NeuroImage 141 (2016) 542–555

Contents lists available at ScienceDirect

NeuroImage

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

Bayesian longitudinal segmentation of hippocampal substructures in brain MRI using subject-specific atlases

Juan Eugenio Iglesias[a,b,*], Koen Van Leemput[c,d], Jean Augustinack[c], Ricardo Insausti[e], Bruce Fischl[c,f], Martin Reuter[c,f], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Basque Center on Cognition, Brain and Language, Spain
[b] Translational Imaging Group, University College London, United Kingdom
[c] Massachusetts General Hospital and Harvard Medical School, USA
[d] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[e] Human Neuroanatomy Laboratory, University of Castilla-La Mancha, Spain
[f] MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), USA

ARTICLE INFO

Article history:
Received 18 March 2016
Accepted 7 July 2016
Available online 15 July 2016

Keywords:
Hippocampal subfields
Longitudinal modeling
Segmentation
Bayesian modeling

ABSTRACT

The hippocampal formation is a complex, heterogeneous structure that consists of a number of distinct, interacting subregions. Atrophy of these subregions is implied in a variety of neurodegenerative diseases, most prominently in Alzheimer's disease (AD). Thanks to the increasing resolution of MR images and computational atlases, automatic segmentation of hippocampal subregions is becoming feasible in MRI scans. Here we introduce a generative model for dedicated longitudinal segmentation that relies on subject-specific atlases. The segmentations of the scans at the different time points are jointly computed using Bayesian inference. All time points are treated the same to avoid processing bias. We evaluate this approach using over 4700 scans from two publicly available datasets (ADNI and MIRIAD). In test–retest reliability experiments, the proposed method yielded significantly lower volume differences and significantly higher Dice overlaps than the cross-sectional approach for nearly every subregion (average across subregions: 4.5% vs. 6.5%, Dice overlap: 81.8% vs. 75.4%). The longitudinal algorithm also demonstrated increased sensitivity to group differences: in MIRIAD (69 subjects: 46 with AD and 23 controls), it found differences in atrophy rates between AD and controls that the cross sectional method could not detect in a number of subregions: right parasubiculum, left and right presubiculum, right subiculum, left dentate gyrus, left CA4, left HATA and right tail. In ADNI (836 subjects: 369 with AD, 215 with early cognitive impairment — eMCI — and 252 controls), all methods found significant differences between AD and controls, but the proposed longitudinal algorithm detected differences between controls and eMCI and differences between eMCI and AD that the cross sectional method could not find: left presubiculum, right subiculum, left and right parasubiculum, left and right HATA. Moreover, many of the differences that the cross-sectional method already found were detected with higher significance. The presented algorithm will be made available as part of the open-source neuroimaging package FreeSurfer.

Introduction

Background

The study of the human hippocampus has traditionally attracted considerable attention from the neuroscience and neuroimaging communities due to its connection with memory (Eldridge et al., 2000; Scoville and Milner, 1957) and an array or neurological disorders, especially Alzheimer's disease (AD) (Apostolova et al., 2006; Du et al., 2001; Laakso et al., 1998). Limits in MR acquisition have for many years forced *in vivo* studies to treat the hippocampus as a single structure. However, the hippocampus consists of a number

* Corresponding author at: Translational Imaging Group, University College London, United Kingdom.
E-mail address: e.iglesias@bcbl.eu (J. Iglesias).

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

http://dx.doi.org/10.1016/j.neuroimage.2016.07.020

of subregions that have been shown to have different memory functions using animal models (Kesner, 2007; Rolls, 2010). In humans, there is increasing evidence that hippocampal subregions play different roles in memory (Gabrieli et al., 1997; Kesner, 2007; Knierim et al., 2006; Zeidman and Maguire, 2016), and that they are differently affected by AD (Arnold et al., 1991; Braak and Braak, 1991). Therefore, *in vivo* analysis of hippocampal subregions holds great promise to improve our understanding of normal aging and AD, as well as to deliver more sensitive biomarkers of AD and other neurological disorders.

Recent advances in MRI acquisition have made it possible to study the hippocampal subregions *in vivo*. Earlier studies had to rely on manual segmentations (Burggren et al., 2008; Mueller et al., 2007), typically performed on T2 scans acquired coronally with high in-plane resolution and relatively thick slices. Automated methods have since been proposed to bypass the manual segmentation procedure, which requires extensive expertise, is extremely time consuming, and cannot be reproduced easily. Yushkevich et al. (2010b, 2015) proposed a multi-atlas segmentation algorithm using a library of manually labeled T1 and T2 scans, whose output was refined by a machine learning bias correction strategy. Wang et al. (2006, 2015) employed a surface-based atlas approach. Our group, in previous work, used a probabilistic atlas to produce segmentations with a Bayesian inference algorithm within a generative framework. In a first version (Van Leemput et al., 2009), the atlas was constructed using high-resolution *in vivo* MRI scans (coronal slices with .38 mm in-plane resolution, .8 mm slice separation). More recently, we acquired ultra-high resolution *ex vivo* MRI, which enabled us to produce very detailed manual segmentations and, in turn, a much more accurate atlas (Iglesias et al., 2015). It is the use of generative techniques that enables the application of *ex vivo* atlases to the segmentation of *in vivo* scans, since they do not require the intensity characteristics of the training and test datasets to match — in contrast with registration-based algorithms such as Yushkevich's and Wang's.

Many large scale studies, including the Alzheimer's Disease Neuroimaging Initiative (ADNI), are now collecting longitudinal MRI data. Since they remove the confounding inter-subject variability, longitudinal studies enable us to accurately quantify within-subject neuroanatomical changes, and provide higher sensitivity than their cross-sectional counterparts (Fitzmaurice et al., 2012). However, until now, no dedicated method exists (to the best of our knowledge) for the longitudinal segmentation of hippocampal subregions.

In this paper, we introduce a novel Bayesian approach for the joint segmentation of hippocampal subregions across multiple time points. The method is based on a generative model of longitudinal MRI scans, extending our cross-sectional approach (Iglesias et al., 2015) to longitudinal datasets. Rather than by a population-wide atlas, the scans at the different time points are assumed to have been generated by a subject-specific atlas, which introduces a statistical dependence between the time points and ensures that the different images and corresponding segmentations are similar to each other. This subject-specific atlas is simply a deformed version of the population-wide atlas. Within this framework, the segmentations of all time points are computed simultaneously with a Bayesian inference algorithm; the subject-specific atlas is obtained as a by-product. Due to its generative nature and unsupervised intensity model, the algorithm is robust against changes in MRI contrast.

Further related work on longitudinal segmentation

Longitudinal segmentation algorithms exploit the prior knowledge that a set of images belongs to the same subject, in order to produce more accurate and consistent segmentations than when the images are processed independently. A crucial aspect of longitudinal methods is the need to keep them unbiased: algorithms that do not treat all time points the same way introduce processing bias due to the additional processing steps applied to selected images (Reuter and Fischl, 2011).

Many longitudinal segmentation approaches rely on a non-linear, group-wise registration that brings the images from the different time points into a common coordinate space. The registration should be computed in an intermediate space (Smith et al., 2002), in order to avoid biases due to image resampling in the space to a selected scan — typically the baseline (Thompson et al., 2011; Yushkevich et al., 2010a). In some methods, the group-wise alignment is precomputed with a registration algorithm. For example, Gao et al. (2014) used pre-aligned scans to optimize a cost function that included an intensity correction term matching the intensity profiles across time points. Other approaches integrate the registration into the segmentation framework. For instance, Shi et al. (2010b) used a multi-channel (T1/T2) segmentation algorithm guided by prior tissue probability maps; the spatial mapping of the tissue maps across time points was estimated simultaneously with the segmentation using an expectation maximization algorithm. Xue et al. (2006, 2010) proposed a similar approach, which iteratively used the estimate of the segmentations to update the registrations and vice versa.

Some approaches do not require non-linear registration to produce the segmentations — though rigid registrations are still used to bring the images into rough alignment. In the context of whole hippocampus segmentation, Wolz et al. (2010) built a 4D graph in which a voxel had 6 spatial neighbors and 2 temporal neighbors (from the preceding and following time points). In their model, unary terms included intensity and anatomical priors, whereas pair-wise terms were engineered to enforce spatial and temporal smoothness in the segmentation. The segmentation of all time points was then computed simultaneously with graph cuts. In a similar framework, Bauer et al. (2014) used a random forest classifier in the unary term. Other papers have exploited expert knowledge to drive the segmentation. For example, Wang et al. (2011) constrained the distance across the serial images to remain within a biologically plausible range, and used a similar strategy in a more recent paper Wang et al. (2013) to segment the brain cortex (keeping the thickness within a reasonable range).

Finally, some longitudinal segmentation approaches have used a subject-specific atlas to produce consistent segmentations. In the context of neonate brain segmentation, Shi et al. (2010a) registered a population-wide atlas to the latest time point, which is normally the most reliable one in infants (least motion, and most contrast between gray and white matter), in order to produce subject-specific tissue probability maps. Rather than using a single time point as the target of the registration, Aubert-Broche et al. (2013) built a subject-specific atlas by non-linearly coregistering the time points; then, they registered a population-wide atlas to the output to obtain subject-specific probability maps.

Contribution: an unbiased, longitudinal segmentation method for hippocampal subregions based on a subject-specific atlas

The contribution of this article is twofold. In first place, it presents the first available automated algorithm for longitudinal segmentation of the hippocampal subregions; prior works have only addressed the longitudinal segmentation of the hippocampus as a whole (Wolz et al., 2010). Additionally, it presents a novel generative model for longitudinal segmentation based on subject-specific atlases, which is unbiased and adaptive to changes in MRI contrast. The models assumes that the images are generated by a hidden subject-specific atlas, which is in turn generated by a population-wide atlas. Even though the idea of using subject-specific atlases is not original, our model is novel: as opposed to works like Aubert-Broche et al. (2013), we estimate the subject-specific atlas along with the registrations and segmentations in a probabilistic framework, rather than precomputing it based solely on image intensities. This has the advantage

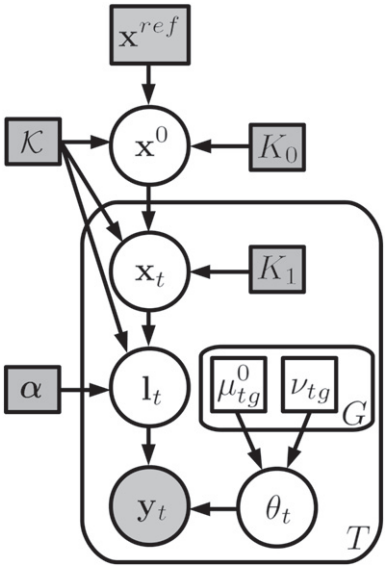

Fig. 1. Generative model for longitudinal MRI data. Random variables are in circles, parameters in squares. Shaded variables are observed. Plates indicate replication.

