NeuroImage 125 (2016) 834–847

Contents lists available at ScienceDirect

NeuroImage

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

Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease

Andrea Chincarini [a,*], Francesco Sensi [a], Luca Rei [a,b], Gianluca Gemme [a], Sandro Squarcia [a,b], Renata Longo [g,e], Francesco Brun [f,e], Sabina Tangaro [d], Roberto Bellotti [c,d], Nicola Amoroso [c,d], Martina Bocchetta [h,i], Alberto Redolfi [h], Paolo Bosco [h], Marina Boccardi [h], Giovanni B. Frisoni [k,h], Flavio Nobili [j], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Istituto Nazionale di Fisica Nucleare, Sezione di Genova, I-16146 Genova, Italy
[b] Dipartimento di Fisica, Università degli Studi di Genova, I-16146 Genova, Italy
[c] Dipartimento Interateneo di Fisica, Università degli Studi di Bari, Italy
[d] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
[e] Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, Italy
[f] Dipartimento di Ingegneria e Architettura, Università degli Studi di Trieste, Italy
[g] Dipartimento di Fisica, Università degli Studi di Trieste, Italy
[h] IRCCS Centro San Giovanni di Dio Fatebenefratelli, I-25125 Brescia, Italy
[i] Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
[j] Neurofisiologia Clinica, Dipartimento di Neuroscienze, Oftalmologia e Genetica, Azienda Ospedale-Università S. Martino, Genova I-16132, Italy
[k] University Hospitals and University of Geneva, Geneva, Switzerland

ARTICLE INFO

Article history:
Received 4 June 2015
Accepted 1 October 2015
Available online 26 October 2015

Keywords:
MRI
Image analysis
Longitudinal measure
Alzheimer's disease
Hippocampus

ABSTRACT

Background: Structural MRI measures for monitoring Alzheimer's Disease (AD) progression are becoming instrumental in the clinical practice, and more so in the context of longitudinal studies. This investigation addresses the impact of four image analysis approaches on the longitudinal performance of the hippocampal volume.
Methods: We present a hippocampal segmentation algorithm and validate it on a gold-standard manual tracing database. We segmented 460 subjects from ADNI, each subject having been scanned twice at baseline, 12-month and 24 month follow-up scan (1.5 T, T1 MRI). We used the bilateral hippocampal volume v and its variation, measured as the annualized volume change $\Lambda = \delta v/year(mm^3/y)$. Four processing approaches with different complexity are compared to maximize the longitudinal information, and they are tested for cohort discrimination ability. Reference cohorts are Controls vs. Alzheimer's Disease (CTRL/AD) and CTRL vs. Mild Cognitive Impairment who subsequently progressed to AD dementia (CTRL/MCI-*co*). We discuss the conditions on v and the added value of Λ in discriminating subjects.
Results: The age-corrected bilateral annualized atrophy rate (%/year) were: −1.6 (0.6) for CTRL, −2.2 (1.0) for MCI-*nc*, −3.2 (1.2) for MCI-*co* and −4.0 (1.5) for AD. Combined (v, Λ) discrimination ability gave an Area under the ROC curve (*auc*) = 0.93 for CTRL vs AD and *auc* = 0.88 for CTRL vs MCI-*co*.
Conclusions: Longitudinal volume measurements can provide meaningful clinical insight and added value with respect to the baseline provided the analysis procedure embeds the longitudinal information.

Abbreviations: AD, Alzheimer's Disease; ADNI, Alzheimer's disease Neuroimaging Initiative; *AUC*, Area Under Curve; CTRL, Control Subjects; MCI (*-nc*/*-co*), Mild Cognitive Impairment (non-progressing to AD/progressing to AD); MNI, Montreal Neurological Institute; MRI, Magnetic Resonance Imaging; ROC, Receiver Operating Characteristic; SVM, Support Vector Machine; VOI, Volume Of Interest.

* Corresponding author at: INFN, via Dodecaneso 33, I-16146 Genova, Italy. Fax: +39 010 313358.
E-mail address: andrea.chincarini@ge.infn.it (A. Chincarini).

[1] 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: adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2015.10.065
1053-8119/

Introduction

Among image-based markers, structural information is considered highly informative in the quantification of progression to Alzheimer's disease (AD). This is becoming even more important in the context of longitudinal studies where substantial literature (Hogan et al., 2004; Bateman et al., 2012; McEvoy et al., 2011; Spulber et al., 2013; Lobanova et al., 2014; Leung et al., 2010; Schuff et al., 2009; Rusinek et al., 2003; Fox and Schott, 2004) suggests that longitudinal trend may be pivotal in discriminating a population at risk.

In addition, there is enough scientific evidence supporting the use of the hippocampal geometrical properties (such as the hippocampal volume) as biomarker of early / progression of AD, and the reader is referred to Frankó et al. (2013), Chincarini et al. (2011), Gerardin et al. (2009), Fennema-Notestine et al. (2009) for a sample of studies in the field.

There are now a number of methods to automatically segment the hippocampal structure, many of them featuring high accuracy and reliability (Shen et al., 2002; Morra et al., 2008; Pruessner et al., 2000; Bishop et al., 2011; Wolz et al., 2010b, 2014). In addition, the recently concluded segmentation harmonization effort (see Frisoni et al. (2014), Apostolova et al. (2015)) delivered a set of gold-standard tracings to be used as reference for both human and automatic readers (Bocchetta et al., 2014; Boccardi et al., 2015).

Despite the use of gold-standard segmentations, the reliability and the clinical usefulness of a longitudinal measurement can be hindered by several confounding factors, namely: technical errors (acquisition noises, artifacts, data analysis and algorithmic instabilities) and physiological variability (both intrinsic and due to external conditions such as hydration, lipidic balance, nutrition and hormonal concentration, Duning et al. (2005), Maclaren et al. (2014)). The goal of longitudinal analysis though is to find the long-term trend due to either normal or pathological aging, neglecting the nuisances of both intrinsic and extrinsic variabilities.

