 | $< 0.001$ |
| | FS | 0.624(0.031) | 0.659(0.047)$^a$ | 0.677(0.047)$^a$ | 0.698(0.047)$^{a,b}$ | 52.66 | $< 0.001$ |
| $SR_0$ | CC | 0.643(0.019) | 0.665(0.031)$^a$ | 0.675(0.032)$^a$ | 0.689(0.031)$^{a,b,c}$ | 47.18 | $< 0.001$ |
| | LC | 0.639(0.020) | 0.639(0.020)$^a$ | 0.673(0.033)$^a$ | 0.687(0.031)$^{a,b,c}$ | 48.06 | $< 0.001$ |
| | LL | 0.644(0.020) | 0.667(0.032)$^a$ | 0.679(0.033)$^{a,b}$ | 0.693(0.032)$^{a,b,c}$ | 49.16 | $< 0.001$ |
| | FS | 0.37(0.47) | 0.90(0.62)$^a$ | 1.31(1.32)$^{a,b}$ | 1.62(0.74)$^{a,b}$ | 39.09 | $< 0.001$ |
| $\Lambda_{SR}$ | CC | 0.22(0.54) | 0.41(0.43) | 0.83(0.56)$^{a,b}$ | 1.10(0.40)$^{a,b,c}$ | 59.03 | $< 0.001$ |
| [%] | LC | 0.19(0.32) | 0.54(0.40)$^a$ | 0.83(0.57)$^{a,b}$ | 1.11(0.45)$^{a,b,c}$ | 76.37 | $< 0.001$ |
| | LL | 0.25(0.32) | 0.60(0.43)$^a$ | 0.84(0.56)$^{a,b}$ | 1.28(0.50)$^{a,b,c}$ | 82.56 | $< 0.001$ |

Data as represented as mean and standard deviation (SD). MANCOVA with Bonferroni pairwise comparisons is used for analysing the markers at baseline and atrophy rates. Statistical significance is considered with $p - value < 0.01$. [a] Significant compared to NC. [b] Significant compared to sMCI. [c] Significant compared to pMCI. NC= Normal control; sMCI= Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease; HV= Hippocampal volume; SR= Surface roughness; $\Lambda$ = Annual atrophy rate in (%); FS= FreeSurfer; CC= Cross-sectional; LC= Hybrid framework; LL = proposed framework belonging to the clinical groups; i.e. the value of a flag is 1 if the subject is in the $k$-group and 0 otherwise.

The statistical analysis of the markers at baseline and their annual atrophy rates over the clinical groups and implementations are shown in Table 4. The annual atrophy rates were computed using Eq. (7). Figure 5 shows the distributions of these measurements. As the lowess plots for the two hippocampal markers have already shown, the HV atrophies indicate a loss of volume in the time, which is accelerated with the progression of the disease. The SR atrophies show an increase in the roughness over time, which is also accelerated with AD progression.

MANCOVA with Bonferroni pairwise comparisons was used for analyzing the markers with age as covariate. Both markers at baseline (i.e. $HV_0$, $SR_0$) show a similar discrimination capacity between the four implementations,

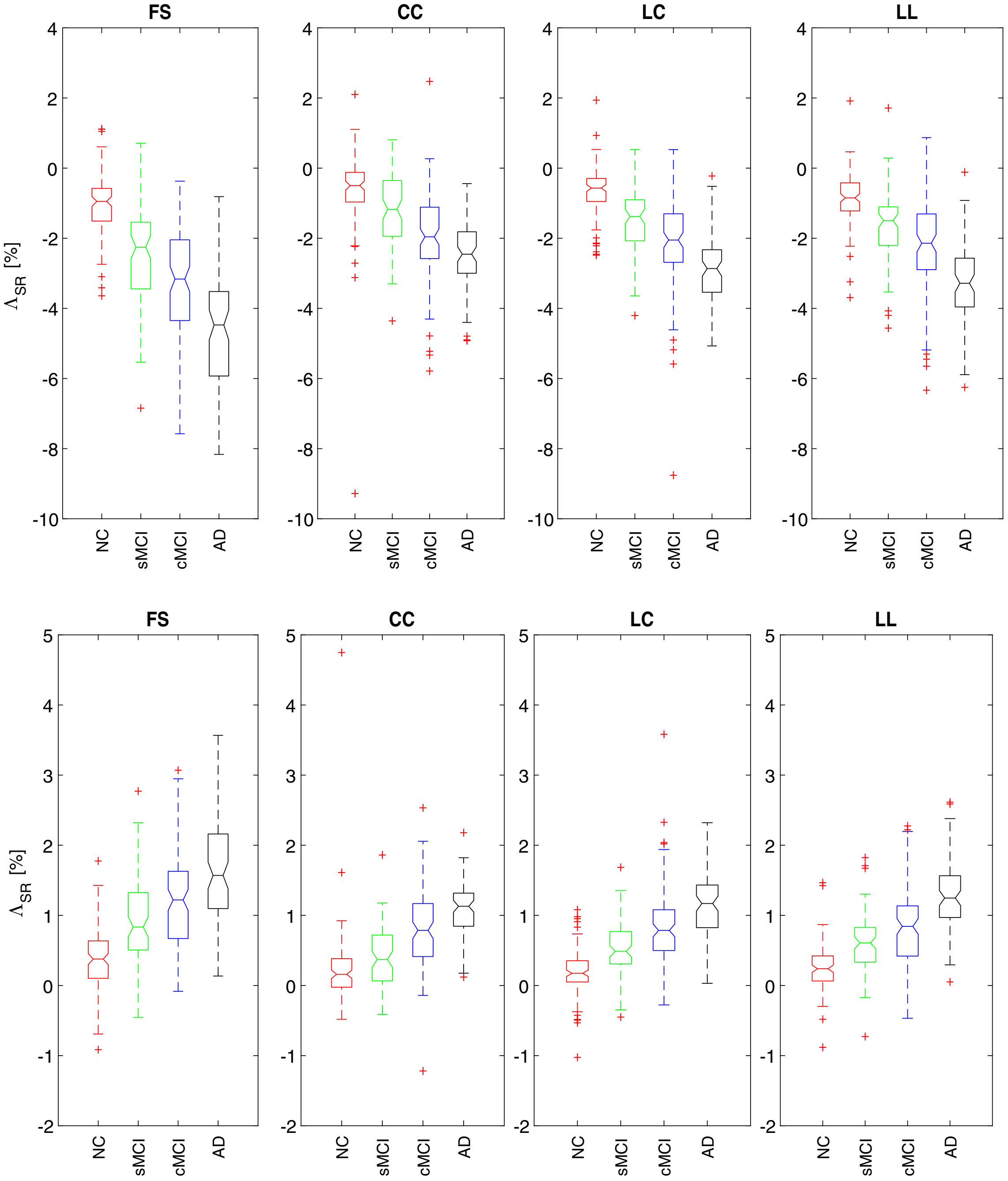

**Fig. 5** Annual atrophy rate distributions in (%) for each of the four clinical groups in the markers using the competing methods: **a** Hippocampal volume (HV) and **b** Hippocampal surface roughness (SR). NC= Normal control; sMCI= Stable Mild cognitive impairment; pMCI= Progressive Mild cognitive impairment; AD= Alzheimer disease; FS= FreeSurfer; CC= Cross-sectional; LC= Hybrid framework; LL= proposed framework; $\Lambda$ = Annual atrophy rate in (%)

with a slight improvement in the LL method compared to the CC approach. Regarding the annual atrophy rates of the markers, there is an improvement in the power of discrimination in the LL method ($\Lambda_{HV}$: F $=$ 84.38, $p <$ 0.001; $\Lambda_{SR}$: F $=$ 82.56, $p < 0.001$) with respect to the CC approach and even with regard to FS method, especially in relation to the clinical groups of MCI and AD. We also note that the marker distributions from each clinical group using our approach show less dispersion than the Freesurfer framework.