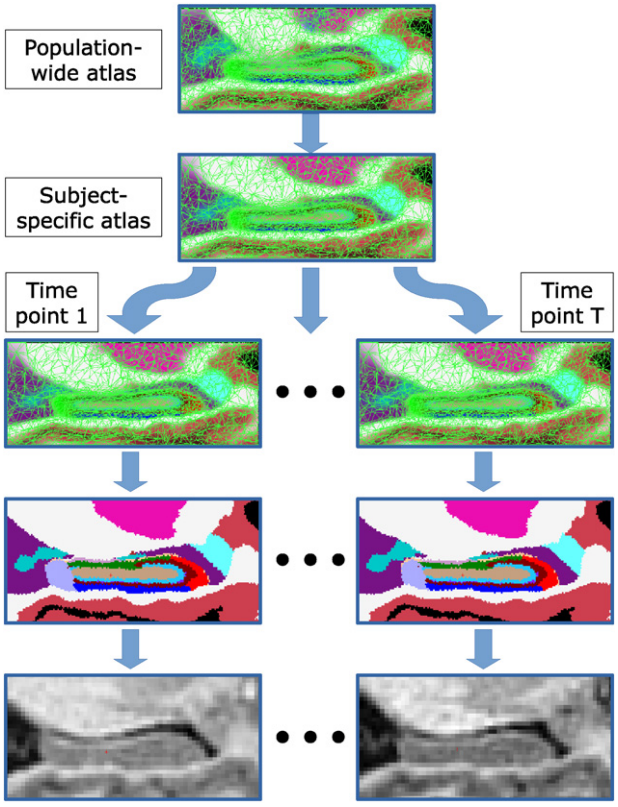

Fig. 2. Illustration of the generative process through which the longitudinal MRI data are assumed to be generated: the population-wide atlas is first deformed into a subject-specific atlas, which is subsequently deformed T times — once per time point. Segmentations are sampled from these deformed atlases, and image intensities are generated from the segmentations through a Gaussian mixture model.

that the segmentation and registration can iteratively improve each other.

The rest of this paper is organized as follows. Section "Methods" describes the generative framework that our proposed approach is based on, as well as the Bayesian inference algorithm that we used to obtain the segmentations. In Section "Experiments and Results", we describe a set of experiments that evaluated the test–retest reliability and sensitivity to group differences; since the hippocampal subregions cede to neurodegenerative pathology that worsens over time, we tested our approach on two public MRI datasets of AD patients (ADNI and MIRIAD). The experiments compared our algorithm with two competing methods; the results are further analyzed in Section "Discussion", while Section "Conclusion" closes the article.

Methods

Our segmentation framework is based on a generative model of longitudinal MRI data. In this section, we first describe the forward generative model, in which longitudinal MRI scans are assumed to have been generated by a probabilistic atlas of anatomy. Then, we present an inference algorithm that "inverts" the model with Bayes rule in order to estimate longitudinal segmentations from MRI data.

Forward generative model of longitudinal MRI scans

Let $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T\}$ be the image intensities of a set of T longitudinal MRI scans from the same subject. Each scan is represented by a vector of intensities corresponding to J voxels, i.e., $\boldsymbol{y}_t = [y_{t1},\ldots,y_{tJ}]$. Here we follow the literature of probabilistic atlases with unsupervised intensity models (Ashburner and Friston, 2005; Pohl et al., 2006; Van Leemput, 2009; Van Leemput et al., 1999), but modify the framework in order to adapt it to the longitudinal nature of the data. The image intensities are assumed to have been generated by the following process (the graphical model is displayed in Fig. 1, and further illustrated in Fig. 2):

i) We are given a probabilistic, population-wide atlas of anatomy, which is encoded as a tetrahedral mesh (Van Leemput, 2009) that covers the region of interest (in our case, a cuboid containing the hippocampus). The mesh is defined by its position $\boldsymbol{x}_{ref}$ (a vector with the coordinates of its N nodes) and its connectivity $\mathcal{K}$. Each node n has a corresponding vector of label probabilities $\boldsymbol{\alpha}_n = [\alpha_{n1},\ldots,\alpha_{nL}]$, where α_{nl} is the frequency with which label l is expected at node n, and L is the number of neuroanatomical labels modeled by the atlas.

ii) The mesh is deformed from its reference position $\boldsymbol{x}_{ref}$ to a new position $\boldsymbol{x}_0$, which is specific to the subject at hand, and yields the corresponding *subject-specific atlas*. The deformation is governed by a prior probability distribution that penalizes deformations and explicitly forbids collapsing tetrahedra, thereby preserving the topology of the mesh (Ashburner et al., 2000):

$$p(\boldsymbol{x}_0) \propto \exp\left[-K_0 \sum_d U_d^{\mathcal{K}}(\boldsymbol{x}_0, \boldsymbol{x}_{ref})\right], \quad (1)$$

where d loops over the tetrahedra in the mesh, K_0 is the stiffness parameter, and $U_d^{\mathcal{K}}(\boldsymbol{x}_0, \boldsymbol{x}_{ref})$ is the cost of deforming the d^{th} tetrahedron (see further details in Ashburner et al. (2000)).

iii) The mesh in position $\boldsymbol{x}_0$ (i.e., the subject-specific atlas) is further deformed T times to positions $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\}$ (corresponding to the T time points) — but this time using $\boldsymbol{x}_0$ as reference position:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_0) \propto \exp\left[-K_1 \sum_d U_d^{\mathcal{K}}(\boldsymbol{x}_t, \boldsymbol{x}_0)\right], \quad (2)$$

for $t = 1,\ldots,T$. Note that the deformed mesh positions $\{\boldsymbol{x}_t\}$ are conditionally independent given the subject-specific atlas $\boldsymbol{x}_0$, which is the variable that creates the statistical dependence between the time points. A consequence of this conditional independence is that no particular temporal trajectory (e.g., atrophy) is assumed. This choice increases the flexibility of the method, by enabling it to model trajectories that involve changes in trend over time (e.g., crossover studies or cyclic patterns).

iv) Using the deformed mesh positions, label probabilities at each time point and voxel are computed by interpolating the values at the vertices of the tetrahedron enclosing the voxel. Let $\boldsymbol{r}_j$ be the 3D coordinates of voxel j, and let ϕ_{tn}^{K} be a deformed interpolation basis function linked to node n at time point t. The interpolated label probabilities at voxel j of time point t are then given by[2]

$$p_j\left(l|\boldsymbol{x}_t\right) = \sum_{n=1}^{N} \alpha_{nl}\phi_{tn}^{K}\left(\boldsymbol{r}_j;\boldsymbol{x}_t\right).$$

Segmentation images $\{\boldsymbol{l}_1,\ldots,\boldsymbol{l}_T\}$ are then created by independently sampling these categorical distributions at each voxel:

$$p\left(\boldsymbol{l}_t|\boldsymbol{x}_t\right) = \prod_{j=1}^{J} p_j\left(l_{tj}|\boldsymbol{x}_t\right)$$

where l_{tj} is the label of voxel j in time point t.

v) The intensities of the voxels are generated following three assumptions. First, that they are conditionally independent, given the segmentations. Second, that they follow a Gaussian distribution for each label and time point. And third, that labels describing structures of the same tissue type share their Gaussian parameters (means and variances) through G global classes. For example, gray matter structures such as the amygdala, the cerebral cortex, and many of the hippocampal subregions will belong to the same global class (see details in Section "Implementation Details"). Under these assumptions, the probability of observing the image at time point t is

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{l}_t,\boldsymbol{\theta}_t) &= \prod_{j=1}^{J} p(y_{tj}|l_{tj},\boldsymbol{\theta}_t) \\ &= \prod_{j=1}^{J} \mathcal{N}(y_{tj};\mu_{t\mathcal{G}(l_{tj})},\sigma^2_{t\mathcal{G}(l_{tj})}), \end{aligned}$$

where $\mathcal{N}$ is the Gaussian distribution, $\mathcal{G}(l) \in \{1,\ldots,G\}$ is the global class corresponding to label l, (μ_{tg},σ^2_{tg}) are the Gaussian parameters for time point t and global class g, and $\boldsymbol{\theta}_t = \{\{\mu_{tg}\},\{\sigma^2_{tg}\}\}$ represents all Gaussian parameters for time point t. Note that we allow the Gaussian parameters to be different for each time point, which removes the need to standardize the intensities across time points, and also models possible changes in contrast induced by disease. The parameters of each Gaussian (μ_{tg},σ^2_{tg}) are assumed to be independent samples of normal-inverse gamma (NIG) distributions, which is the conjugate prior for a Gaussian distribution with unknown mean and variance:

$$\begin{aligned} p(\boldsymbol{\theta}_t) &= \prod_{g=1}^{G} p\left(\mu_{tg},\sigma^2_{tg}\right) \\ p\left(\mu_{tg},\sigma^2_{tg}\right) &= NIG\left(\mu^0_{tg},\nu_{tg},0,0\right) \\ &= \mathcal{N}\left(\mu_{tg};\mu^0_{tg},\sigma^2_{tg}/\nu_{tg}\right), \end{aligned}$$

where we have assumed that the variance-related parameters of the NIG are equal to zero (i.e., the prior on the variance is a uniform distribution), and the remaining hyperparameters μ^0_{tg} and ν_{tg} encode any prior knowledge that we might have on the image intensities of each time point: μ^0_{tg} represents the expected mean of class g at time point t, which is assumed to have been obtained as the sample mean of ν_{tg} prior observations. Details on how these hyperparameters are computed are given in Section "Implementation Details" and Table 1.

Table 1
Global tissue classes grouping structures with similar image intensity properties. GC-DG stands for granule cell layer of the dentate gyrus, and HATA for hippocampus-amygdala transition area.

Global class	Structures
Gray matter	Cerebral cortex, amygdala, parasubiculum, presubiculum, subiculum, CA1, CA2/3, CA4, GC-DG, HATA
White matter	Cerebral white matter, fimbria
Cerebrospinal fluid	Ventricle, hippocampal fissure
Dicencephalon	Diencephalon
Thalamus	Thalamus
Pallidum	Pallidum
Putamen	Putamen
Choroid plexus	Choroid plexus

Segmentation as Bayesian inference

Given the model described above, segmentation can be cast as a Bayesian inference problem:

$$\left\{\hat{\boldsymbol{l}}_t\right\} = \arg\max_{\{\boldsymbol{l}_t\}} \; p\left(\{\boldsymbol{l}_t\}|\{\boldsymbol{y}_t\}\right).$$

Solving this problem exactly leads to an intractable integral over the model parameters, so we make the standard approximation that the posterior distribution of the parameters is heavily peaked. If we group all Gaussian parameters in $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_T\}$, and all deformations (subject-specific atlas and time points) in $\boldsymbol{x} = \{\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\}$, we have

$$\begin{aligned} \{\hat{\boldsymbol{l}}_t\} &= \arg\max_{\{\boldsymbol{l}_t\}} \int_{\mathbf{x}}\int_{\boldsymbol{\theta}} p\left(\{\boldsymbol{l}_t\}|\boldsymbol{x},\boldsymbol{\theta},\{\boldsymbol{y}_t\}\right) p\left(\boldsymbol{x},\boldsymbol{\theta}|\{\boldsymbol{y}_t\}\right) d\boldsymbol{x}d\boldsymbol{\theta} \\ &\approx \arg\max_{\{\boldsymbol{l}_t\}} p\left(\{\boldsymbol{l}_t\}|\hat{\boldsymbol{x}},\hat{\boldsymbol{\theta}},\{\boldsymbol{y}_t\}\right), \end{aligned}$$

where the point estimates of the model parameters are given by

$$\left\{\hat{\boldsymbol{x}},\hat{\boldsymbol{\theta}}\right\} = \arg\max_{\boldsymbol{x},\boldsymbol{\theta}} \; p\left(\boldsymbol{x},\boldsymbol{\theta}|\{\boldsymbol{y}_t\}\right).$$

[2] Linear barycentric interpolation leads to simpler solutions and provides satisfactory results in our case, but more complex models could be used, e.g. Pohl et al. (2007) and Ashburner and Friston (2009).