Our investigation here looks for possible implementations of a segmentation-based longitudinal marker, aiming at the reduction of variabilities other than the long-term aging contribution. First, we develop a segmentation algorithm on a separate dataset, delivering the hippocampal volume. Then, we segment a large number of MR images from ADNI and use the hippocampal volume to construct a longitudinal marker. This marker is implemented with four algorithmic variations of increasing complexity, meant to enhance the robustness and accuracy of the segmentation over the longitudinal scans. Finally, we assess the marker prognostic potential and estimate under which conditions the longitudinal information is clinically relevant.

Materials and methods

Dataset

MRI scans (1.5 T, T1-weighted) were selected from the ADNI database [2] and downloaded in the original format (DICOM). The subjects' id list is provided in supplemental table S1.

We selected 460 subjects having four scans: two scans at baseline (hereafter labeled *baseline* and *repeat*), 12-month and 24-month scans for a total of $460 \times 4 = 1840$ images.

According to the ADNI evaluators, subjects were grouped in three cohorts consisting of 148 Controls (CTRL), 216 Mild Cognitive Impairment (MCI) and 96 Alzheimer's Disease (AD) (clinical label given at baseline). Coarse statistical description is summarized in Table 1.

MCI subjects were further divided into 121 "MCI progressing to AD" (MCI-*co*) and 95 stable MCI, or"MCI non-progressing" (MCI-*nc*) according to the clinical follow-up which stretched up to 96 months after the baseline scan. A few MCI subjects (8) received more than two labels during follow-up (MCI / AD / normal cognition). They were treated considering the first and the latest evaluation only.

On average, time to AD occurred after 48 (24–84) months (90% confidence bounds) from the baseline.

Image processing

Image processing closely follows the method detailed in Chincarini et al. (2011), save for two procedural differences. We summarize here the main steps applied to each MR image up to the extraction of its Volumes of Interest (VOI), which were used as starting points of the segmentation algorithm.

MR images underwent a series of filters designed for bias B-field reduction, volume normalization, anatomical structure registration and gray level intensity equalization. The two novelties with respect to Chincarini et al. (2011) are the lack of the pyramidal noise filter and the addition of the B-field bias reduction, the latter implemented with the BET algorithm (Smith, 2002). The noise filtering step was avoided to keep the intensity contrast between the hippocampus structure and the adjacent structures (amigdala mainly), which could be impaired by the pyramidal filter. Similarly, the B-field bias correction was introduced to improve on the deformable registration cost function used in the segmentation process.

As result of the pre-processing steps, images were aligned with a 12-parameters affine transformation to the Montreal Neurological Institute (MNI, mazziotta et al. (1995)) space and the mean gray level intensities of the three major brain constituents – cerebro-spinal fluid (CSF), gray matter (GM) and white matter (WM) – were matched to reference values. In addition, aligned images are spatially sampled as the MNI template, that is with isotropic voxels of 1 mm.

Each image was then sampled with 2 VOIs with dimension $30 \times 80 \times 40$ mm each, which were placed in both Medial Temporal Lobes (MTL) so that the hippocampi are anatomically aligned to the VOIs sagittal axes (see Fig. 1 for an example of VOI positioning and content).

Finally, a finer intra-cranial volume correction (icv) is computed by non-linear mapping of the segmented MNI brain mask (provided with the template) onto the affine-registered image and the mask volume is weighted by the affine transformation jacobian. This number is a minor factor (of the order of the unity) and it does not correct for the native volume versus the MNI-space one, as the spatial normalization already compensated for it. It is rather used to adjust for the possible deviations that escape the affine registration. This non-linear-based intra-cranial volume adjustment is used as a hippocampal volume correction factor after the segmentation process.

Segmentation algorithm

The main ground for developing our own segmentation procedure instead of using an existing one was the choice to have it under control and to use a probabilistic atlas approach rather than voxel-based classification techniques.

The procedure (referred in the following as *GDIseg*) requires only the hippocampal VOI in input and it is not too dissimilar from that proposed by Wolz et al. (2010a), save for some details. It was developed on a MR set consisting of 100 T1-weighted MR images and tracings (Frisoni et al. (2014), preliminary release) from the "harmonized protocol for hippocampal volumetry" project (HarP, www.hippocampal-protocol.net), subjects not included in the dataset presented in this investigation.

For the purpose of this investigation we require only two outputs from *GDIseg*: the bilateral hippocampal volume v and a spatial probability map A, which should ideally peak on the hippocampi voxels and quickly fade to zero on all other brain structures. The *GDIseg* algorithm is described in Appendix A.

[2] The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Table 1
Demographics of the test dataset from ADNI.

Cohort	Sample size	M/F	Age [y] (at baseline)	MMSE		
				Baseline	Month 12	Month 24
CTRL	148	77/71	76.5 (70.2–85.9)	29.0 (27.9–30.0)	30.0 (27.0–30.0)	29.0 (27.0–30.0)
MCI-*nc*	95	64/31	77.2 (62.8–86.2)	28.0 (24.0–30.0)	28.0 (23.0–30.0)	28.0 (22.2–30.0)
MCI-*co*	121	74/47	74.7 (63.9–86.0)	27.0 (24.0–30.0)	26.0 (20.0–29.0)	24.0 (18.0–29.0)
AD	96	50/46	76.7 (63.6–87.3)	23.0 (20.0–26.0)	22.0 (13.0–27.0)	19.5 (6.2–27.0)

CRTL = Controls; AD = Alzheimer's Disease; MCI-*nc* = MCI non-converters; MCI-*co* = MCI converters. Number within parentheses show the 90% confidence interval.

Implementations

We implemented the longitudinal analysis procedure with four progressive steps, starting with a naive approach in which all scans are treated separately, to a fully integrated one in which image processing and segmentation are intertwined. A schematic comparison of the four implementations is given in Fig. 2.

All descriptions regarding the hippocampal VOIs have no explicit laterality labels but it is intended that they are run on the left and right VOI separately.

A: independent processing and segmentation

Each scan is treated independently. The icv correction is also computed separately on the four scans; no longitudinal (i.e., time) information is used (Fig. 2A). This implementation serves as base comparison to assess the performance increase of more sophisticated approaches.