A comparison of the trends in the annual atrophy ratios in other publications can only be given on hippocampal volume. We have no references about SR atrophies. The $\Lambda_{HV}$ distributions by our LL method agree well with previously published data. Annual rates of volume atrophy vary across studies but they can be estimated at about $3\%-5\%$ for AD and around 1% for healthy aging (Fox et al. 2011). Wolz et al. (2010) reported that healthy aging was characterized by an annual average atrophy rate of 0.85%, sMCI showed an average annual atrophy rate of 1.72%, and pMCI were characterized by an annual average atrophy rate of 3.23%, which was similar to the 3.85% atrophy rate observed in AD. Iglesias et al. (2016) have published on the ADNI that for control subjects, the annual average atrophy rate is 1.5%. For sMCI, they reported rate of 2%, and for late MCI, a rate of 3.5% was reported. In the AD group, the rate ranged from 3.6% to 4%. The $\Lambda_{HV}$ values reported by the LL approach are in line with the values published by Wolz et al. The FS approach shows similar trends in the atrophy values to those published by Iglesias et al.

These accordances confirm: (1) The proposed LL approach follows the work of Wolz et al. Our proposal also includes a hippocampal segmentation based on graph cuts and a common CRF model; (2) Iglesias et al. used the FreeSurfer longitudinal framework to calculate hippocampal atrophy; thus, it was expected that the FS results would be similar to those reported by these authors; and (3) The population sample chosen in this work is representative, as for identical or similar implementations, the atrophy results among the clinical groups of the AD study are similar.

## Classification

The hippocampal markers were subsequently analyzed using linear discrimination, which fits a normal density to each group with a pooled estimate of covariance. The use of a simple, linear classifier ensured that the classification accuracy was primarily determined by the quality of the input data rather than by stochastic variations in the classifier.

The classifiers were trained using four features extracted from the markers: (a) the markers at baseline ($HV_0$ and $SR_0$), (b) the annual atrophy rates ($\Lambda_{HV}$ and $\Lambda_{SR}$), (c) the combination of the two previous features and (d) our proposal of the average residue between the $i$-longitudinal trajectory of the marker and the LME model (see Eq. (8), $l_i^{HV}$ and $l_i^{SR}$).

Given a feature of a marker, the classifications of paired data were experienced between the three main clinical groups: NC vs MCI, NC vs AD and MCI vs AD. Additionally, the discriminating power of the markers between sMCI vs pMCI patients was also evaluated, simultaneously showing the separation capacity between the pMCI subjects with respect to the AD group. The purpose of the comparison between pMCI vs AD is to show a certain amount of within-group homogeneity, especially when the pMCI is a high-risk cohort; that is, these subjects will undergo the onset of AD in the near future (Sabuncu et al. 2014).

For each comparison between two clinical groups, implementations and hippocampal features, receiver operating characteristic (ROC) curves were calculated. The discriminant value of the corresponding ROC curve was estimated using the area under curve (AUC). DeLong's test was applied to compare AUCs between methods (DeLong et al. 1988). Additionally, sensitivity (SEN), specificity (SPE) and accuracy (ACC) scores of the classifiers were also computed (Cuingnet et al. 2011).

A bootstrapping approach was used to evaluate the classification rate between pairs of clinical groups: for each group, 75% of the subjects were randomly selected for training. The remaining 25% were then classified according to the features used in the training sets. The classification scores and interval confidences for different groups displayed in the tables are after 5000 runs.

The first experiment evaluated the ability to discriminate between pairs of clinical groups by means of the four features of the HV and SR markers using the AUC performance of each implementation. Table 5 summarizes the AUC for each pair of clinical groups to be compared according to the features of the markers and the method used. Given an implementation and two groups to discriminate, in general it can be said that there are no differences between the classification results using the same feature of HV or SR marker. The combination of the marker at baseline with the annual atrophy ratio gives better results than classifiers trained with these features independently. This conclusion is in line with other publications (Chincarini et al. 2016). The best classification results were obtained when training the classifiers with $l_i^{HV}$ and $l_i^{SR}$.

Significant improvements in classification were observed with $l_i^{HV}$ and $l_i^{SR}$ from the longitudinal analysis compared to their features at baseline, i.e. $HV_0$ and $SR_0$, respectively, in any of the clinical groups to be discriminated and in any of the used implementations. Regarding the

**Table 5** For each comparison between two clinical groups, the performance AUC from each feature and each implementation (FS,CC,LC,LL)