Using Bayes' rule, we can rewrite this problem as

$$\{\hat{\boldsymbol{x}},\hat{\boldsymbol{\theta}}\} = \arg\max_{\boldsymbol{x},\boldsymbol{\theta}} p\left(\{\boldsymbol{y}_t\}|\boldsymbol{x},\boldsymbol{\theta}\right) p\left(\boldsymbol{x}\right) p\left(\boldsymbol{\theta}\right)$$
$$= \arg\max_{\boldsymbol{x},\boldsymbol{\theta}} p\left(\boldsymbol{x}_0\right) \prod_{t=1}^{T} p\left(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}_t\right) p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) p\left(\boldsymbol{\theta}_t\right).$$

Finally, taking the logarithm of this expression, and expanding

$$p\left(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\theta}_t\right) = \sum_{\boldsymbol{l}_t} p\left(\boldsymbol{y}_t|\boldsymbol{l}_t,\boldsymbol{\theta}_t\right) p\left(\boldsymbol{l}_t|\boldsymbol{x}_t\right)$$
$$= \sum_{\boldsymbol{l}_t} \prod_{j=1}^{J} p\left(y_{tj}|l_{tj},\boldsymbol{\theta}_t\right) \prod_{j=1}^{J} p_j\left(l_{tj}|\boldsymbol{x}_t\right)$$
$$= \prod_{j=1}^{J} \sum_{l=1}^{L} p\left(y_{tj}|l,\boldsymbol{\theta}_t\right) p_j\left(l|\boldsymbol{x}_t\right),$$

we obtain the following objective function of the variables $\boldsymbol{x}_0$, $\{\boldsymbol{x}_t\}$, and $\{\boldsymbol{\theta}_t\}$:

$$\log p(\boldsymbol{x}_0) + \sum_{t=1}^{T} \log p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) + \sum_{t=1}^{T} \log p\left(\boldsymbol{\theta}_t\right) + \sum_{t=1}^{T}\sum_{j=1}^{J} \log\left[\sum_{l=1}^{L} p\left(y_{tj}|l,\boldsymbol{\theta}_t\right) p_j\left(l|\boldsymbol{x}_t\right)\right]. \quad (3)$$

The optimization of this objective function solves a joint registration, segmentation and subject-specific atlas estimation problem. We use a coordinate ascent scheme, in which one variable is updated at a time in an iterative fashion. In the rest of this section, we first describe the optimization procedure for each of the variables; then, we describe how the final segmentation is obtained once the point estimates have been computed; next, we provide details on our implementation; and finally, we close the section with a description of our strategy to avoid biases in the longitudinal analysis.

Optimization of $\boldsymbol{x}_t$, $t > 0$

The deformations of the individual time points can be updated independently of each other. Dropping any terms that are independent of $\boldsymbol{x}_t$ in Eq. (3), the problem reduces to

$$\arg\max_{\boldsymbol{x}_t} \log\ p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right) + \sum_{j=1}^{J} \log\left[\sum_{l=1}^{L} p\left(y_{tj}|l,\boldsymbol{\theta}_t\right) p_j\left(l|\boldsymbol{x}_t\right)\right]. \quad (4)$$

This is a registration problem, which includes a regularization term (the first) and a data term (the second). As in Iglesias et al. (2015), we solve this problem directly with a conjugate gradient optimizer. The problem is actually identical to that of Iglesias et al. (2015), with the only difference that the node positions of the population-wide atlas $\boldsymbol{x}_{ref}$ are replaced by those of the subject-specific atlas $\boldsymbol{x}_0$.

Optimization of $\boldsymbol{\theta}_t$

As with $\boldsymbol{x}_t$, the Gaussian parameters can be updated one time point at a time. The problem of Eq. (3) becomes

$$\arg\max_{\boldsymbol{\theta}_t} \log\ p(\boldsymbol{\theta}_t) + \sum_{j=1}^{J} \log\left[\sum_{l=1}^{L} p\left(y_{tj}|l,\boldsymbol{\theta}_t\right) p_j\left(l|\boldsymbol{x}_t\right)\right], \quad (5)$$

which can be solved with an Expectation-Maximization (EM) algorithm Dempster et al. (1977). The method iterates between an expectation (E) and a maximization (M) step until convergence. In the E step, a lower bound of the objective function in Eq. (5) that touches it at the current estimate of $\boldsymbol{\theta}_t$ is built, which involves computing a soft classification of each voxel in the image corresponding to the time point t:

$$W_{tjl} = \frac{p\left(y_{tj}|l,\boldsymbol{\theta}_t\right) p_j\left(l|\boldsymbol{x}_t\right)}{\sum_{l'=1}^{L} p\left(y_{tj}|l',\boldsymbol{\theta}_t\right) p_j\left(l'|\boldsymbol{x}_t\right)}. \quad (6)$$

In the subsequent M step, this bound is optimized with respect to $\boldsymbol{\theta}_t$, thereby guaranteeing to improve the original objective function of Eq. (5) compared to the previous iteration (Dempster et al., 1977). Taking derivatives and setting them to zero, we obtain the following update equations:

$$\mu_{tg} \leftarrow \frac{\nu_{tg}\mu_{tg}^0 + \sum_{j=1}^{J} \Omega_{tjg} y_{tj}}{\nu_{tg} + \sum_{j=1}^{J} \Omega_{tjg}}, \quad (7)$$

$$\sigma_{tg}^2 \leftarrow \frac{\nu_{tg}\left(\mu_{tg} - \mu_{tg}^0\right)^2 + \sum_{j=1}^{J} \Omega_{tjg}\left(y_{tj} - \mu_{tg}\right)^2}{\sum_{j=1}^{J} \Omega_{tjg}}, \quad (8)$$

where we have defined $\Omega_{tjg} = \sum_{\mathcal{G}(l)=g} W_{tjl}$.

Optimization of $\boldsymbol{x}_0$

Considering only terms depending on $\boldsymbol{x}_0$, Eq. (3) becomes

$$\arg\max_{\boldsymbol{x}_0} \left[\log p(\boldsymbol{x}_0) + \sum_{t=1}^{T} \log p\left(\boldsymbol{x}_t|\boldsymbol{x}_0\right)\right],$$

which is independent of the image intensities. Since the function $U_d^{\mathcal{K}}(\mathbf{x}_a,\mathbf{x}_b)$ in Eqs. (1) and (2) is symmetric (Ashburner et al., 2000), we can rewrite

$$\arg\min_{\boldsymbol{x}_0} \sum_{d} \left[K_0 U_d^{\mathcal{K}}\left(\boldsymbol{x}_0,\boldsymbol{x}_{ref}\right) + K_1 \sum_{t=1}^{T} U_d^{\mathcal{K}}\left(\boldsymbol{x}_0,\boldsymbol{x}_t\right)\right]. \quad (9)$$

Eq. (9) can be seen as a weighted "average" of the mesh positions of the time points and that of the population-wide atlas $\boldsymbol{x}_{ref}$. The atlas essentially plays the role of an additional time point, though with a different weight (K_0, rather than K_1). We solve this problem numerically with a conjugate gradient algorithm.

Computation of final segmentation

Once the point estimates of the model parameters have been computed, the conditional posterior label probabilities for each voxel are given by the soft classifications provided by the E step of the EM algorithm used to update the Gaussian parameters, i.e., Eq. (6):

$$p\left(\{\boldsymbol{l}_t\}|\hat{\boldsymbol{x}},\hat{\boldsymbol{\theta}},\{\boldsymbol{y}_t\}\right) = \prod_{t=1}^{T} p\left(\boldsymbol{l}_t|\hat{\boldsymbol{x}}_t,\hat{\boldsymbol{\theta}}_t,\boldsymbol{y}_t\right)$$
$$= \prod_{t=1}^{T}\prod_{j=1}^{J} W_{tjl_{tj}}. \quad (10)$$

If we desire to compute discrete segmentations, the MAP (maximum-a-posteriori) estimate can be computed voxel by voxel as

$$\hat{l}_{tj} = \arg\max_{l} W_{tjl},$$

whereas if we are interested in the volumes of the structures, their expected value can be shown to be equal to

$$V_{tl} = \sum_{j=1}^{J} W_{tjl},$$

where V_{tl} is the volume of the structure with label l in the image acquired at time point t.

Implementation details

Given a set of longitudinal scans, we first preprocess the data using the FreeSurfer (Dale et al., 1999; Fischl, 2012; Fischl et al., 2002, 1999) longitudinal stream (Reuter and Fischl, 2011; Reuter et al., 2012). The longitudinal stream creates an unbiased within-subject template space and image (base) (Reuter et al., 2012) using an inverse consistent registration method (Reuter et al., 2010). This template is a robust representation of the average subject anatomy and is processed with a modified FreeSurfer pipeline. The original time point images are conformed and resampled to the template space via a single cubic b-spline interpolation step. Several processing steps of the FreeSurfer pipeline are then initialized for each time point with common information from the subject template to increase reliability and thus statistical power. The normalized, bias-field corrected, skull-stripped images (*norm.mgz*) corresponding to the different time points are then used as input for the proposed longitudinal segmentation algorithm (i.e., $\{\boldsymbol{y}_t\}$).

To initialize the mesh positions, we first use an affine registration procedure to align the modeled image region with the cuboid in which the population-wide atlas is defined. As reference image, the registration uses a binary hippocampal mask extracted from the automated segmentation (FreeSurfer's "*aseg.mgz*") of the subject template. As moving image, the registration uses a soft segmentation of the hippocampus estimated from $\boldsymbol{x}_{ref}$. After the affine registration, we further deform the mesh non-linearly with Eq. (1) to the same automated segmentation of the subject template. This mesh deformation is used to initialize the node positions of subject-specific atlas $\boldsymbol{x}_0$, as well as the deformations of the time points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$.

The hyperparameters of the different time points and global tissue classes are computed from the corresponding *norm* and *aseg* images as follows: for each global class g, we extract the intensities of the voxels of *norm* labeled as any of the compatible labels by *aseg* (i.e., $l, s.t. \mathcal{G}(l) = g$). We set μ^0_{tg} to the median value of such intensities, and ν_{tg} to a conservative value equal to one half of the number of such voxels. The complete mapping of labels to global tissue classes is detailed in Table 1. Note that voxels from *outside* the hippocampus to estimate the intensity properties of the hippocampal subregions, which makes the algorithm more robust. For example: since they both consist of white matter, the intensity distribution of the fimbria can be more easily estimated from the cerebral white matter, which is much bigger and easier to segment.