B: unified image processing

In this implementation image preprocessing is merged by generating an unbiased within-subject template space, while segmentation follows on each VOI independently (Fig. 2B).

The within-subject template is constructed by generating an average intermediate image H from the 4 scans (baseline, repeat, month 12 and month 24) using robust, inverse consistent registration (Reuter et al., 2012a). The intermediate within-subject template is processed up to the extraction of the hippocampal VOIs according to the Image processing section. The relevant parameters (registration onto the MNI reference, VOI positions and intensity normalization) are passed back to the original scans so that the actual VOIs can be extracted.

This implementation ensures that all 4 scans are treated uniformly and the VOIs are extracted with a very high degree of reproducibility. The icv correction is computed on H only.

C: atlas matrix re-normalization

This implementation shares the same image processing as in "B" but it adds a refinement to the segmentation algorithm (Fig. 2C). This is based on the construction of a single deformation field f^* that summons the main longitudinal variation of the hippocampal shape. Implementation "C" supplements the *GDIseg* algorithm by adding the temporal information in the form of a post-processed probabilistic map A.

Consider the four scans of a single subject and let b_i be the hippocampal VOI extracted from scan i and A_i the related probabilistic atlas. Let also f_{ij} be a deformation field that maps b_i onto b_j ($i, j = 1..4$).

We can define the 4×4 matrix $\mathbf{f}$ whose elements are the f_{ij} and which contains the identity transformation I on the diagonal, with the requirement that $f_{ij} + f_{ji} = I$. Similarly, we can define a matrix $\mathbf{a}$ of probabilistic maps whose elements are $a_{ij} = f_{ij}(A_i)$, i.e., the application of the field f_{ij}

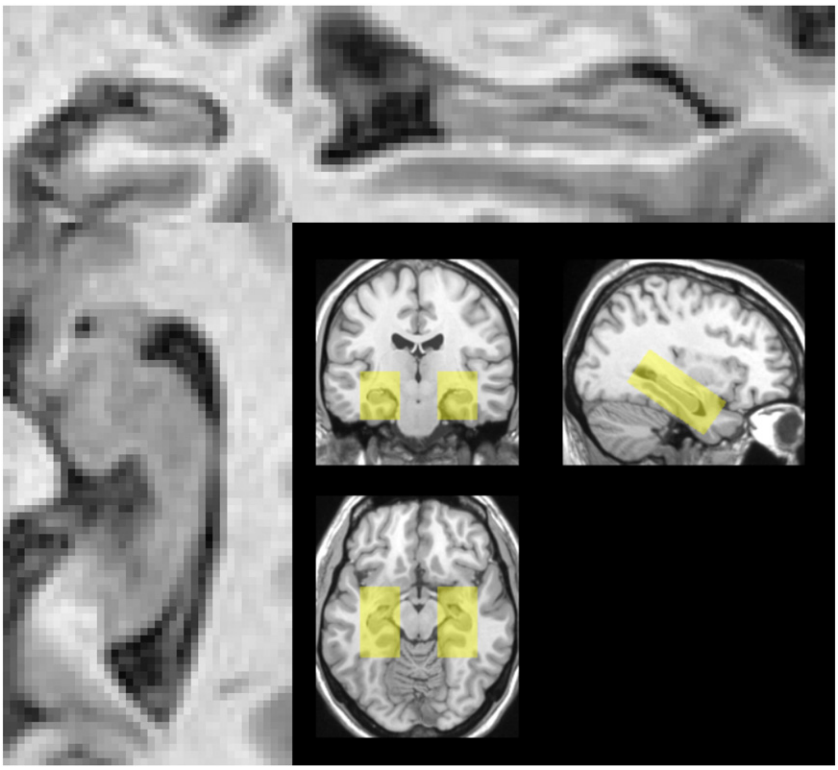

Fig. 1. Positioning and content of a sample hippocampal VOI.

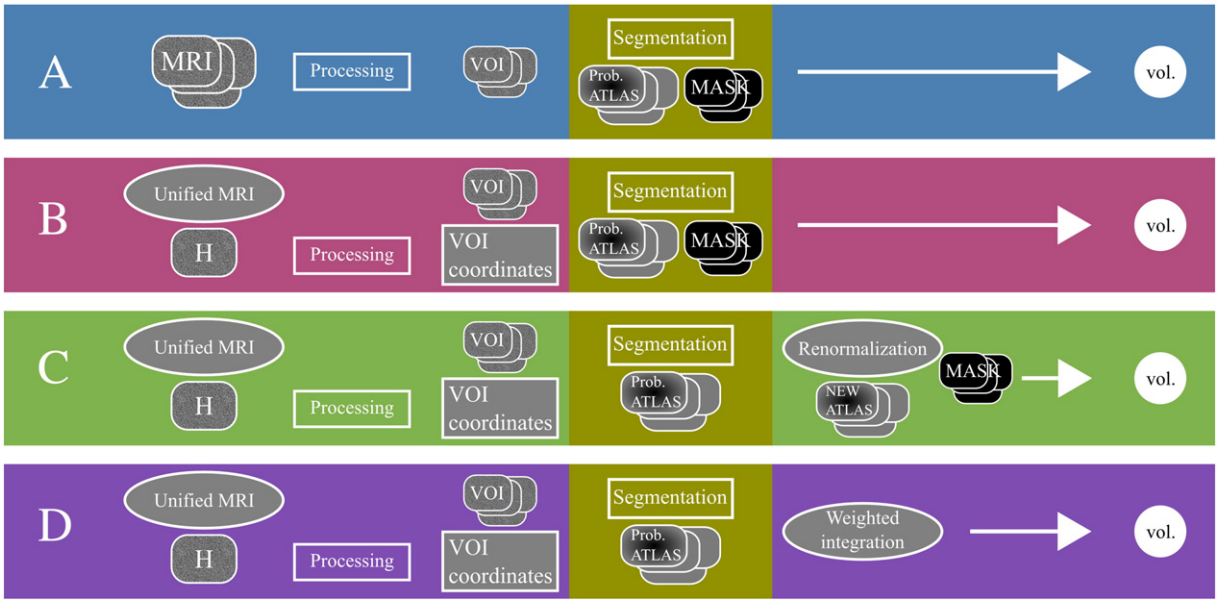

Fig. 2. Schematic flowchart of the four implementations. The four MRI drawings represents the baseline, repeat, month 12 and month 24 scans. In implementation A (A: independent processing and segmentation) all four images follow a separate preprocessing and segmentation path. In implementation B (B: unified image processing) an intermediate image *H* is generated and preprocessing is performed on it; parameters are then mapped back onto the original images to extract the VOIs. In implementation C (C: atlas matrix re-normalization) the VOIs extracted with the B procedure are segmented together with atlas re-normalization. Implementation D (D: weighted integration) avoids the shape segmentation and delivers an equivalent volume only.