| Feature | Method | NC vs MCI | p-value | NC vs AD | p-value | MCI vs AD | p-value |
|---|---|---|---|---|---|---|---|
| $HV_0$ | FS | 0.785(0.784, 0.786) | < 0.001 | 0.898(0.897, 0.899) | < 0.001 | 0.654(0.652, 0.656) | 0.006 |
| | CC | 0.783(0.782, 0.784) | 0.013 | 0.896(0.894, 0.897) | < 0.001 | 0.657(0.655, 0.658) | 0.006 |
| | LC | 0.789(0.787, 0.790) | 0.023 | 0.898(0.897, 0.899) | < 0.001 | 0.657(0.655, 0.659) | 0.005 |
| | LL | 0.794(0.793, 0.796) | 0.064 | 0.910(0.909, 0.911) | < 0.001 | 0.659(0.657, 0.661) | 0.005 |
| $\Lambda_{HV}$ | FS | 0.712(0.711, 0.714) | < 0.001 | 0.867(0.866, 0.869) | < 0.001 | 0.684(0.683, 0.686) | 0.274 |
| | CC | 0.656(0.655, 0.658) | < 0.001 | 0.795(0.794, 0.797) | < 0.001 | 0.607(0.605, 0.608) | < 0.001 |
| | LC | 0.707(0.706, 0.709) | < 0.001 | 0.830(0.829, 0.832) | < 0.001 | 0.663(0.661, 0.665) | 0.059 |
| | LL | 0.687(0.686, 0.689) | < 0.001 | 0.857(0.856, 0.859) | < 0.001 | 0.703(0.701, 0.705) | 0.362 |
| $HV_0, \Lambda_{HV}$ | FS | 0.811(0.810, 0.812) | 0.423 | 0.932(0.931, 0.933) | 0.251 | 0.701(0.700, 0.703) | 0.484 |
| | CC | 0.787(0.786, 0.789) | 0.021 | 0.912(0.911, 0.914) | 0.008 | 0.663(0.661, 0.665) | < 0.001 |
| | LC | 0.802(0.800, 0.803) | 0.353 | 0.934(0.933, 0.935) | 0.274 | 0.680(0.678, 0.682) | 0.057 |
| | LL | 0.795(0.794, 0.796) | 0.049 | 0.935(0.934, 0.936) | 0.303 | 0.686(0.684, 0.688) | 0.267 |
| $l_i^{HV}$ | FS | 0.820(0.819, 0.821) | 1 | 0.941(0.941, 0.942) | 0.700 | 0.691(0.689, 0.692) | 0.201 |
| | CC | 0.803(0.801, 0.804) | 0.182 | 0.928(0.927, 0.928) | 0.008 | 0.679(0.677, 0.681) | 0.057 |
| | LC | 0.806(0.805, 0.807) | 0.215 | 0.936(0.935, 0.936) | 0.073 | 0.694(0.692, 0.695) | 0.318 |
| | LL | 0.805(0.804, 0.806) | 0.223 | 0.947(0.946, 0.947) | 1 | 0.720(0.718, 0.722) | 1 |
| $SR_0$ | FS | 0.785(0.784, 0.787) | < 0.001 | 0.905(0.904, 0.906) | < 0.001 | 0.672(0.670, 0.674) | 0.038 |
| | CC | 0.776(0.775, 0.778) | < 0.001 | 0.896(0.895, 0.898) | < 0.001 | 0.670(0.668, 0.672) | 0.033 |
| | LC | 0.777(0.776, 0.778) | < 0.001 | 0.901(0.900, 0.902) | < 0.001 | 0.670(0.668, 0.672) | 0.036 |
| | LL | 0.780(0.779, 0.781) | < 0.001 | 0.908(0.907, 0.909) | < 0.001 | 0.675(0.673, 0.677) | 0.068 |
| $\Lambda_{SR}$ | FS | 0.698(0.697, 0.700) | < 0.001 | 0.822(0.820, 0.824) | < 0.001 | 0.655(0.653, 0.657) | 0.061 |
| | CC | 0.668(0.666, 0.669) | < 0.001 | 0.831(0.829, 0.832) | < 0.001 | 0.648(0.646, 0.650) | 0.046 |
| | LC | 0.696(0.694, 0.697) | < 0.001 | 0.813(0.811, 0.815) | < 0.001 | 0.651(0.649, 0.653) | 0.040 |
| | LL | 0.687(0.686, 0.689) | < 0.001 | 0.841(0.839, 0.842) | < 0.001 | 0.672(0.670, 0.674) | 0.049 |
| $SR_0, \Lambda_{SR}$ | FS | 0.804(0.802, 0.805) | 0.221 | 0.923(0.922, 0.924) | 0.123 | 0.696(0.694, 0.698) | 0.283 |
| | CC | 0.781(0.780, 0.782) | 0.006 | 0.916(0.914, 0.917) | 0.085 | 0.681(0.679, 0.683) | 0.117 |
| | LC | 0.804(0.803, 0.805) | 0.277 | 0.927(0.926, 0.928) | 0.301 | 0.698(0.696, 0.700) | 0.302 |
| | LL | 0.785(0.784, 0.787) | 0.021 | 0.928(0.927, 0.929) | 0.344 | 0.699(0.697, 0.700) | 0.319 |
| $l_i^{SR}$ | FS | 0.818(0.817, 0.819) | 1 | 0.945(0.944, 0.946) | 0.909 | 0.700(0.699, 0.702) | 0.314 |
| | CC | 0.796(0.794, 0.797) | 0.083 | 0.937(0.936, 0.938) | 0.089 | 0.697(0.695, 0.698) | 0.296 |
| | LC | 0.799(0.797, 0.800) | 0.132 | 0.942(0.941, 0.943) | 0.220 | 0.701(0.698, 0.703) | 0.314 |
| | LL | 0.799(0.798, 0.800) | 0.189 | 0.946(0.944, 0.947) | 1 | 0.712(0.710, 0.714) | 1 |

The p-values of paired DeLong's test are shown comparing the best feature in AUC respect to the rest of the features. Numbers within parentheses are the 95% confidence interval. NC = Normal control; MCI = Mild cognitive impairment; AD = Alzheimer disease; HV= Hippocampal volume; SR = Surface roughness; Λ = Annual atrophy rate in (%); FS= FreeSurfer; CC = Cross-sectional; LC= Hybrid framework; LL= proposed framework

implementations, the FS method generates the highest AUC values in the classification between NC vs MCI, whereas our LL approach best classifies the groups NC vs AD and MCI vs AD. The classification results from $l_i^{HV}$ or $l_i^{SR}$ with the LL implementation show improvements in the AUC with respect to the CC and LC methods, which become significant improvements, even when compared with the CC method trained with the combination of the marker features.

Given the LL implementation, the best AUC values were obtained by training the classifiers with $l_i^{HV}$ or $l_i^{SR}$. Figure 6 shows a comparison of the ROC curves using the different features used to classify the patients from the LL implementation. The worst AUC values were provided when training the classifiers with annual atrophy ratios, coinciding with other authors (Chincarini et al. 2016). However, in our LL approach, in the discrimination between MCI vs AD, atrophy measures improve the classification results compared to the FS method. As already indicated, similar AUC values were observed between the HV and SR markers. However, differences appear when comparing the ROC curves between the HV and SR markers, indicating that these measurements have complementary properties. These indications will be confirmed in the following experiment.