We set the stiffness parameters to $K_0 = K_1 = 0.05$, which is the default value for the cross-sectional method currently implemented in FreeSurfer (Iglesias et al., 2015). We rasterize (i.e., interpolate) the mesh at 0.333 mm isotropic resolution, which is also the default value in the current FreeSurfer implementation. This resolution represents the voxel size at which the final segmentations are obtained.

For the optimization, we use the following scheme: we first alternately update $\{\boldsymbol{\theta}_t\}$ and $\{\boldsymbol{x}_t\}$ 10 times. Each update of $\boldsymbol{\theta}_t$ iterates between the E and M steps until the change in the objective function is less than 10^{-5}, whereas each update of $\boldsymbol{x}_t$ takes at most 20 iterations of the conjugate gradient method (it stops early if the maximum shift across mesh nodes is less than 10^{-5}). Next, $\boldsymbol{x}_0$ is updated with the conjugate gradient algorithm (maximum 100 steps; the early termination criterion is the same as for $\boldsymbol{x}_t$). The optimization then returns to the update of $\{\boldsymbol{\theta}_t\}$, starting a new external iteration. We set the maximum number of external iterations to 10. The complete segmentation algorithm is summarized in Algorithm 1.

Algorithm 1. Longitudinal segmentation.

Compute $\mu^0_{tg}, \nu_{tg}, \boldsymbol{x}_0$ with *norm.mgz*, *aseg.mgz*
$\mu_{tg} \Leftarrow \mu^0_{tg}, \forall t, g$; $\sigma^2_{tg} \Leftarrow 100, \forall t, g$; $\boldsymbol{x}_t \Leftarrow \boldsymbol{x}_0, \forall t > 0$
for $its = 1$ to 10 **do**
  for $t = 1$ to T **do**
    $LogPprev \Leftarrow -\infty$; $LogPcurr \Leftarrow 0$
    while $LogPcurr - LogPprev > 10^{-5}$ **do**
      $LogPprev \Leftarrow LogPcurr$
      $LogPcurr \Leftarrow$ Eq. 5
      $W_{tjl} \Leftarrow$ Eq. 6; $\mu_{tg} \Leftarrow$ Eq. 7; $\sigma^2_{tg} \Leftarrow$ Eq. 8
    end while
    if $its < 10$ **then**
      $itDef \Leftarrow 0$; $maxDef \Leftarrow \infty$
      while $itDef < 20$ **and** $maxDef > 10^{-5}$ **do**
        $itDef \Leftarrow itDef + 1$
        $(\boldsymbol{x}_t, maxDef) \Leftarrow$ conjugate gradient on Eq. 4
      end while
    end if
  end for
  if $its < 10$ **then**
    $itDef \Leftarrow 0$; $maxDef \Leftarrow \infty$
    while $itDef < 20$ **and** $maxDef > 10^{-5}$ **do**
      $itDef \Leftarrow itDef + 1$
      $(\boldsymbol{x}_0, maxDef) \Leftarrow$ conjugate gradient on Eq. 9
    end while
  end if
end for
$\hat{l}_{tj} \Leftarrow \mathrm{argmax}_l W_{tjl}, \forall t, j$
$V_{tl} \Leftarrow \sum_{j=1}^{J} W_{tjl}, \forall t, l$

Avoiding biases

As mentioned in the introduction, processing bias can be introduced if all the time points are not treated in exactly the same way. In our algorithm, the initialization is computed with the output from the FreeSurfer longitudinal pipeline, which is designed to avoid processing bias (Reuter et al., 2010, 2012). The segmentation algorithm is also unbiased, since all images are treated identically. Moreover, subjects with a single time point are treated as if they were longitudinal, which makes the measures derived from them comparable with those obtained from subjects with multiple time points. More specifically, the FreeSurfer longitudinal pipeline includes a pose normalization step that introduces resampling artifacts and a subject template, and the hippocampal segmentation estimates the mesh position for a subject-specific atlas (rather than using the population-wide atlas directly). This procedure makes it possible to include all subjects in analyses that support single time point data, such as linear mixed effects models (Bernal-Rusiel et al., 2013).

Experiments and results

MRI data

We used two publicly available datasets in the experiments in this study: MIRIAD and ADNI. The MIRIAD dataset consists of T1-weighted brain MRI scans of AD patients ($n = 46$) and cognitively normal (CN) controls ($n = 23$) acquired at intervals from two weeks to two years. All 69 subjects were scanned at 0, 2, 6, 14, 26, 38 and 52 weeks from baseline; 39 subjects were also scanned at 18 months; 22 of these 39 were further scanned at 24 months. At 0, 6 and 38 weeks, two back-to-back scans were conducted without

removing the subject from the scanner in between. The mean age at baseline of the subjects was 69.6±6.9 years. All the scans were acquired on the same 1.5 T scanner (GE Signa) with an IR-FSPGR sequence (coronal slices with 0.9375×0.9375 mm resolution, 1.5 mm slice thickness, TR=15 ms, TE=5.4 ms, TI=650 ms, flip angle 15°). Further information can be found at https://www.ucl.ac.uk/drc/research/miriad-scan-database.

The ADNI dataset consists of longitudinal T1-weighted scans from 836 subjects of the ADNI dataset. The subjects are divided into four classes: elderly controls ($n = 252$), early mild cognitive impairment (eMCI, $n = 215$), late MCI (lMCI, $n = 176$), and AD ($n = 193$). The subjects were scanned on average 4.8 times (minimum: a single time; maximum: 11 times; 4013 scans in total), with a mean interval between scans equal to 286 days (minimum: 23 days, maximum: 1567 days). The mean age at baseline of the subjects was 75.1±6.6 years. Since the ADNI project spans multiple sites, different scanners were used to acquire the images; further details on the acquisition can be found at http://www.adni-info.org.

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The main goal of ADNI is to test whether MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to analyze the progression of MCI and early AD. Markers of early AD progression can aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as decrease the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California — San Francisco. ADNI is a joint effort by co-investigators from industry and academia. Subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. These three protocols have recruited over 1500 adults (ages 55–90) to participate in the study, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the corresponding protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2.

Experimental setup

Competing methods

We compared the performance of our algorithm with that of two other approaches. The competing methods were:

1. *Cross-sectional segmentation* (henceforth "C-S"): the algorithm described in Iglesias et al. (2015) was used to segment each time point independently of the others in a cross-sectional fashion (i.e., as if they were different subjects).
2. *Cross-sectional segmentation with longitudinal initialization* (henceforth "L-INIT"): same as C-S, but initializing the algorithm with the automated segmentation (*aseg*) from the longitudinal FreeSurfer stream (rather than the cross-sectional *aseg*).
3. *Longitudinal segmentation* (henceforth "LONG"): the algorithm described in this paper was used to segment all the time points corresponding to each subject simultaneously.

The motivation for testing L-INIT is twofold. First, it is currently the recommended setup for longitudinal hippocampal subfield segmentation in FreeSurfer. And second, it enables us to isolate the contribution of our proposed generative model to the results of LONG, separating it from the contribution of the longitudinal initialization.

In order to assess the segmentation accuracy of the methods, we would ideally use ground truth labels obtained from manual delineations of the hippocampal substructures made on the *in vivo* MRI scans. However, such manual annotations would have to be made with the protocol that we used to build the ultra-high resolution *ex vivo*, which is not possible. Instead, we validated the method indirectly through two sets of experiments: test–retest reliability, and group differentiation with linear mixed effect (LME) modeling.

Experiment 1: test–retest reliability

In order to evaluate the test–retest reliability of the methods, we used them to segment the scan–rescan data of the MIRIAD dataset. For each subject, we took the scan–rescan session corresponding to the first time point (therefore including both AD subjects and controls). After segmenting each of the $n = 69$ pairs of scans with the three competing algorithms, we compared their performance with two different metrics. First, we measured the absolute difference in volume estimates for each of the segmented hippocampal subregions. The smaller this difference, the larger the agreement across the two scans. Second, we computed the Dice overlap between the MAP segmentations of each subregion in the two scans. The Dice coefficient between two binary masks X and Y is defined as

$$Dice(X,Y) = 2\frac{|X \cap Y|}{|X| + |Y|},$$

and is bounded by 0 (no overlap) and 1 (perfect overlap). When the C-S method is used, computing the Dice overlap requires a rigid registration between the two scans, which was computed with the robust registration tool in FreeSurfer (Reuter et al., 2010). In order to mitigate the effect of image resampling on the Dice overlaps in this scenario, we used linear resampling to warp the scans to the intermediate space (the base) and replaced the Dice coefficient by a soft counterpart:

$$Dice_s\,(X_s, Y_s) = 2\frac{\sum_{\boldsymbol{r}} X_s(\boldsymbol{r}) Y_s(\boldsymbol{r})}{\sum_{\boldsymbol{r}} X_s(\boldsymbol{r}) + \sum_{\boldsymbol{r}} Y_s(\boldsymbol{r})},$$

where $\boldsymbol{r}$ represents spatial locations, and $X_s(\boldsymbol{r})$, $Y_s(\boldsymbol{r})$ are resampled masks defined between zero and one[3] .

Experiment 2: group analysis with LME

The test–retest experiment described above only evaluated one aspect of the longitudinal algorithms: their ability to produce consistent segmentations. Additionally, it is necessary to test the performance when capturing the temporal evolution of the segmented structures. For example, an algorithm that always produces the same output yields perfect test–retest reliability, but also fails to capture any anatomical changes over time or differentiate groups based on such changes.

We carried out two experiments using group analyses: one with MIRIAD, and one with ADNI. The setup was identical in both cases, with the only difference that the datasets have different numbers of classes. For each hippocampal subregion, we built an LME model for the estimated volume in which subject intercept and slope were random effects, intracranial volume (ICV) and age at baseline were fixed effects, and each group had its own (fixed) bias and slope. The model fit and computation of p values for F tests comparing the fixed slopes of the different groups was done with the LME toolbox in FreeSurfer (Bernal-Rusiel et al., 2013). We then took the ability of the measurements to separate the (fixed) slopes of the groups as a measure of

[3] Despite using soft Dice, some bias against the C-S method is still introduced; this is further discussed in Section "Discussion".

the sensitivity of the longitudinal segmentation to detect anatomical change associated with disease

For the ADNI dataset, we chose to merge the late MCI and AD classes into a single class (lMCI/AD). This choice was motivated by the fact that a pilot LME analysis using whole hippocampal volumes from FreeSurfer's longitudinal stream did not reveal any differences in atrophy rates between the two classes. This is consistent with the results of other studies based on manual (Jack et al., 2000, 2004) and automated segmentations (Risacher et al., 2009). This lack of differences between the late MCI and AD groups may be explained by the continuous nature of pathology; current *in vivo* imaging technology cannot identify the subtle differences in atrophy rates between the two groups. It is necessary to examine the patient serially to be sure of the clinical findings, and 10–20% of patients with MCI will worsen and convert to AD (in fact, many lMCI subjects are diagnosed as AD at other time points in ADNI). In addition, the presence of comorbidities and other dementia etiologies (e.g., vascular dementia or dementia of the Lewy body disease, Schneider et al., 2009) makes it difficult to decipher the stage of the pathology at this point with *in vivo* imaging.