to A_i. By definition, the diagonal elements are $a_{ii} = A_i$. Addition, subtraction and multiplication by a constant on the deformation field f are intended to be applied voxel-by-voxel to the displacement vector components. The identity operator I components are by definition all zero.

We now assume that the main contribution to the longitudinal trend can be captured by a linear map of a new operator f^*. The intent of f^* is to capture the mean, long term drift by averaging over the paths from the baseline to the last follow-up scan, so that

$$f_{ij} \simeq \alpha_{ij} f^*, \alpha_{ij} \in [0, 1]$$

A possible choice for α_{ij} could be

$$\alpha_{ij} = \frac{t_j - t_i}{\max_{i,j=1..4}[t_j - t_i]}$$

where t_i is the time of the i^{th} scan. In order to find f^* we average the deformation fields on all paths connecting the earliest to the latest scan.

The generalized expression is

$$f^* = \frac{1}{1 + n_1 + n_2 + \ldots}\left(f_{xy} + \sum_{x<k<y}\left(f_{xk} + f_{ky}\right) + \sum_{x<k<h<y}\left(f_{xk} + f_{kh} + f_{hy}\right) + \ldots\right)$$

where n_r are the number of possible paths from x to y using r intermediates. The simplified expression for 4 scans (taking into account that $t_2 = t_1$) is

$$f^* = \frac{1}{4}(f_{14} + f_{24} + (f_{13} + f_{34}) + (f_{23} + f_{34}))$$

We can now compute the new matrix **f** with elements $\alpha_{ij} f^*$, and hence the new atlas matrix **a**.

We have re-normalized the probabilistic maps a_{ij} to comply with a single deformation field that links the VOIs extracted from the longitudinal scans. The re-normalized a_{ij} are averaged over the columns and then thresholded, to get the binary masks. Then, we apply the icv correction the same way as in implementation "B".

D: weighted integration

In this last implementation images are preprocessed as in "B" and segmentation undergoes a post-processing step, this time though we drop the requirement of an actual binary mask per VOI, in favor of the volume information alone (Fig. 2 D).

For each subject and bilateral VOI we define two new maps:

$$A_p = \prod_{j=1..4} A_j$$
$$A_m = \max_j A_j$$

where j is the index to the baseline, repeat, 12 month and 24 month scans; the 'max' is taken voxel-wise over the four A_j. If x represents the gray intensity in any voxel, the quantity:

$$W(k, y) = \sum_{x \in \mathrm{VOI}_k} x\, A_y$$

is the weighted sum of the intensity values over the volume VOI_k. We now define the longitudinal volumes as:

$$v_j = \hat{v}\, \frac{W(j, m)\, W(1, p)}{W(1, m)\, W(j, p)}$$

The normalization constants $\hat{v}$ is the mean volume over the baseline and repeat scans, as given by *GDIseg*.

In short, this formulation modulates the intensities in the bigger map (A_m, which includes the hippocampal boundary) with the inner intensity values (A_p, where all segmentations agree).

Performance metrics

We checked the performance of all described procedures with four metrics. The first one (reliability) is simply a quality control to assess the robustness of *GDIseg* on a large number of images. Then we looked at the test/re-test performance (reproducibility) and at the longitudinal trend. Finally we checked whether the longitudinal information can

improve on the accuracy when used as combined biomarker together with the volume.

Reliability

The segmentation procedure was applied without human intervention to 1840 images from the ADNI database. A quality control test checks whether and on how many images the procedure crudely fails. This control does not imply a "correct" hippocampus segmentation – in terms of harmonized protocol – it only points out possible failures in the preprocessing and in the segmentation procedure. To perform this test we construct two identical statistics Re_{voi} and Re_{mask}:

$$Re_{voi} = \min_{t,L,R}\left\{ \max_{i}[r(VOI, TB_i)] \right\}$$
$$Re_{mask} = \min_{t,L,R}\left\{ \max_{j}[r(mask, TM_j)] \right\}$$

where r is the Pearson correlation coefficient, the 'max' is taken on the templates and the 'min' is taken among scans (t) and laterality (L,R). Template Boxes (TB) and Template Masks (TM) are the hippocampal VOIs and manual tracings on the HarP image dataset (see Appendix A).

This test computes the best correlation coefficient among the VOI intensities and each TB, as well as among the segmented mask and each TM, then keeping the lowest among these values with respect to the number of scans and laterality. In other words, from each subject we get 8 VOIs and 8 hippocampal tracings (bilateral regions on 4 scans). If either one or more are too distant from its nearest template (in terms of correlation coefficient), the subject is flagged for visual inspection. This formulation assumes that the HarP subjects are sampled as to represent all relevant physiological variability.

Mishaps in image processing (intensity normalization for instance), in the VOI extraction (registration) and in the segmentation algorithm will result in either one or both statistics to be significantly impaired. Visual inspection of outliers and most extreme values follows, to understand the reasons of failure and ensure that outliers are indeed the only images on which the automatic procedure failed. Subjects failing this test are discarded.