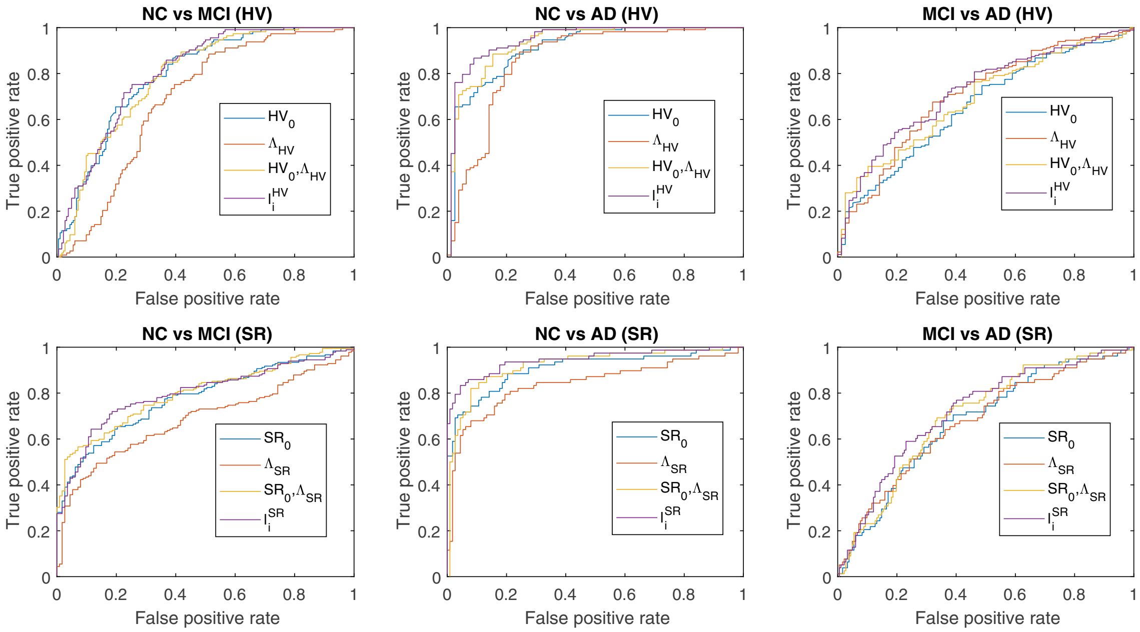

**Fig. 6** The four ROC curves from the features of the HV ($HV_0$, $\Lambda_{HV}$, $\{HV_0, \Lambda_{HV}\}$, and $l_i^{HV}$) and SR ($SR_0$, $\Lambda_{SR}$, $\{SR_0, \Lambda_{SR}\}$, and $l_i^{SR}$) markers using the LL method for detecting diagnosis between clinical groups: NC vs MCI, NC vs MCI and MCI vs AD. HV= Hippocampal volume; SR= Surface roughness; $\Lambda$ = Annual atrophy rate in (%); NC= Normal control; MCI= Mild cognitive impairment; AD= Alzheimer disease; LL=proposed framework

According to the AUC values reported, the classifiers trained with $l_i^{HV}$ or $l_i^{SR}$ exhibit the best performance. Thus, the next experiment was to analyze the SEN, SPE and ACC of the classifiers between pairs of clinical groups to be discriminated using $l_i^{HV}$ and $l_i^{SR}$. Table 6 summarizes the results of the classification in pairs of clinical groups for the four implementations using $l_i^{HV}$ and $l_i^{SR}$. The training of classifiers with $l_i^{HV}$ or $l_i^{SR}$ does not produce significant differences between the four competing implementations in any of the paired groups to be discriminated. The classifiers trained with $l_i^{HV}$ are verified to have greater SEN, while those classifiers trained with $l_i^{SR}$ offer higher values of SPE. The FS implementation can better discriminate between sMCI vs pMCI. The LL approach distinguishes the best between pMCI vs. AD, especially when trained with $l_i^{HV}$.

Finally, Fig. 7 shows the segmentation results of the FreeSurfer longitudinal framework and our 4D approach for one typical patient with scans at the baseline and a 12-month visit. The subject identification of the ADNI data is: 003_S_1057 (MCI). The first row shows the three views and the scans are overlapped with the segmentations of the FreeSurfer (FS) and our approach (LL) at the baseline. The second row shows the results at 12 months. The results show that the proposed segmentations preserve the shape learned from HarP atlases and the hippocampal surfaces are smoothed relative to the FreeSurfer segmentations. As a consequence of these results, the HV and SR markers in each clinical group, extracted from the proposed hippocampal segmentations, produce less dispersion than the markers obtained with FreeSufer method (see Table 4 and Fig. 5).

## Discussion

In this study, we evaluated the impact of using the longitudinal data from MRI scans for AD study. We compared four pipelines for extracting AD markers in a longitudinal context. Each method involved different combinations of registration and segmentation steps. The proposed longitudinal method (LL, longitudinal registration and segmentation) was compared with a cross-sectional implementation (CC) and a hybrid approach (LC) that uses the longitudinal registration and independent segmentation of each 3D ROI. This manner allowed us to analyze how each procedure contributes to improving predictions during the progression of AD. These results were also compared with the Freesurfer longitudinal framework (FS).

Our hippocampal segmentation algorithm is based on a) use of the HarP atlases, b) longitudinal registration and c) 4D segmentation. The main hypothesis in the simultaneous segmentation is based on the robust estimation of the probability model in the inner voxels of the hippocampus and an individualized estimation in the voxels around the hippocampal surface.

**Table 6** Results of the classifications between patients with different statuses using $l_i^{HV}$ and $l_i^{SR}$ in the selected ADNI database by means FreeSurfer v5.3 (FS), and the three implemented methods (CC, LC and LL)