Results

Test–retest

Fig. 3 displays the absolute differences (in %) between the volumes of the hippocampal subregions estimated from the scan–rescan data of the MIRIAD dataset. The average differences across structures are as follows: 6.5% for C-S, 5.9% for L-INIT, and 4.5% for LONG. L-INIT provides a slight improvement over the purely cross-sectional (C-S) method, thanks to the implicit regularization introduced by the use of the FreeSurfer longitudinal stream in the initialization. Despite being quite consistent across subregions, this improvement is only significant (as measured with a two-tailed paired t-test) for one of them: the left granule cell layer of the dentate gyrus (DG). The proposed longitudinal method (LONG), which explicitly regularizes the segmentations, produces the lowest difference for all structures except for the right fimbria. The improvements over the C-S method are statistically significant for all structures except for the presubiculum and fimbria (both sides); left molecular layer; and left whole hippocampus. In absolute terms, the errors are below 5% for all structures except for the parasubiculum, hippocampus-amygdala transition area (HATA) and fimbria. These three subregions suffer from the highest variability in volume estimates: the parasubiculum because it represents the transition of the hippocampus with the entorhinal cortex, and its boundaries are not well defined; the HATA because it is a transitional region with the head of the hippocampus (dorsal subiculum) and amygdala; and the fimbria due to its occasional low contrast.

Fig. 4 displays the Dice coefficient for the different hippocampal subregions and competing methods. The averages across subregions are 0.754 for C-S, 0.775 for L-INIT, and 0.818 for LONG. L-INIT outperforms C-S for nearly all structures, in a statistically significant manner in most cases (once more, significance was assessed with a two-tailed paired t-test). LONG provides the highest Dice for all subregions except for the left tail, right tail and right fimbria. Moreover, it yields a statistically significant increase with respect to the other

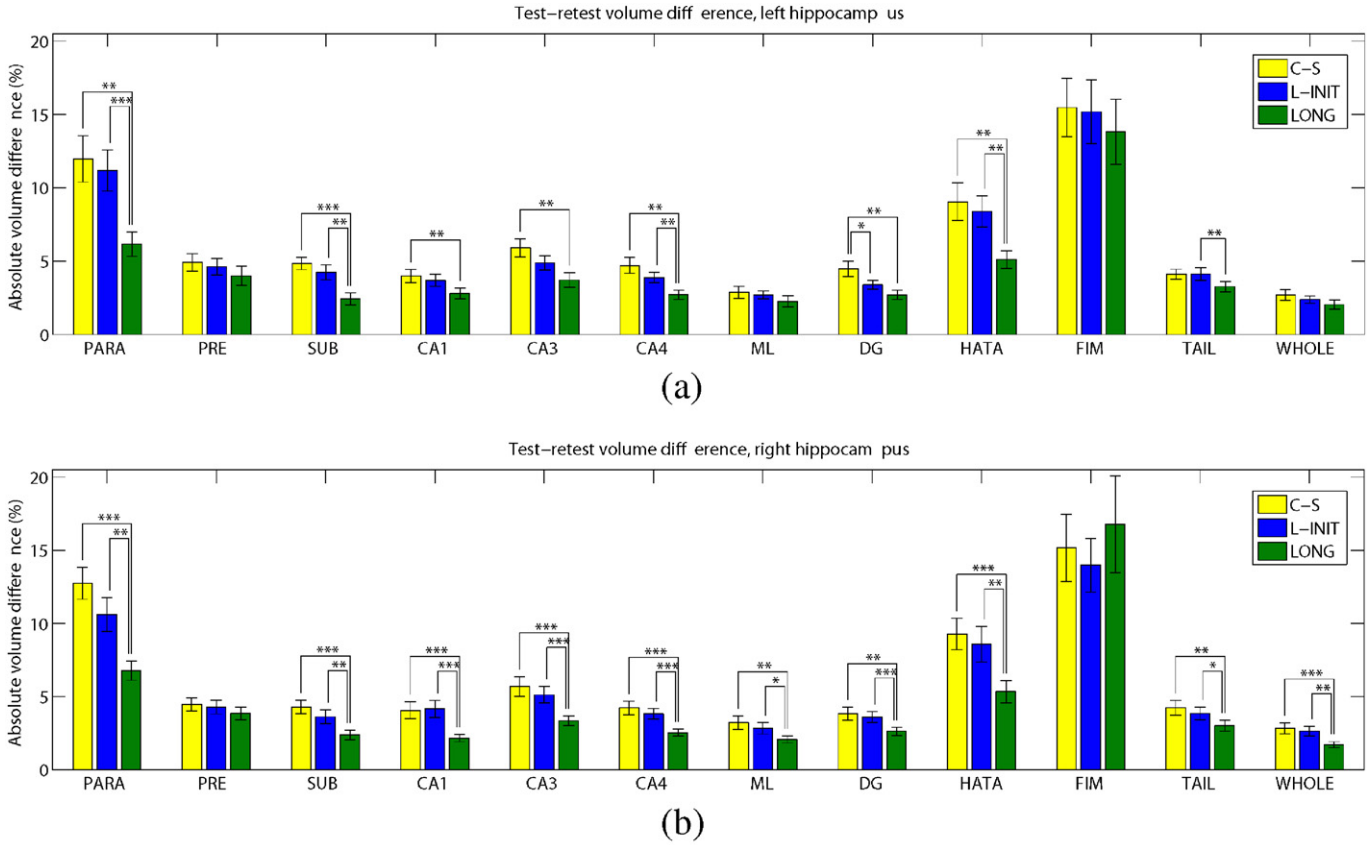

Fig. 3. Absolute volume differences (in % of total volume) for the hippocampal subregions in the back-to-back scans of the MIRIAD dataset: (a) left hippocampus; and (b) right hippocampus. The bars represent the mean, and the error bars, one standard deviation. A two-tailed paired t-test was used to assess whether there were significant differences between the methods: one asterisk represents $p < 0.05$, two asterisks represent $p < 0.01$, and three asterisks represent $p < 0.001$. The abbreviations of the hippocampal subregions are as follows: SUB = subiculum, PRE = presubiculum, PARA = parasubiculum, ML = molecular layer, DG = granule cell layer of the dentate gyrus, CA3 = CA2+CA3, FIM = fimbria, HATA = hippocampus-amygdala transition area, and WHOLE = whole hippocampus. For anatomical and morphological definitions of these subregions, see Iglesias et al. (2015).

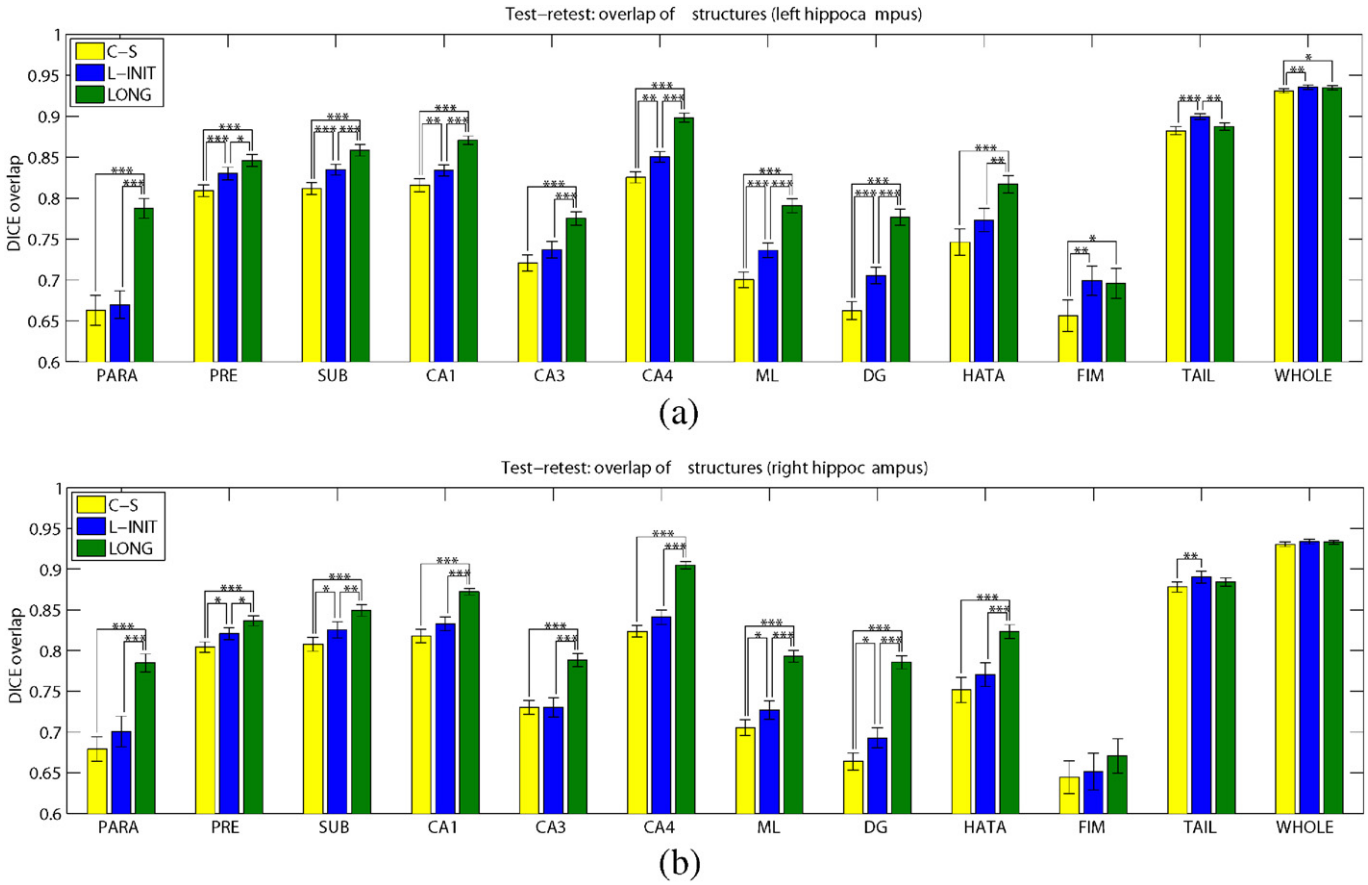

Fig. 4. Dice overlaps for the different subregions in the back-to-back scans of the MIRIAD dataset: (a) left hippocampus; and (b) right hippocampus. Please see the caption of Fig. 3 for the abbreviations of the hippocampal subregions and the convention for representation of statistical significance.

two methods in all hippocampal subregions except for the tail and fimbria. It is worth noting that the Dice scores for C-S are negatively affected (to a very small extent) by the resampling that is required to compute them.

Fig. 5 shows a coronal slice of a test–retest scan illustrating the differences between the algorithms. In this sample subject, C-S undersegments the superior region of the hippocampus (pointed red arrow) only in the first scan, creating a large difference with the second scan. While this issue is fixed by L-INIT, some undersegmentation still occurs in the subicular region of the first scan (blue arrow), and some inconsistencies are observed in the presubiculum and molecular layer (green arrow). The proposed longitudinal framework (LONG), on the other hand, produces segmentations that are more consistent with each other.