Reproducibility

We addressed the statistics of the segmentation volumetry comparing baseline and repeat scans. This tests is crucial for informed use in both research and clinical settings. Test/re-test reproducibility — i.e., how the outcome measure varies when computed over two repeat scans acquired in the absence of plausible biological variability — is a critical measure for reliable biomarkers. The considered quantity is

$$\Delta = 2\frac{v_r - v_b}{v_r + v_b} = \frac{v_r - v_b}{\hat{v}}$$

where v_b and v_r are the baseline and repeat hippocampal volumes respectively.

Longitudinal trend

The annualized volume change Λ (expressed in mm^3/year) is defined as the slope of the least-squares linear fit of the longitudinal volume measures v_i versus time:

$$v_i - \xi_i = \Lambda\, t_i + \beta$$

where ξ_i and β are the residuals and the intercept respectively, and $i = 1..4$ tags the baseline, repeat, 12-month and 24-month scans. To make Λ more robust we did not choose to split measures into 0–12 m and 12 m–24 m intervals as in Schuff et al. (2009).

A linear model using age, sex and cohort as predictors found cohort and age as significant ($p < 10^{-4}$). We adjusted Λ for age using de-correlation.

Then, we used de-correlation to cross-check whether Λ maintains significant prognostic performance after the adjustment for $\hat{v}$ and mini-mental state examination (MMSE) score.

Combined markers

The added complexity to derive a longitudinal biomarker – albeit a simple one based on the hippocampal volume drift over time – should be balanced by an increased prognostic potential.

ROC analysis on the combined volume and trend indexes was computed with a linear discriminant. We used a support vector machine (SVM) classifier on the feature set $(\hat{v}, \Lambda)$ and we considered the distance from the separating plane as the new marker. Its performance was compared to that of $\hat{v}$ and Λ alone.

Software and statistics

Image processing was carried out on a dedicated computational farm running the LONI pipeline software (www.loni.ucla.edu), using MATLAB (www.mathworks.com) and ITK (www.itk.org) as algorithm libraries. All statistical analyses were carried out within the MATLAB environment.

The Λ score was adjusted for specific variables by de-correlation using linear regression in the following manner:

$$\Lambda_i^{adj} = \Lambda_i - \left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right)$$

where Λ_i is the score from the ith subject, x_{ij} is variable j of subject i to be adjusted for, and $\hat{\beta}_i$ is estimated using a least squares fit $\Lambda_i = \beta^0 + \sum_j \beta_j x_{ij}$ to the considered dataset. We adjusted for either age or for MMSE, as they are among the major confounders and we checked whether Λ^{adj} still carried information. No dominant non-linear relationships were observed when inspected by scatter plots. Consequently, a linear adjustment was considered sufficient.

A SVM classifier with linear kernel was trained on CTRL vs. MCI-*co* cohorts. The trained classifier was used to assess the AD and MCI-*nc* cohorts. The combined marker was the distance from the SVM separating plane. ROC analysis of the combine marker $(\hat{v}, \Lambda)$ on CTRL vs. MCI-co are given with a 20-fold cross-validation method. Right and left structures were treated separately.

Confidence intervals on AUC values in Table 3 were computed by bootstrapping (1000 times) and by using the bias-corrected percentile method (Martinez, 2011). Statistical significance in Table 4 versus the null AUC and among different markers was carried on according to Hanley and McNeil (1982, 1983).

The estimation of confidence intervals on the AUC can be carried out with several methods, each delivering slightly different values. Hence the comparison and compatibility among tests in Tables 3 and 4 should take into consideration that confidence intervals are method-dependent estimates. We considered seven methods, parametric and non-parametric: Hanley–McNeil (parametric); Mann–Whitney, Logit and Bootstrap (non-parametric, Gengsheng Qin and Hotilovac (2007)); Maximum variance (non-parametric, Cortes and Mohri (2004)); and Wald, and Wald/continuity-corrected (non-parametric, Kottas et al. (2014)).

For instance, the width of the confidence interval on $\hat{v}_L$ for the CTRL/AD cohorts (implementation D, AUC = 0.89 in Table 3) ranges from 0.06 (Hanley–McNeil) to 0.09 (Mann–Whitney); in numbers 0.86 — 0.92 and 0.84 — 0.93. Another example with Λ_R, implementation C and CTRL/MCI-*co* (AUC = 0.78) shows a substantially similar interval width of all methods (0.74 — 0.82 Hanley, Mann–Whitney; 0.73 — 0.84 Maximum Variance). The bias-corrected percentile bootstrap was regarded as a safe estimate as it did not require any assumption about the normality of the log-transformed AUC (Ahn and Yim, 2009).

Results

Results on volume and longitudinal feature ($\hat{v}$ and Λ) are given after correction for age (de-correlation). Hippocampal volumes are given after correction for icv and in the MNI space with spatial sampling of $1 \times 1 \times 1$ mm.

Quality control

Fig. 3 shows the distribution of Re_{voi} and Re_{mask} for all 460 subjects. There are three distinctive outliers which are excluded from subsequent analyses and whose inconsistent VOIs and tracings are shown aside (Figs. 3a, b and c). Potential outliers – placed in the low value regions of the Re_{voi}/Re_{mask} scatter plot – are visually screened to ensure that they are correctly classified as proper VOI and hippocampal tracings.

One of the outliers (Fig. 3a) stems from a blind injection: a null image (white noise only) was placed in the analysis pipeline on purpose, in order to test the reliability of the whole analysis procedure. Another outlier (Fig. 3b) is due to incorrect brain spatial registration, causing the VOIs to be misplaced. The third one (Fig. 3c) is due to the peculiar atrophy conditions, which has no related template in the HarP subject selection.

Reproducibility

The relative volume variation over baseline and repeat scan is given for the A, B, C and D implementations in percent units (%), mean and standard deviation: $\Delta_A = -0.1 \pm 3.5$, $\Delta_B = -0.1 \pm 2.7$, $\Delta_C = 0.0 \pm 0.1$ and $\Delta_D = 0.1 \pm 1.2$. The absolute value of the standard deviation σ_v over the quantity $v_r - v_b$ is: $\sigma_v^A = 156$, $\sigma_v^B = 128$, $\sigma_v^C = 5$ and $\sigma_v^D = 68$ (units in mm^3).