| Subjects | Feature | Method | SEN (%) | SPE (%) | ACC (%) | AUC |
|---|---|---|---|---|---|---|
| NC & MCI | $l_i^{HV}$ | FS | 70.4(70.2, 70.6) | 75.6(75.4, 75.8) | 72.4(72.2, 72.5) | 0.820(0.819, 0.821) |
| | | CC | 70.0(69.8, 70.1) | 76.7(76.5, 76.9) | 72.5(72.4, 72.7) | 0.803(0.801, 0.804) |
| | | LC | 71.7(71.5, 71.9) | 75.4(75.2, 75.6) | 73.1(73.0, 73.3) | 0.806(0.805, 0.807) |
| | | LL | 70.9(70.7, 71.1) | 75.5(75.3, 75.7) | 72.7(72.5, 72.8) | 0.805(0.804, 0.806) |
| | $l_i^{SR}$ | FS | 69.2(69.0, 69.4) | 79.9(79.7, 80.1) | 73.3(73.2, 73.4) | 0.818(0.817, 0.819) |
| | | CC | 66.8(66.6, 67.0) | 79.4(79.2, 79.6) | 71.6(71.5, 71.7) | 0.796(0.794, 0.797) |
| | | LC | 64.9(64.7, 65.1) | 80.1(79.9, 80.3) | 70.7(70.6, 70.8) | 0.799(0.797, 0.800) |
| | | LL | 65.5(65.3, 65.7) | 79.1(78.9, 79.3) | 70.7(70.6, 70.8) | 0.799(0.798, 0.800) |
| NC & AD | $l_i^{HV}$ | FS | 86.1(85.9, 86.4) | 85.7(85.5, 85.9) | 83.5(83.3, 83.7) | 0.941(0.941, 0.942) |
| | | CC | 82.8(82.6, 83.1) | 89.7(89.6, 89.9) | 86.9(86.8, 87.0) | 0.928(0.927, 0.928) |
| | | LC | 81.9(81.6, 82.1) | 89.6(89.5, 89.8) | 86.4(86.3, 86.6) | 0.936(0.935, 0.936) |
| | | LL | 81.5(81.3, 81.8) | 90.4(90.2, 90.5) | 86.7(86.6, 86.9) | 0.947(0.946, 0.947) |
| | $l_i^{SR}$ | FS | 82.6(82.4, 82.9) | 90.7(90.6, 90.9) | 87.4(87.3, 87.5) | 0.945(0.944, 0.946) |
| | | CC | 84.4(84.2, 84.6) | 94.0(93.9, 94.1) | 90.1(90.0, 90.2) | 0.937(0.936, 0.938) |
| | | LC | 82.1(81.9, 82.3) | 94.9(94.8, 95.0) | 89.7(89.6, 89.8) | 0.942(0.941, 0.943) |
| | | LL | 81.3(81.1, 81.5) | 95.3(95.2, 95.4) | 89.6(89.4, 89.7) | 0.946(0.944, 0.947) |
| MCI & AD | $l_i^{HV}$ | FS | 72.8(72.6, 73.1) | 59.1(58.9, 59.3) | 63.2(63.1, 63.4) | 0.691(0.689, 0.692) |
| | | CC | 64.8(64.5, 65.1) | 60.0(59.8, 60.3) | 61.5(61.3, 61.6) | 0.679(0.677, 0.681) |
| | | LC | 64.7(64.4, 65.0) | 64.5(64.3, 64.7) | 64.6(64.4, 64.7) | 0.694(0.692, 0.695) |
| | | LL | 67.9(67.6, 68.2) | 64.8(64.6, 65.0) | 65.7(65.6, 65.9) | 0.720(0.718, 0.722) |
| | $l_i^{SR}$ | FS | 67.7(67.4, 68.0) | 67.4(67.2, 67.6) | 67.5(67.3, 67.6) | 0.700(0.699, 0.702) |
| | | CC | 60.1(59.8, 60.4) | 67.6(67.4, 67.8) | 65.4(65.2, 65.5) | 0.697(0.695, 0.698) |
| | | LC | 61.6(61.2, 61.9) | 68.3(68.1, 68.5) | 66.3(66.1, 66.4) | 0.701(0.698, 0.703) |
| | | LL | 65.1(64.8, 65.4) | 66.9(66.8, 67.1) | 66.4(66.3, 66.5) | 0.712(0.710, 0.713) |
| sMCI & pMCI | $l_i^{HV}$ | FS | 65.5(65.3, 65.8) | 58.4(58.1, 58.7) | 62.4(62.2, 62.5) | 0.617(0.615, 0.619) |
| | | CC | 57.6(57.4, 57.9) | 56.8(56.5, 57.1) | 57.3(57.1, 57.4) | 0.587(0.585, 0.589) |
| | | LC | 56.1(55.8, 56.3) | 56.2(55.9, 56.5) | 56.1(56.0, 56.3) | 0.589(0.586, 0.591) |
| | | LL | 55.9(55.7, 56.2) | 56.4(56.1, 56.7) | 56.2(56.0, 56.3) | 0.590(0.588, 0.592) |
| | $l_i^{SR}$ | FS | 57.8(57.5, 58.0) | 59.8(59.5, 60.1) | 58.7(58.5, 58.8) | 0.615(0.613, 0.617) |
| | | CC | 51.3(51.0, 51.5) | 57.6(57.3, 57.9) | 54.1(53.9, 54.3) | 0.589(0.587, 0.591) |
| | | LC | 53.3(53.0, 53.6) | 58.8(58.5, 59.2) | 55.8(55.6, 56.0) | 0.591(0.589, 0.593) |
| | | LL | 52.7(52.4, 53.0) | 58.8(58.5, 59.1) | 55.4(55.2, 55.6) | 0.594(0.592, 0.596) |
| pMCI & AD | $l_i^{HV}$ | FS | 69.1(68.9, 69.4) | 56.2(55.9, 56.5) | 61.8(61.7, 62.0) | 0.640(0.638, 0.642) |
| | | CC | 63.2(62.9, 63.5) | 57.1(56.9, 57.4) | 59.8(59.6, 59.9) | 0.640(0.638, 0.642) |
| | | LC | 63.8(63.5, 64.1) | 54.2(54.0, 54.5) | 58.4(58.2, 58.5) | 0.641(0.639, 0.643) |
| | | LL | 64.2(63.9, 64.5) | 60.4(60.1, 60.7) | 62.1(61.9, 62.3) | 0.670(0.668, 0.672) |
| | $l_i^{SR}$ | FS | 64.9(64.6, 65.2) | 65.0(64.7, 65.2) | 65.0(64.8, 65.1) | 0.651(0.649, 0.653) |
| | | CC | 57.2(56.9, 57.5) | 64.9(64.6, 65.2) | 61.6(61.4, 61.7) | 0.654(0.652, 0.656) |
| | | LC | 58.1(57.8, 58.4) | 65.2(64.9, 65.4) | 62.1(61.9, 62.2) | 0.654(0.652, 0.656) |
| | | LL | 59.4(59.1, 59.7) | 64.0(63.8, 64.3) | 62.0(61.8, 62.2) | 0.653(0.651, 0.655) |

Numbers within parentheses are the 95% confidence interval. NC = Normal control; sMCI = Stable Mild cognitive impairment; pMCI = Progressive Mild cognitive impairment; AD= Alzheimer disease; SEN= sensitivity; SPE = specificity; ACC= accuracy; AUC = Area under curve

The test-retest experiments with the MIRIAD data demonstrated that the LL pipeline gave smaller volume errors and significant improvements in the dice overlap comparison compared with the FS, CC and LC pipelines. The longitudinal registration consistently aligned all time points of a subject, increasing the test/retest reliability. As