Group analysis

Figs. 6 and 7 show the atrophy rates for the MIRIAD dataset (computed for each group as the fixed slope divided by the fixed intercept) as estimated by the three competing methods. The cross sectional method (C-S) can detect the differences in some of the subregions and in the whole hippocampal volume, particularly in the right hemisphere (which is known to atrophy at a faster rate, Thompson et al., 2004). When L-INIT is used, effects that the C-S method could not detect are now found: moderate effects on the right tail and subiculum, and mild effects on the left dentate gyrus and CA4, though a strong effect is lost for the left subiculum. Our new algorithm (LONG) improves group differentiation even further: in addition to all the effects that the other two approaches could detect combined, it also finds a strong effect on the left presubiculum, a moderate effect on the right presubiculum, and mild effects on the left HATA and right parasubiculum.

Figs. 8 and 9 show the atrophy rates for the ADNI dataset. When comparing the controls with the lMCI/AD group, strong effects are found by all three methods for almost every hippocampal subregion (except for the highly variable fimbria). However, when comparing controls with eMCI and lMCI/AD with eMCI, the longitudinal methods reveal differences that the cross-sectional version could not find. Initializing with the longitudinal FreeSurfer segmentation (L-INIT) yields stronger signal for a number of subregions, such as the left CA3, left HATA, and right subiculum. The proposed longitudinal model (LONG) detects even more effects, such as slight differences in the left subiculum and presubiculum, and the right parasubiculum. LONG also detects stronger effects for many other subregions, such as the left DG, left CA4, or right CA1.

Discussion

The model we propose in this paper assumes that longitudinal scans of a certain individual have been generated by a hidden subject-specific atlas. This spatio-temporal approach allows a completely symmetric setup (all time points are treated identically), thus avoiding potential processing bias. The subject-specific atlas explicitly regularizes the segmentation across scans from different time points, which consistently increases the test–retest reliability while improving sensitivity. Perfect reliability can, of course, be enforced by reporting the same result across time independent of the image (over-constraining the method). However, this will prevent the detection of longitudinal changes and group differences. The

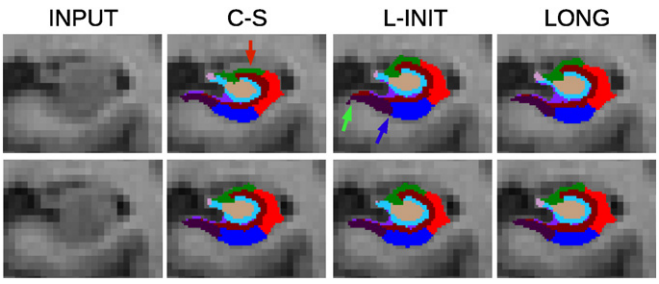

Fig. 5. Registered coronal slices of back-to-back scans of a sample subject of the MIRIAD dataset. From left to right: input, cross-sectional segmentation, segmentation with FreeSurfer longitudinal initialization, and proposed longitudinal method. The top row corresponds to the first scan, and the bottom row to the second scan.

presented approach aims at optimizing the trade-off between noise reduction and over-regularization by keeping the model flexible enough to follow temporal morphometric changes.

The proposed longitudinal segmentation method was evaluated against a purely cross-sectional implementation (C-S) and a variant of it (L-INIT) that uses the FreeSurfer longitudinal stream in the initialization. The test–retest experiments revealed that taking advantage of the longitudinal stream already enabled L-INIT to consistently outperform C-S in terms of volume error and Dice coefficient. The generative model takes the performance one step further, and enables our proposed method (LONG) to outperform L-INIT for both metrics and nearly every hippocampal subregion. It is worth noting that the Dice coefficients computed for C-S are negatively affected by the registration it requires. However, given that all other metrics (including the sensitivity to differences in atrophy rates) support the superiority of L-INIT and LONG, and given that we used a soft version of the Dice coefficient to reduce the impact of resampling, there is no reason to believe that the observed differences can be attributed exclusively to interpolation artifacts.

When comparing atrophy rates across disease groups, we observed a similar trend as in the test–retest experiments. L-INIT revealed effects that C-S could not detect, and we also demonstrated that the regularization scheme in LONG increases the ability to separate various groups in the two datasets (MIRIAD and ADNI) even further. This is essential as significance in group comparisons is affected both by the measurement noise and the effect size.

In absolute terms, the three competing methods yielded approximately the same annual rates of atrophy for the whole hippocampus in controls: 1% in MIRIAD, and 1.5% in ADNI. For early MCI (in ADNI), they all produced similar estimates as well (2%). In the AD group, however, the rates dropped from 3.75% to 3.35% in MIRIAD and from 4% to 3.6% in ADNI for the proposed method. This could indicate that the regularization scheme used by our method (i.e., the subject-specific atlas) might slightly oversmooth trajectories corresponding to larger atrophy rates (i.e., those corresponding to AD patients).

We also need to emphasize that higher atrophy rates do not necessarily correspond to more accurate segmentations. Ideally, one would evaluate such accuracy directly with the help of manual delineations, but this was not possible in this study because the 1 mm *in vivo* images cannot be manually annotated with our *ex vivo* delineation protocol. Nevertheless, the atrophy rates estimated by our method agree well with previously published data. In MIRIAD, our estimates are very similar to those reported by Cash et al. (2015), who surveyed the output from 13 automated methods, and reported 0.7% for controls and 3.8% for AD. In ADNI, our estimates for late MCI/AD are also very close to those reported by Jack et al. (2000) (3.5%) and Jack et al. (2004) (3.3%–3.6%) using manual segmentations, even though higher values have also been reported by other studies (e.g., Henneman et al. (2009) reported 4.0%). A more thorough analysis of hippocampal atrophy rates estimated with neuroimaging can be found in Barnes et al. (2009).

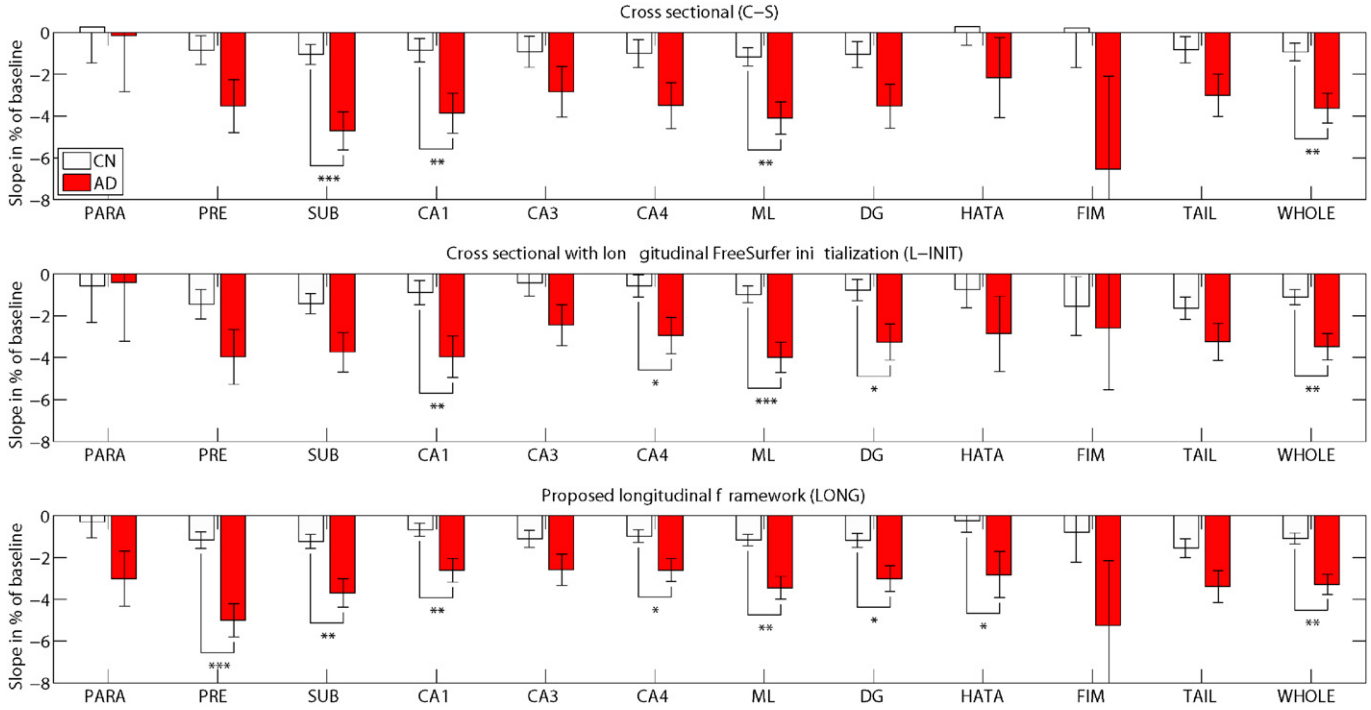

Fig. 6. MIRIAD dataset: atrophy rates (percentage of baseline, per year) for the hippocampal subregions of the left hemisphere as estimated by the three competing methods. The abbreviations for the subregions and the conventions for statistical significance can be found in Fig. 3.

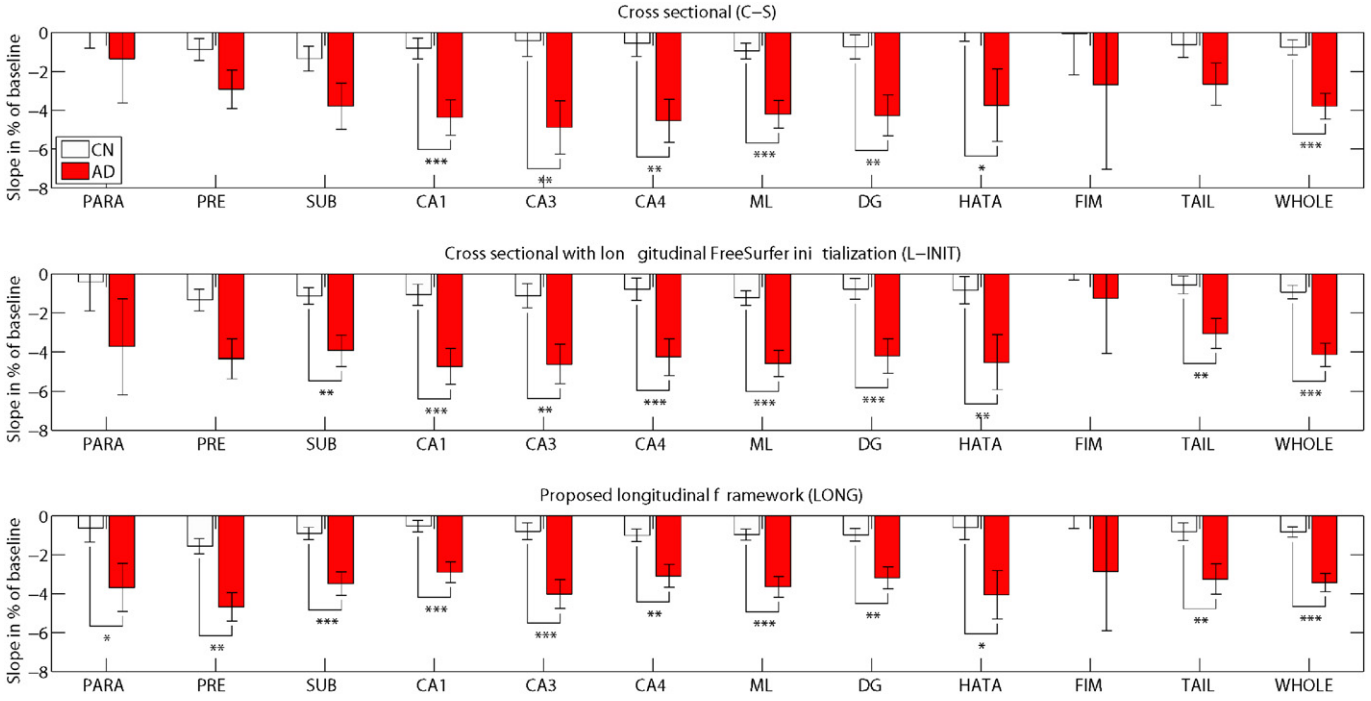

Fig. 7. MIRIAD dataset: atrophy rates for the hippocampal subregions of the right hemisphere. The abbreviations for the subregions and the conventions for statistical significance can be found in Fig. 3.