Longitudinal trend

Mean Λ values over cohorts and implementations are shown in Table 2.

Λ is significantly correlated to the baseline volume $\hat{v}$ in implementations B, C and D. The Pearson correlation r is $r_A = 0.05$ ($p = 0.12$), $r_B = 0.09$ ($p = 0.01$), $r_C = 0.41$ ($p < 10^{-4}$) and $r_D = 0.37$ ($p < 10^{-4}$). In words, volume loss is higher (in absolute value, i.e., more negative numbers) in smaller structures.

In terms of cohort discrimination, Fig. 4 shows the distribution and ROC curves of Λ for the right and left hippocampus separately, where it is apparent that the AUC steadily increases with the implementation complexity (from A → D). Comprehensive results on the AUC of $\hat{v}$ and Λ are summarized in Table 3.

The average bilateral AUC remained significant ($p < 10^{-4}$) after de-correlating baseline MMSE score ($AUC_A = 0.64$, $AUC_B = 0.64$, $AUC_C = 0.67$ $AUC_D = 0.70$) and volume $\hat{v}$ ($AUC_A = 0.66$, $AUC_B = 0.66$, $AUC_C = 0.63$ $AUC_D = 0.68$).

A derived alternative marker is the bilateral average of the relative annualized volume loss

$$\lambda = \frac{1}{2}\left([\Lambda/\hat{v}]_R + [\Lambda/\hat{v}]_L\right)$$

expressed in *%/year*. Values (mean and standard deviation) are: $\lambda = -1.6(0.55)$ for CTRL, $\lambda = -2.2(1.0)$ for MCI-*nc*, $\lambda = -3.2(1.2)$ for MCI-*co* and $\lambda = -4.0(1.5)$ for AD (λ results are calculated on implementation D).

In order to better specify the expected levels of relative annualized loss in potentially pathological subjects, the CTRL cohort is compared to an 'AD-like' cohort consisting of subjects with AD together with subjects who subsequently developed AD (MCI-*co*). Using implementation D, we selected three cut-offs relevant for accuracy (*acc*), sensitivity (*sens*) and specificity (*spec*): $\lambda = -2.19$ (*sens* = 0.83, *spec* = 0.85, *acc* = 0.84, maximum accuracy criterion); $\lambda = -1.28$(*sens* = 0.32, *spec* = 0.95, *acc* = 0.69); $\lambda = -2.94$ (*sens* = 0.95, *spec* = 0.69, *acc* = 0.80). In this example the area under the ROC curve is *AUC* = 0.90 and a graphical representation of the two distributions is shown in Fig. 6.

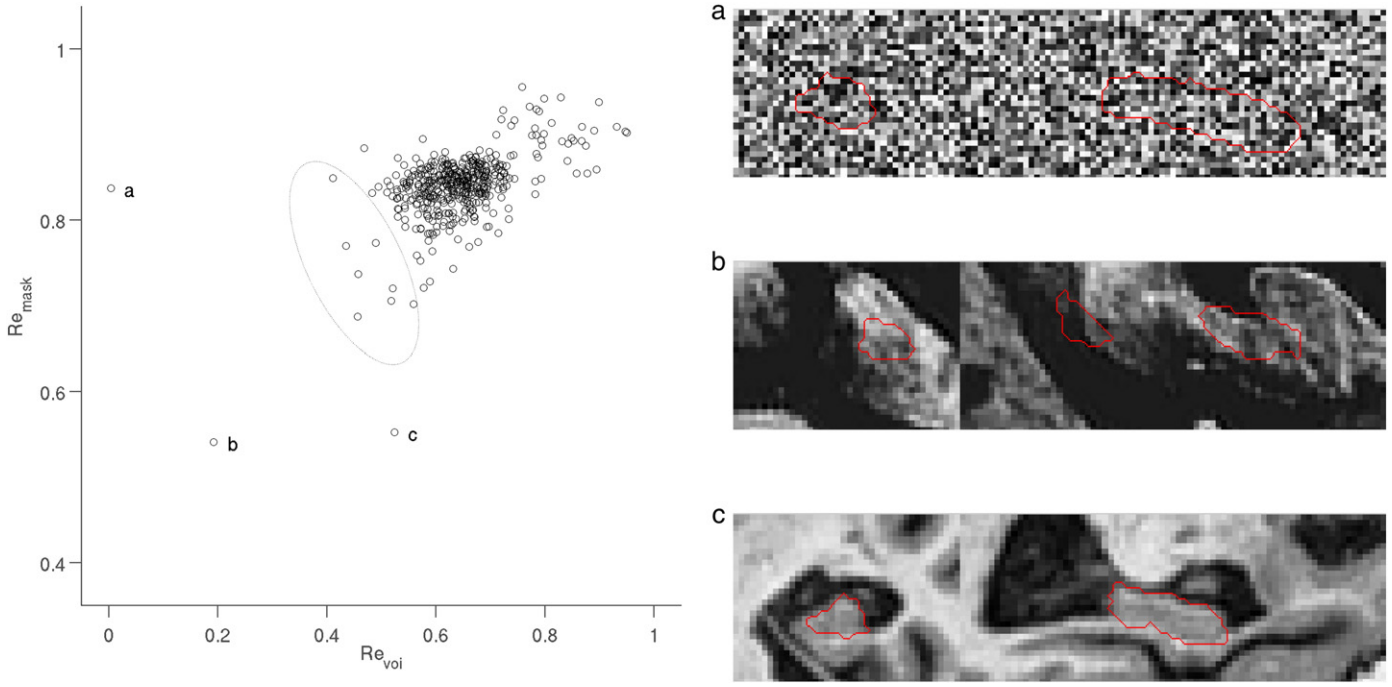

Fig. 3. Left: reliability scatter plot over VOIs (x-axis) and hippocampal masks (y-axis). Each circle represents a subject. Lower scores are an indication of either improper image processing or biased template sampling. *a,b* and *c* are outliers. The dotted outline shows the subject who underwent visual inspection. Right: coronal and sagittal view of the three outlier VOIs. The red outline shows the *GDIseg* hippocampal tracing.