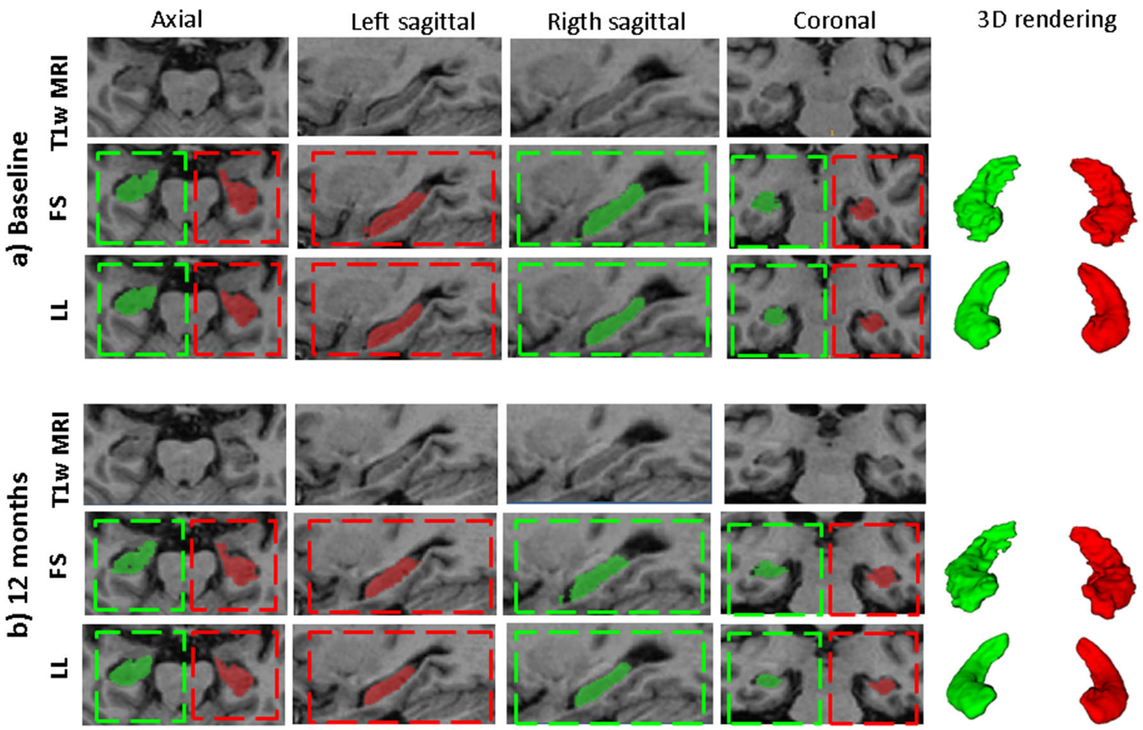

**Fig. 7** Comparison between FreeSurfer longitudinal framework (FS) and the proposed method (LL).The first row shows the three views and the scans are overlapped with the segmentations of the FS and our approach at the baseline. The ROIs are also illustrated by dashed line boxes. The second row shows the results at 12 months

the registration was better, the positions of the HarP atlases were more consistent across time points, thus reducing one source of variance in the hippocampal segmentations.

Better reliability can be obtained by reporting the same segmentation across time points (over-constraining the method). However, this improvement produces limits the detection of longitudinal changes and group differences. To overcome this issue, a longitudinal approach should be a trade-off between noise reduction and over-regularization by keeping the model flexible enough to capture hippocampal atrophies over time (Iglesias et al. 2016). The LL implementation also showed the highest power of discrimination using annual hippocampal atrophy rates compared to the other methods. Moreover, the annual hippocampal volume atrophy rates, i.e. $\Lambda_{HV}$, for the four clinical groups of AD were in the range of those reported by previous publications. Regarding the classification of subjects between the paired clinical groups of AD, the LL method showed the best AUC performance between the NC vs AD, MCI vs AD and pMCI vs AD groups.

Bernal-Rusiel et al. (2013) have performed a longitudinal analysis using T1w MRI for AD study. In this work, 791 subjects with a follow-up of up to 4 years and with a total number of 3177 scans were analyzed from the ADNI repository. The data were processed with the Freesurfer longitudinal framework. The authors used two markers: a) Mean thickness within the entorhinal cortex and b) HV. Demographic, clinical and markers data are publicly available and are included in the LME code distribution.

Since our population sample of study is smaller than the one analyzed by Bernal et al., we speculate whether our results are consistent when applying the designed tools of the analysis of the data to the new enlarged population. The algorithms for the calculation of HV atrophy and the classification between clinical groups from $l_i^{HV}$ were applied to the Bernal data (see Eqs. (7) and (8), respectively).

The HV marker was used as a reference to compare the results of the analysis and classification extracted from the Bernal data and the data used in this work. Nevertheless, the Bernal data report the HV in the native space of each scan. Therefore, HV markers were normalized by the intracranial volume of each subject and rescaled to the MNI152 space. Table 7 shows the statistical analysis of the HV at baseline and their annual hippocampal volume atrophy over the clinical groups calculated using the Bernal data. The $HV_0$ and $\Lambda_{HV}$ distributions among the clinical groups coincide with those obtained with our population using the FS method. We also found that the annual HV atrophy discriminates better than the HV marker at baseline, showing significant differences between the sMCI vs pMCI and pMCI vs AD groups. This conclusion was also observed in our population sample (see Table 4).

Table 8 shows the classification scores between the clinical groups of AD applied to the Bernal data using

**Table 7** MANCOVA with Bonferroni pairwise comparisons is used for analysing the HV marker at baseline and atrophy rates from the Bernal data ($HV_0$, $\Lambda_{HV}$)

| Feature | Method | NC | sMCI | pMCI | AD | F | p |
|---|---|---|---|---|---|---|---|
| $HV_0$ | FS | 8820(1132) | 7883(1300)$^{a}$ | 6965(1119)$^{a,b}$ | 6687(1244)$^{a,b}$ | 126.19 | $< 0.001$ |
| $\Lambda_{HV}$ | FS | −1.28(0.73) | −2.07(0.94)$^{a}$ | −3.51(1.22)$^{a,b}$ | −4.26(1.15)$^{a,b,c}$ | 354.81 | $< 0.001$ |

Data as represented as mean and standard deviation (SD). Statistical significance is considered with $p - value < 0.01$. $^a$ Significant compared to NC. $^b$ Significant compared to sMCI. $^c$ Significant compared to pMCI. NC= Normal control; sMCI = Stable Mild cognitive impairment; pMCI = Progressive Mild cognitive impairment; AD= Alzheimer disease; HV = Hippocampal volume; SR = Surface roughness; $\Lambda$ = Annual atrophy rate in (%); FS=FreeSurfer

$l_i^{HV}$ as the feature for training the classifiers. The results of the SEN, SPE, ACC and AUC values among the three main groups (i.e. NC vs MCI, NC vs AD and MCI vs AD) are almost identical with those obtained with our population sample using the FS method (see Table 6). They show discrepancies in the discrimination between the sMCI vs pMCI and pMCI vs AD groups. While better discernment between the sMCI vs pMCI groups appears in the Bernal data, this discriminating ability falls substantially between the pMCI vs AD groups. As already mentioned, the pMCI patient group is a high-risk cohort in which the subjects will undergo an onset of AD in the near future. Then, the within-group homogeneity assumption is violated, which in turns impacts statistical inference. Therefore, since the classification scores between MCI vs AD in the two databases are similar, the discrimination discrepancies between sMCI vs pMCI and pMCI vs AD can be explained by the difficulty associated with diagnosing the pMCI group.

The highest consumption of computational time in our approach occurs in the longitudinal registration step, which is performed by the FreeSurfer longitudinal framework.

Regarding the hippocampal segmentation algorithm and with the objective of calculating unary potentials, the computational complexity is primarily due to the non-rigid registrations of the selected HarP atlases into the 3D ROI for each time-point. The task of non-rigid registrations has been parallelized. The registration of the first 15 HarP atlases in a 3D ROI requires less than 45 seconds ($N_R = 15$ (Platero and Carmen Tobar 2015), [Dual CPU] Intel Xeon E5520 @ 2.27 GHz with 24 GB of RAM). The label fusion method using non-rigid registrations calculates the unary and pairwise potentials of the CRF model (Eq. (1)), and the labeling is found by applying the min-cut/max-flow algorithm of Boykov and Kolmogorov (2004). The scripts used in this study are available at https://www.nitrc.org/projects/longhippsegm/.