Conclusion

In this article, we have proposed a novel Bayesian longitudinal segmentation algorithm for hippocampal subregions based on a hidden subject-specific atlas. The method is general and could in principle be applied to other brain regions, though such a setup would require further evaluation in future work. Also, the method does not make any assumptions on the shape or temporal smoothness of the trajectories, i.e., it treats all time points the same way. This design increases the flexibility of the proposed segmentation method. Further information on ordering and time spacing, as well as further assumptions on the shape of the trajectories (e.g., linear) can be exploited by the statistical tools that are used to analyze the output of the segmentation. For example, in this study, we used a linear

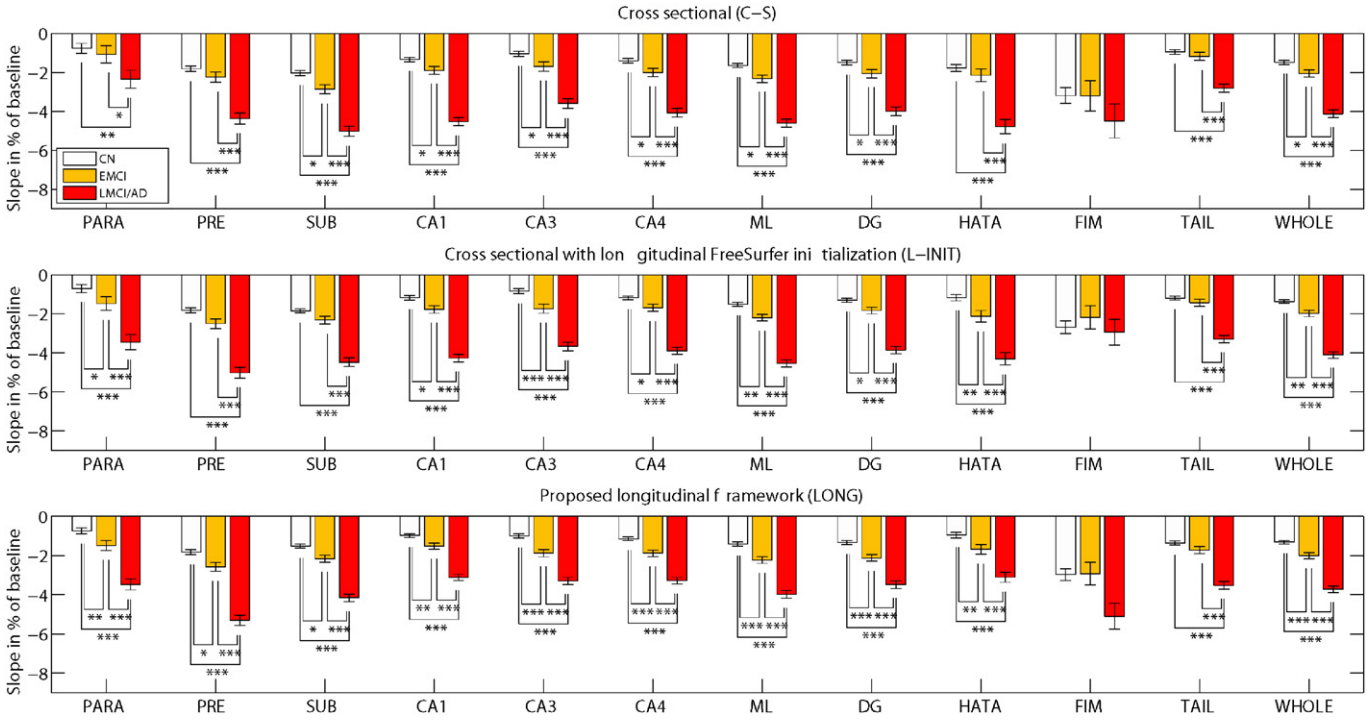

Fig. 8. ADNI dataset: atrophy rates (percentage of baseline, per year) for the hippocampal subregions of the left hemisphere as estimated by the three competing methods. The abbreviations for the subregions and the conventions for statistical significance can be found in Fig. 3.

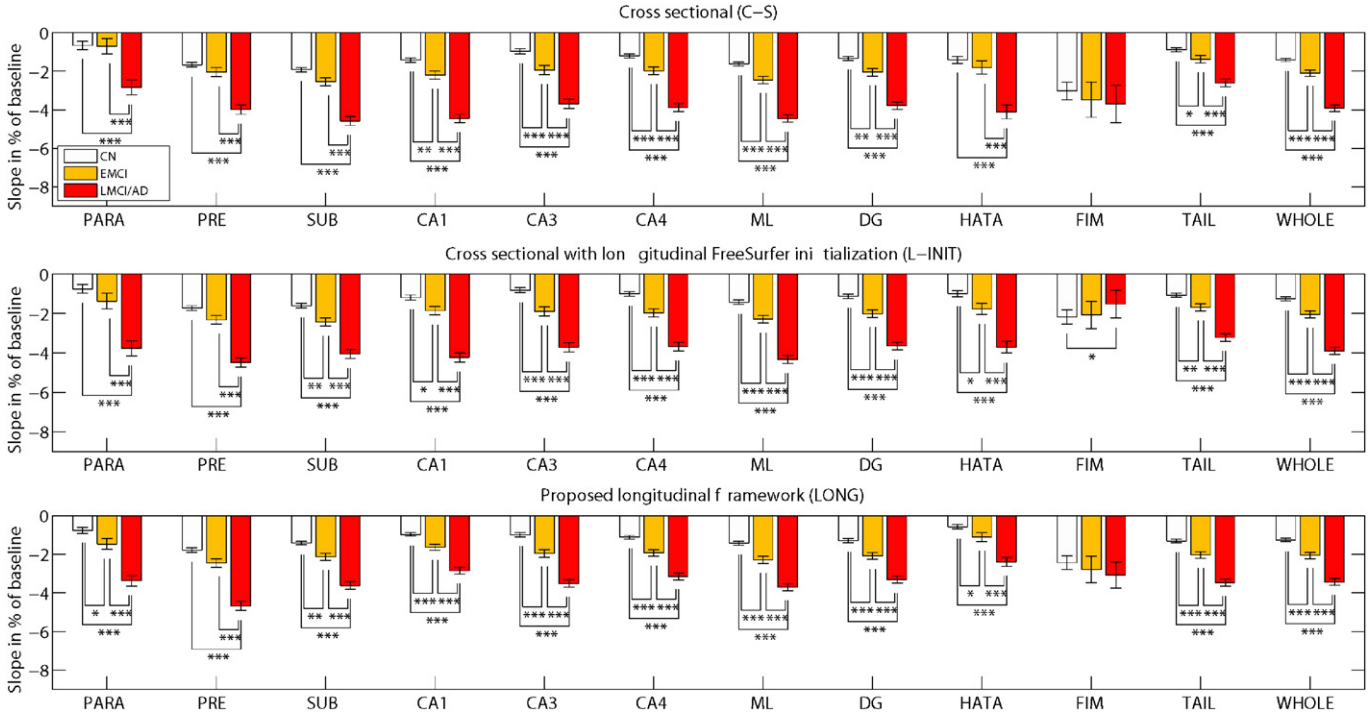

Fig. 9. ADNI dataset: atrophy rates (percentage of baseline, per year) for the hippocampal subregions of the right hemisphere. See caption of Fig. 8 for an explanation of this figure.

mixed effect model that accounted for the time spacing a correlations between repeated measures, while assuming linear trajectories (which approximately holds in atrophy studies).

Our approach builds on the literature of Bayesian segmentation with unsupervised intensity models, and inherits the robustness of such methods against changes in MRI contrast — which stems from the fact that intensity properties are inferred directly from the images to be segmented. This is actually a requirement if the atlas is constructed using *ex vivo* data (which enables ultra-high-resolution), since fixation and death radically change MRI contrast. Therefore, the algorithm does not require and intensity standardization across time points, and can handle changes in contrast induced by disease. That said, if the image intensities at all time points are know to be normalized and not affected by pathology, the robustness of the algorithm could be enhanced by forcing the Gaussian parameters to be equal across time points, i.e., $\boldsymbol{\theta}_t = \bar{\boldsymbol{\theta}}, \forall t$. However, the potential gain would be minimal because there are sufficient voxels in each time point to estimate $\boldsymbol{\theta}_t$ with high certainty (Iglesias et al., 2013).

Another advantage of Bayesian segmentation with probabilistic atlases that our algorithm also inherits is its computational efficiency. Our implementation runs in approximately $15T - -20T$ min on a modern desktop, where T is the number of time points[4] . The implementation will be publicly shared as part of the popular neuroimaging package FreeSurfer, and will be (to the best of our knowledge) the first available method to longitudinally segment the hippocampal subregions.

As in the original cross-sectional method (Iglesias et al., 2015), the volumetric results from individual subfields need to be interpreted with caution when segmenting 1 mm images; at that resolution, the molecular layer is not visible, and the fitting of the internal boundaries of the hippocampal atlas relies mostly on the prior. In that sense, the statistical dependence introduced by the subject-specific atlas helps increase the stability of the segmentation of such internal boundaries across time points. Nevertheless, we would only recommend complex analyses (e.g., shape analysis) of the segmentations if the proposed method is applied to longitudinal data acquired at a higher resolution (e.g., $0.4 \times 0.4 \times 2.0$ mm scans as in Iglesias et al., 2015).