Table 2
Mean Λ values.

		CTRL	MCI-*nc*	MCI-*co*	AD
R	A	−75.90 (84.62)	−80.04 (89.81)	−135.80 (93.15)	−135.54 (90.59)
	B	−72.60 (67.46)	−96.99 (69.46)	−129.63 (87.30)	−140.09 (83.22)
	C	−69.32 (47.40)	−98.39 (66.19)	−131.29 (67.50)	−154.58 (73.46)
	D	−76.27 (23.40)	−91.96 (37.74)	−124.41 (45.34)	−143.10 (54.22)
L	A	−61.83 (79.76)	−73.76 (96.06)	−111.40 (88.74)	−108.91 (85.86)
	B	−56.48 (53.35)	−61.14 (86.43)	−95.81 (72.96)	−101.47 (91.23)
	C	−59.82 (43.71)	−72.01 (54.72)	−113.99 (59.09)	−133.65 (60.39)
	D	−63.34 (25.19)	−77.70 (40.32)	−108.19 (44.21)	−122.80 (47.32)

Annualized volume change (Λ) in $mm^3/year$ (mean and standard deviation).

Combined markers

The benefit of adding the trend information Λ to the average baseline volume $\hat{v}$ is summarized in Table 4 and graphically shown in Fig. 5. In each comparison, we marked whether the combined information fared significantly better than either factors. Considering a total of 3 (group comparisons) × 4 (implementations) × 2 (laterality) = 24 tests, adding atrophy rate information to the baseline volume resulted in a significantly higher AUC (compared to that of the volume alone) in 14 tests.

Sample size calculation

To determine the power of the different implementations in detecting effects on hippocampal volume loss over time we estimated the sample size needed in a hypothetical treatment trial to measure a 25% slowing in Λ with $\alpha = 0.05$ significance level and a power $1 - \beta = 0.8$.

Using the formula

$$n = \frac{2\sigma^2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta^2}$$

we chose $\delta = 0.25\overline{\Lambda}$ where $\overline{\Lambda} = (\Lambda_R + \Lambda_L)/2$ is the bilateral mean atrophy rate of the corresponding clinical group, σ their standard deviation and z values are $z_{1-\alpha/2} \simeq 1.96$ and $z_{1-\beta} \simeq 0.84$ respectively. For each patient group, the estimated sample sizes are displayed in Table 5.

Discussion

In this study we evaluated the impact of using the longitudinal information deriving from serial MRI scans as an added value compared to 'spot' baseline scans in patients with MCI or AD as compared to controls. The assumption was that atrophy rate with time could be a neurodegeneration marker independent of single atrophy measures. We showed that with a 2-y observation time this is true only if adequate post-processing is performed. On the other side, this means that 2-y repeated measures are useless when only a raw estimate of atrophy rate is performed 'on the fly', that is with a simple algorithm that does not embed the longitudinal information.

We compared four possible algorithmic implementations of a volume marker in a longitudinal context, where the longitudinal information is taken into account with different degrees both in the pre-processing and post-processing steps. The first implementation (A) is considered for comparison only.

The longitudinal information is translated into a simple measure Λ, which estimates the hippocampal volume drift (atrophy rate) in time; Λ is then used as a biomarker – alone and in combination with the average baseline volume $\hat{v}$ – to assess its potential in discriminating among relevant clinical groups.

All procedures are fully automated and implement an internal quality check.

Conceptually, the most similar work to this one is Wolz et al. (2010b) – where the longitudinal (i.e., time) information is embedded in the segmentation workflow – and partially similar to McEvoy et al. (2011). We conclude that clinical insight into AD development of subject initially classified as MCI can be derived from quantitative measures processed simultaneously from multiple time points, and that these measures are more consistent than single-time point ones.

To further reduce the atrophy rate uncertainties we could have used several more time points. This however would be an impractical protocol to implement outside clinical trials. Similarly, using two time points only (i.e., 0–12 m) would result in a larger error and a lower discrimination power (Wolz et al., 2010b).

Quality control

All procedures need a stable segmentation, which in turns depends on an accurate VOI placing. Segmentation accuracy with respect to the expert tracing is comparable to results in literature: the LEAP method (Wolz et al., 2010a) DICE index ≃ 0.85; adaboost, ada-SVM and Freesurfer (Morra et al., 2010) Precision ≃ 0.71 − 0.84, Recall ≃ 0.73 − 0.87; and in Lötjönen et al. (2011) DICE index ≃ 0.87.

In this study the supplemental Re_{voi} and Re_{mask} statistics are used as warning indicators of outliers as they compare a new VOI and related segmentation with the reference templates. If the templates do not sample the population extensively enough we may incur in extreme statistic values. In the particular example shown in Fig. 3c, the VOI and its segmentation are not necessarily outliers per se; they are rather given a low rank due to the lack of similar templates. In facts, while Re_{voi} captures structure other than the hippocampus, Re_{mask} refers to the segmentation alone, therefore its score is below the average.

Other VOIs with significant and widespread atrophy dwell in the lower *Re* region for the same reason. Although these cases might bear little clinical significance, an extension of the template database would favorably impact the finding of true outliers.

In the case of the purely noisy image (blank test) of Fig. 3a, Re_{mask} value still ranks among acceptable numbers while $Re_{voi} = 0$; this is explained because *GDIseg* is based on atlas deformation and the transformation constraints on the deformation field (such as the use of the demons algorithm and the smoothing parameters) are weakly affected by noise. In addition, the use of the intra-subject template and the averaged deformation field avoid the pitfalls of overestimating the changes in the atrophy rate (Thompson and Holland, 2011).

Reproducibility

The standard deviations in implementation A and B are rather conspicuous, that is in comparison to the volume change one would want to measure to discriminate among cohorts. Implementation C has a definitely lower mark, but this value is heavily biased by the re-normalization algorithm and doesn't represent the true variability. Rather, it represents the error due to the threshold algorithm when applied to the averaged probability matrix a_{ij}.