Future work will follow four directions. First, the 4D hippocampal segmentation algorithms will be improved by means of the patch-based label fusion methods, which can be extended to the longitudinal analysis. In this manner, PatchMatch (Barnes et al. 2009) will be used in the field of the label fusion methods and will also be based on a multi-feature framework (Giraud et al. 2016; Platero and Tobar 2017). Second, our experiments have shown that the SR marker has similar discrimination power as the HV marker. The advantage of the SR marker is that it also offers a local analysis of atrophy on the hippocampal surface. A voxel-based analysis will be performed to find regions of the hippocampal surface that are the most discriminating among clinical groups for AD study (Chételat et al. 2008; La Joie et al. 2010; Perrotin et al. 2015). Third, we should evaluate the utility of longitudinal neuroimaging markers for predicting the conversion from MCI to AD. A LME model combined with a Cox regression model will be used to examine the relationship between time-dependent markers and the timing of the conversion to AD (Devanand et al. 2007; Sabuncu et al. 2014). Finally and fourthly, we should investigate the possibility of analyzing the longitudinal residues of several markers simultaneously in order to improve the classification results, as recent

**Table 8** Results of the classifications between patients with different statuses using $l_i^{HV}$ as the feature for training the classifiers on Bernal data by means FreeSurfer v5.1 (FS)

| Subjects | SEN (%) | SPE (%) | ACC (%) | AUC |
|---|---|---|---|---|
| NC & MCI | 71.7(71.6, 71.8) | 76.2(76.1, 76.4) | 73.3 (73.2,73.4) | 0.812(0.811, 0.812) |
| NC & AD | 84.5(84.3, 84.8) | 84.8(84.6, 84.9) | 84.6(84.5, 84.7) | 0.928(0.927, 0.928) |
| MCI & AD | 69.7(69.5, 69.9) | 59.8(59.6, 59.9) | 63.0(62.9, 63.1) | 0.702(0.701, 0.703) |
| sMCI & pMCI | 68.5(68.4, 66.8) | 66.8(66.8, 67.0) | 67.5(67.4, 67.7) | 0.729(0.728, 0.730) |
| pMCI & AD | 64.2(64.0, 64.4) | 52.1(51.9, 52.3) | 58.5(58.4, 58.6) | 0.588(0.587, 0.590) |

NC = Normal control; sMCI = Stable Mild cognitive impairment; pMCI = Progressive Mild cognitive impairment; AD = Alzheimer disease; SEN = sensitivity; SPE = specificity; ACC = accuracy; AUC = Area under curve publications have proposed (Moradi et al. 2015; Eskildsen et al. 2015; Korolev et al. 2016).

## Conclusions

We proposed a longitudinal method to analyze longitudinal datasets from brain T1w MRI for AD study. The framework comprises two innovative parts: a longitudinal segmentation and a longitudinal classification step. The results showed that both steps of the longitudinal pipeline improved the reliability and the accuracy of the detection between clinical groups of AD. The contributions of this work could be summarized as follows: (1) A fast 4D hippocampal segmentation with constraints on atrophy and supported by the HarP atlases has been presented, showing better reliability in test/retest experiments. These hippocampal segmentations also improve the differential diagnosis between NC vs AD, MCI vs AD and pMCI vs AD. (2) The training of the classifiers using the longitudinal trajectory residue produces better results in terms of the AUC than the other classical features employed. (3) The quality control criteria on the longitudinal processing have been expanded when analyzing the consistency between the scans and hippocampal segmentations of a subject sequence. (4) The analysis using MANCOVA of the hippocampal atrophy markers has shown that the proposed method has a greater discriminatory power among the clinical groups of AD than the other implementations. (5) For any of the longitudinal trajectory residues from the longitudinal implementations, compared to the marker at baseline, the classification results show significant improvements. (6) Our population sample from the ADNI repository is sufficiently representative of the progression of AD, as our results show similar trends to those obtained from the Bernal data, which are extended to more subjects and a longer temporal duration follow-up.

## Information Sharing Statement

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

All source used in this manuscript are freely available to the public (Longitudinal neuroimaging hippocampal markers for diagnosing Alzheimer's disease; RRID:SCR_016282). These data are available from the website cited in "Discussion".

**Acknowledgments** Data collection for this study was funded by the ADNI (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company.

## References

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270–279.

Apostolova, L.G., Morra, J.H., Green, A.E., Hwang, K.S., Avedissian, C., Woo, E., Cummings, J.L., Toga, A.W., Jack, C.R., Weiner, M.W., et al. (2010). Automated 3D mapping of baseline and 12-month associations between three verbal memory measures and hippocampal atrophy in 490 ADNI subjects. *NeuroImage*, *51*(1), 488–499.

Artaechevarria, X., Muñoz-barrutia, A., Ortiz-de Solorzano, C. (2009). Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE Transactions on Medical Imaging*, *28*(8), 1266–1277.

Aubert-Broche, B., Fonov, V.S., García-Lorenzo, D., Mouiha, A., Guizard, N., Coupé, P., Eskildsen, S.F., Louis Collins, D. (2013). A new method for structural volume analysis of longitudinal brain mri data and its application in studying the growth trajectories of anatomical brain structures in childhood. *NeuroImage*, *82*, 393–402.

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. (2009). Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, *28*(3), 24.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R. (2013). Alzheimer's Disease Neuroimaging Initiative, et al. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *NeuroImage*, *66*, 249–260.

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J.G., de Leon, M.J., de Toledo-Morrell, L., et al. (2011). Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. *Journal of Alzheimer's Disease*, *26*, 61–75.

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1124–1137.

Bradley, T., Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & Dementia*, *9*(3), 332–337.

Chételat, G., Fouquet, M., Kalpouzos, G., Denghien, I., De La Sayette, V., Viader, F., Mézenge, F., Landeau, B., Baron, J.-C., Eustache, F., et al. (2008). Three-dimensional surface mapping of hippocampal atrophy progression from MCI to AD and over normal aging as assessed using voxel-based morphometry. *Neuropsychologia*, *46*(6), 1721–1731.

Chincarini, A., Sensi, F., Rei, L., Gemme, G., Squarcia, S., Longo, R., Brun, F., Tangaro, S., Bellotti, R., Amoroso, N., et al. (2016). Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease. *NeuroImage*, *125*, 834–847.

Clifford, R., Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Louis Collins, D. (2011). Patch-based segmentation using expert priors application to hippocampus and ventricle segmentation. *NeuroImage*, *54*(2), 940–954.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *NeuroImage*, *56*(2), 766–781.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 837–845.

Devanand, D.P., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G.H., Honig, L.S., Mayeux, R., et al. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment prediction of Alzheimer disease. *Neurology*, *68*(11), 828–836.

Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, *6*(8), 734–746.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Louis Collins, D. (2015). Structural imaging biomarkers of alzheimer's disease: predicting disease progression. *Neurobiology of Aging*, *36*, S23–S31.

Evans, A.C., Janke, A.L., Louis Collins, D., Baillet, S. (2012). Brain templates and atlases. *NeuroImage*, *62*(2), 911–922.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Folstein, M.F., Folstein, S.E., McHugh, P.R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Fox, N.C., Ridgway, G.R., Schott, J.M. (2011). Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials. *NeuroImage*, *57*(1), 15–18.