As a growing number of studies are beginning to collect longitudinal MRI data, the development of dedicated algorithms that exploit the relationship between scans of the same subject is paramount. Longitudinal methods that provide higher sensitivity than their cross-sectional counterparts permit reduction of sample sizes in neuroimaging studies and the detection of much smaller effects. Moreover, longitudinal segmentation algorithms for the hippocampal subregions hold great promise to increase our understanding of AD progression and disease etiology; to provide powerful biomarkers for computer-aided diagnosis at presymptomatic stages; and to allow a highly accurate and localized quantification of treatment response in AD and other neurological disorders.

[4] This is in addition to the processing time required by the main FreeSurfer stream, which is demanding since it produces many other results (cortical thickness, parcellation, etc).

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No.654911 (project "THALAMODEL"), and also from the Spanish Ministry of Economy and Competitiveness (MINECO, reference TEC2014-51882-P). Support for this research was also provided in part by the National Cancer Institute (1K25-CA181632-01), the Genentech Foundation (G-40819) and the Nvidia corporation, which donated a Titan X GPU. Further support was provided by the A.A. Martinos Center for Biomedical Imaging (P41RR014075, P41EB015896, U24RR021382), and was made possible by the resources provided by Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10RR023043. Support was also provided by the National Institute for Biomedical Imaging and Bioengineering (R01EB006758, R21EB018907, R01EB019956), the National Institute on Aging (5R01AG008122,

R01AG016495), the National Institute for Neurological Disorders and Stroke (R01NS0525851, R21NS072652, R01NS070963, R01NS083534, 5U01NS086625) and the Lundbeck Foundation (R141-2013-13117), Additional support was provided by the NIH Blueprint for Neuroscience Research (5U01-MH093765), as part of the multi-institutional Human Connectome Project. In addition, BF has a financial interest in CorticoMetrics, a company whose medical pursuits focus on brain imaging and measurement technologies. BF's interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

The collection and sharing of the MRI data used in the group study based on ADNI was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (U.S. 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: 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. 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Apostolova, L.G., Dinov, I.D., Dutton, R.A., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M., 2006. 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. Brain 129 (11), 2867–2873.

Arnold, S.E., Hyman, B.T., Flory, J., Damasio, A.R., Van Hoesen, G.W., 1991. The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease. Cereb. Cortex 1 (1), 103–116.

Ashburner, J., Andersson, J.L., Friston, K.J., 2000. Image registration using a symmetric prior — in three dimensions. Hum. Brain Mapp. 9 (4), 212–225.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. Neuroimage 45 (2), 333–341.

Aubert-Broche, B., Fonov, V., Garcìa-Lorenzo, D., Mouiha, A., Guizard, N., Coupé, P., Eskildsen, S.F., Collins, D.L., 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, J., Bartlett, J.W., van de Pol, L.A., Loy, C.T., Scahill, R.I., Frost, C., Thompson, P., Fox, N.C., 2009. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. Neurobiol. Aging 30 (11), 1711–1723.

Bauer, S., Tessier, J., Krieter, O., Nolte, L.-P., Reyes, M., 2014. Integrated spatio-temporal segmentation of longitudinal brain tumor imaging studies. Medical Computer Vision. Large Data in Medical Imaging. Springer., pp. 74–83.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., 2013. ADNI, Statistical analysis of longitudinal neuroimage data with linear mixed effects models. Neuroimage 66, 249–260.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82 (4), 239–259.

Burggren, A.C., Zeineh, M., Ekstrom, A.D., Braskie, M.N., Thompson, P.M., Small, G.W., Bookheimer, S.Y., 2008. Reduced cortical thickness in hippocampal subregions among cognitively normal apolipoprotein E e4 carriers. Neuroimage 41 (4), 1177–1183.

Cash, D.M., Frost, C., Iheme, L.O., Ünay, D., Kandemir, M., Fripp, J., Salvado, O., Bourgeat, P., Reuter, M., Fischl, B., et al. 2015. Assessing atrophy measurement techniques in dementia: results from the MIRIAD atrophy challenge. NeuroImage 123, 149–164.

Christensen, A., Alpert, K., Rogalski, E., Cobia, D., Rao, J., Beg, M.F., Weintraub, S., Mesulam, M.-M., Wang, L., 2015. Hippocampal subfield surface deformity in non-semantic primary progressive aphasia. Alzheimers Dement.: Diagn. Assess. Dis. Monit. 1 (1), 14–23.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. Neuroimage 9 (2), 179–194.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 1–38.

Du, A., Schuff, N., Amend, D., Laakso, M., Hsu, Y., Jagust, W., Yaffe, K., Kramer, J., Reed, B., Norman, D., et al. 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 71 (4), 441–447.

Eldridge, L.L., Knowlton, B.J., Furmanski, C.S., Bookheimer, S.Y., Engel, S.A., 2000. Remembering episodes: a selective role for the hippocampus during retrieval. Nat. Neurosci. 3 (11), 1149–1152.

Fischl, B., 2012. Freesurfer. Neuroimage 62 (2), 774–781.

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

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. Neuroimage 9 (2), 195–207.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2012. Applied Longitudinal Analysis. vol. 998. John Wiley & Sons.

Gabrieli, J.D., Brewer, J.B., Desmond, J.E., Glover, G.H., 1997. Separate neural bases of two fundamental memory processes in the human medial temporal lobe. Science 276 (5310), 264–266.

Gao, Y., Prastawa, M., Styner, M., Piven, J., Gerig, G., 2014. A joint framework for 4D segmentation and estimation of smooth temporal appearance changes. Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on, IEEE. pp. 1291–1294.

Henneman, W., Sluimer, J., Barnes, J., Van Der Flier, W., Sluimer, I., Fox, N., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Hippocampal atrophy rates in Alzheimer disease added value over whole brain volume measures. Neurology 72 (11), 999–1007.

Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L., Fischl, B., Van Leemput, K., 2015. A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: application to adaptive segmentation of in vivo MRI. NeuroImage 115, 117–137.

Iglesias, J.E., Sabuncu, M.R., Van Leemput, K., Initiative, A.D.N., et al. 2013. Improved inference in Bayesian segmentation using monte carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778.

Jack, C., Petersen, R., Xu, Y., O'Brien, P., Smith, G., Ivnik, R., Boeve, B., Tangalos, E., Kokmen, E., 2000. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. Neurology 55 (4), 484–490.

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

Kesner, R.P., 2007. A behavioral analysis of dentate gyrus function. Prog. Brain Res. 163, 567–576.

Knierim, J.J., Lee, I., Hargreaves, E.L., 2006. Hippocampal place cells: parallel input streams, subregional processing, and implications for episodic memory. Hippocampus 16 (9), 755.

Laakso, M., Soininen, H., Partanen, K., Lehtovirta, M., Hallikainen, M., Hänninen, T., Helkala, E.-L., Vainio, P., Riekkinen, P., 1998. MRI of the hippocampus in Alzheimers disease: sensitivity, specificity, and analysis of the incorrectly classified subjects. Neurobiol. Aging 19 (1), 23–31.

Mueller, S., Stables, L., Du, A., Schuff, N., Truran, D., Cashdollar, N., Weiner, M., 2007. Measurement of hippocampal subfields and age-related changes with high resolution MRI at 4 T. Neurobiol. Aging 28 (5), 719–726.

Pohl, K.M., Fisher, J., Bouix, S., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2007. Using the logarithm of odds to define a vector space on probabilistic atlases. Med. Image Anal. 11 (5), 465–477.

Pohl, K.M., Fisher, J., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2006. A Bayesian model for joint segmentation and registration. NeuroImage 31 (1), 228–239.

Reuter, M., Fischl, B., 2011. Avoiding asymmetry-induced bias in longitudinal image processing. NeuroImage 57 (1), 19–21.

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

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

Risacher, S.L., Saykin, A.J., West, J.D., Shen, L., Firpi, H.A., McDonald, B.C., 2009. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alzheimer Res. 6 (4), 347.

Rolls, E.T., 2010. A computational theory of episodic memory formation in the hippocampus. Behav. Brain Res. 215 (2), 180–196.

Schneider, J.A., Arvanitakis, Z., Leurgans, S.E., Bennett, D.A., 2009. The neuropathology of probable Alzheimer disease and mild cognitive impairment. Ann. Neurol. 66 (2), 200–208.

Scoville, W.B., Milner, B., 1957. Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry 20 (1), 11.

Shi, F., Fan, Y., Tang, S., Gilmore, J.H., Lin, W., Shen, D., 2010. Neonatal brain image segmentation in longitudinal MRI studies. Neuroimage 49 (1), 391–400.

Shi, F., Yap, P.-T., Gilmore, J.H., Lin, W., Shen, D., 2010. Spatial–temporal constraint for segmentation of serial infant brain MR images. Medical Imaging and Augmented Reality. Springer., pp. 42–50.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. Neuroimage 17 (1), 479–489.

Thompson, P.M., Hayashi, K.M., G.I., de Zubicaray, Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D.H., Gravano, D., Doddrell, D.M., et al. 2004. Mapping hippocampal and ventricular change in Alzheimer disease. Neuroimage 22 (4), 1754–1766.

Thompson, W.K., Holland, D., A.D.N., Initiative, et al. 2011. Bias in tensor based morphometry Stat-ROI measures may result in unrealistic power estimates. Neuroimage 57 (1), 1–4.

Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. IEEE Trans. Med. Imaging 28 (6), 822–837.

Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L.L., Augustinack, J., Dickerson, B.C., Golland, P., Fischl, B., 2009. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19 (6), 549–557.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 897–908.

Wang, L., Miller, J.P., Gado, M.H., McKeel, D.W., Rothermich, M., Miller, M.I., Morris, J.C., Csernansky, J.G., 2006. Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. Neuroimage 30 (1), 52–60.

Wang, L., Shi, F., Li, G., Shen, D., 2013. 4D segmentation of brain MR images with constrained cortical thickness variation. PloS One 8 (7), e64207.

Wang, L., Shi, F., Yap, P.-T., Gilmore, J.H., Lin, W., Shen, D., 2011. Accurate and consistent 4D segmentation of serial infant brain MR images. Multimodal Brain Image Analysis. Springer., pp. 93–101.

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

Xue, Z., Shen, D., Davatzikos, C., 2006. Classic: consistent longitudinal alignment and segmentation for serial image computing. Neuroimage 30 (2), 388–399.

Xue, Z., Wong, K., Wong, S.T., 2010. Joint registration and segmentation of serial lung CT images for image-guided lung cancer diagnosis and therapy. Comput. Med. Imaging Graph. 34 (1), 55–60.

Yushkevich, P.A., Avants, B.B., Das, S.R., Pluta, J., Altinay, M., Craige, C., Initiative, A.D.N., et al. 2010. Bias in estimation of hippocampal atrophy using deformation-based morphometry arises from asymmetric global normalization: an illustration in ADNI 3T MRI data. Neuroimage 50 (2), 434–445.

Yushkevich, P.A., Pluta, J.B., Wang, H., Xie, L., Ding, S.-L., Gertje, E.C., Mancuso, L., Kliot, D., Das, S.R., Wolk, D.A., 2015. Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. Hum. Brain Mapp. 36 (1), 258–287.

Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S., 2010. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53 (4), 1208–1224.

Zeidman, P., Maguire, E.A., 2016. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. Nat. Rev. Neurosci. 17 (3), 173–182.