Frankó, E., & Joly, O. (2013). Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. *PloS one*, *8*(8), e71354.

Fraser, M.A., Shaw, M.E., Cherbuin, N. (2015). A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. *NeuroImage*, *112*, 364–374.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 67–77.

Frisoni, G.B., Jack, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu, Y., Preboske, G., Swihart, T., Blair, M., Cavedo, E., et al. (2015). The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimer's & Dementia*, *11*(2), 111–125.

Giraud, R., Ta, V.-T., Papadakis, N., Manjón, J.V., Louis Collins, D., Coupé, P., ADNI et al. (2016). An optimized patchmatch for multi-scale and multi-feature label fusion. *NeuroImage*, *124*, 770–782.

Iglesias, J.E., Van Leemput, K., Augustinack, J., Insausti, R., Fischl, B., Reuter, M. (2016). ADNI, et al. Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases. *NeuroImage*, *141*, 542–555.

Jack, C.R., Shiung, M.M., Gunter, J.L., Obrien, P.C., Weigand, S.D., Knopman, D.S., Boeve, B.F., Ivnik, R.J., Smith, G.E., Cha, R.H., et al. (2004). Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology*, *62*(4), 591–600.

Jenkinson, M., Bannister, P., Brady, M., Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841.

Kim, J., Valdes-Hernandez, M.D.C., Royle, N.A., Park, J. (2015). Hippocampal shape modeling based on a progressive template surface deformation and its verification. *IEEE Transactions on Medical Imaging*, *34*(6), 1242–1261.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Louis Collins, D., Gee, J., Hellier, P., et al. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, *46*(3), 786–802.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W. (2010). Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, *29*(1), 196–205.

Korolev, I.O., Symonds, L.L., Bozoki, A.C. (2016). Alzheimer?s Disease Neuroimaging Initiative, et al. Predicting progression from mild cognitive impairment to alzheimer's dementia using clinical, mri, and plasma biomarkers via probabilistic pattern classification. *Plos One*, *11*(2), e0138866.

La Joie, R., Fouquet, M., Mézenge, F., Landeau, B., Villain, N., Mevel, K., Pélerin, A., Eustache, F., Desgranges, B., Chételat, G. (2010). Differential effect of age on hippocampal subfields assessed using a new high-resolution 3T MR sequence. *Neuroimage*, *53*(2), 506–514.

Lafferty, J., McCallum, A., Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International conference on machine learning* (pp. 282–289).

Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., Anderson, R.M. (2017). A systematic review of longitudinal studies which measure alzheimers disease biomarkers. *Journal of Alzheimer's Disease*, *59*(4), 1359–1379.

Leung, K.K., Barnes, J., Ridgway, G.R., Bartlett, J.W., Clarkson, M.J., Macdonald, K., Schuff, N., Fox, N.C., Ourselin, S., et al. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *NeuroImage*, *51*(4), 1345–1359.

Lotjonen, J.M.P., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, *49*(3), 2352–2365.

Louis Collins, D., & 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.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M. (2013). MIRIAD - Public release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage*, *70*, 33–36.

Mert, R., Sabuncu, M.R., Desikan, R.S., Sepulcre, J., Yeo, B.T.T., Liu, H., Schmansky, N.J., Reuter, M., Weiner, M.W., Buckner, R.L., Sperling, R.A., et al. (2011). The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of neurology*, *68*(8), 1040–1048.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J. (2015). Alzheimer's Disease Neuroimaging Initiative et al. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. *NeuroImage*, *104*, 398–412.

Mount, D.M., & Arya, S. (2010). Ann: A library for approximate nearest neighbor searching. http://www.cs.umd.edu/mount/ANN/. (Accessed: 20 January 2015). version 1.1.2.

Nestor, S.M., Gibson, E., Gao, F.-Q., Kiss, A., Black, S.E. (2013). A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in Alzheimer's disease. *NeuroImage*, *66*, 50–70.

Osher, S., Fedkiw, R., Piechor, K. (2004). Level set methods and dynamic implicit surfaces. *Applied Mechanics Reviews*, *57*, B15. https://doi.org/10.1115/1.1760520.

Perrotin, A., de Flores, R., Lamberton, F., Poisnel, G., La Joie, R., de la Sayette, V., Mezenge, F., Tomadesso, C., Landeau, B., Desgranges, B., et al. (2015). Hippocampal subfield volumetry and 3D surface mapping in subjective cognitive decline. *Journal of Alzheimer's Disease*, *48*(s1), S141–S150.

Platero, C., & Carmen Tobar, M. (2015). A label fusion method using conditional random fields with higher-order potentials Application to hippocampal segmentation. *Artificial Intelligence in Medicine*, *64*(2), 117–129.

Platero, C., & Carmen Tobar, M. (2016). A fast approach for hippocampal segmentation from t1-MRI for predicting progression in Alzheimer's disease from elderly controls. *Journal of Neuroscience Methods*, *270*, 61–75.

Platero, C., & Tobar, M.C. (2017). Combining a patch-based approach with a non-rigid registration-based label fusion method for the hippocampal segmentation in Alzheimer's Disease. *Neuroinfor matics*, *15*(2), 165–183.

Reuter, M., Diana Rosas, H., Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *NeuroImage*, *53*(4), 1181–1196.

Reuter, M., Schmansky, N.J., Diana Rosas, H., Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, *61*(4), 1402–1418.

Sabuncu, M.R., Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B. (2014). ADNI, et al. Event time analysis of longitudinal neuroimage data. *NeuroImage*, *97*, 9–18.

Schröder, J., & Pantel, J. (2016). Neuroimaging of hippocampal atrophy in early recognition of Alzheimer´s disease–a critical appraisal after two decades of research. *Psychiatry Research: Neuroimaging*, *247*, 71–78.

Sled, J.G., Zijdenbos, A.P., Evans, A.C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging*, *17*(1), 87–97.

Song, Z., Tustison, N., Avants, B., Gee, J. (2006). Integrated graph cuts for brain MRI segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, *4191*, 831–838.

Thompson, W.K., Hallmayer, J., O'Hara, R. (2011). Design considerations for characterizing psychiatric trajectories across the lifespan Application to effects of APOE-$\varepsilon$4 on cerebral cortical thickness in Alzheimer's disease. *American Journal of Psychiatry*, *168*(9), 894–903.

van der Lijn, F., den Heijer, T., Breteler, M., Niessen, W.J. (2008). Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, *43*(4), 708–720.

Viola, P., & Wells, W.M.I. (1997). Alignment by maximization of mutual information. *International Journal of Computer Vision*, *24*(2), 137–154.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 611–623.

Wang, L., Guo, Y., Cao, X., Wu, G., Shen, D. (2016). Consistent multi-atlas hippocampus segmentation for longitudinal MR brain images with temporal sparse representation. In *International workshop on patch-based techniques in medical imaging (pp. 34–42). Springer*.

Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D. (2010). Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. *NeuroImage*, *52*(1), 109–118.