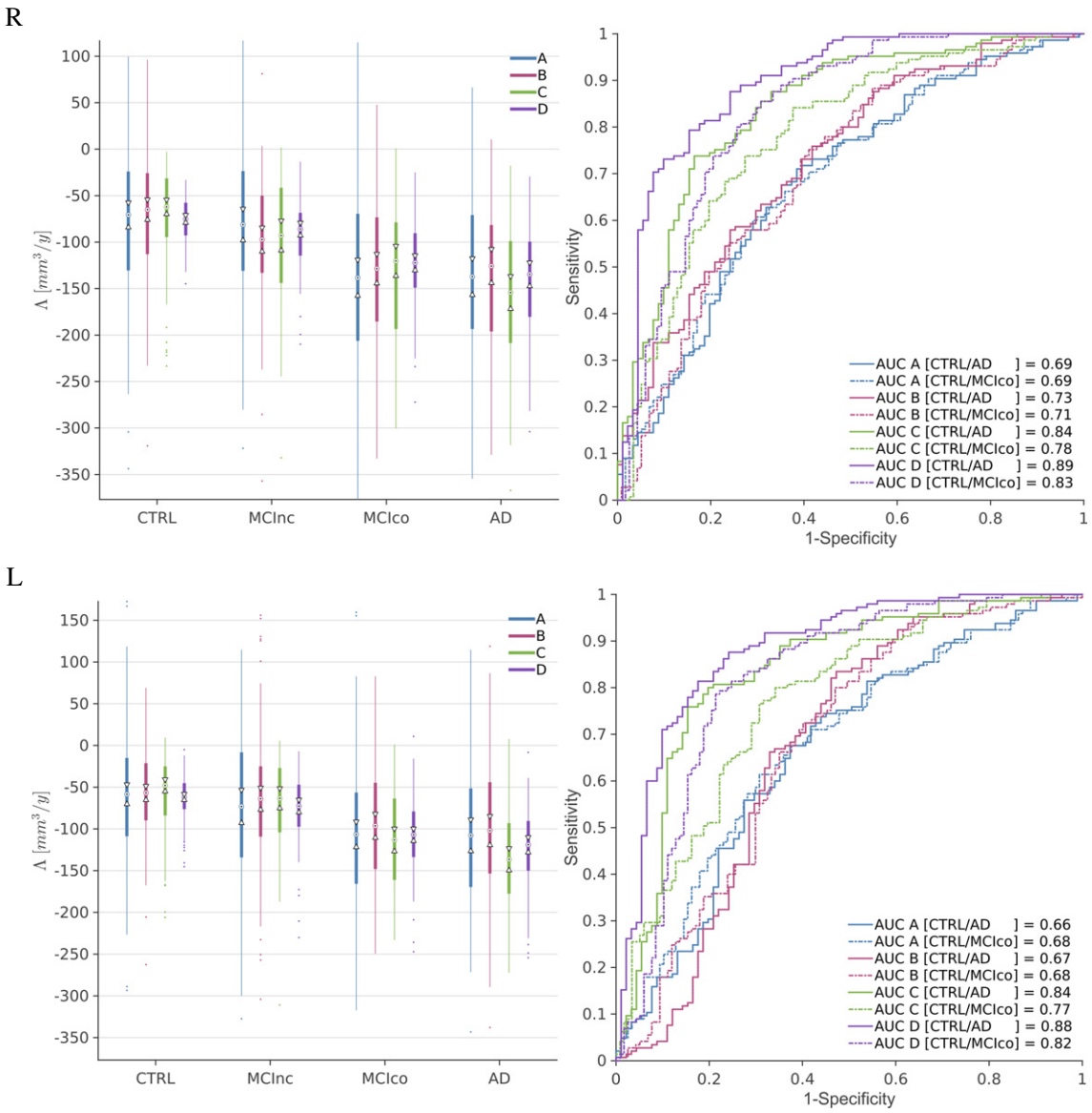

Fig. 4. Distribution of Λ for the right hippocampus (top) and left hippocampus (bottom) on Controls (CTRL), Mild Cognitive Impairment non-converters/converters (MCI-nc/MCI-co) and Alzheimer's Disease (AD) subjects. The median and its 95% conf. interval are marked with a black dot and triangles on each bar. The related ROC curves and area under the curves (AUC) are shown on the right plots.

The value of σ_D^y though reflects the true difference between the baseline and the repeat scan, due to acquisition and processing noises. That is, in implementation D the probability atlas is fixed and there is no threshold step involved.

The difference among implementations can also be appreciated with the normal probability plot for Δ (supplemental Fig. S2), where deviation from the Gaussian distribution is rather marked for implementation A and B.

Comparison to literature shows that results similar to the basic implementations A and B are obtained in Maclaren et al. (2014) (with a total coefficient of variation of ≃ 3 % on the hippocampus and using Freesurfer).

Further methodological considerations

In ADNI, subjects were scanned at different sites and with different MRI equipment. Besides, follow-up images could have been acquired with scanner models other than those used at baseline.

The ADNI protocol goes a great length in assuring reproducibility among sites (Jack et al., 2008) and in addition, other studies showed that ADNI-like acquisitions and optimized analysis procedures (longitudinal processing in particular) are robust across sites, regardless of MRI system differences (see Jovicich et al. (2013) for a detailed analysis). There are though fewer studies combining intra-site and inter-site reproducibility – i.e., measuring the same participants on a variety of scanners – a condition which is relevant in the longitudinal paradigm. In their study, Reig et al. (2009) found that pooling of different sites data can add a significant error compared to intra-site variability, particularly in single-modality (T1) segmentations.

We looked for subjects whose record showed the use of different MRI machines. A survey of the CTRL cohort indicated that 42 out of 148 subjects (≃28 %) were acquired with different scanner models at some follow-up visit (with respect to the MRI system used at baseline).

The potential added variability was gauged with a direct comparison of the statistics using the non-parametric Kolmogorov–Smirnov test. The application to the sample of 106 CTRL (same scanner model across longitudinal measures but different cross-sectionally) and 42 CTRL subjects (different scanner model both in longitudinal measures and cross-sectionally) found no significant difference the Λ statistics, regardless of the implementation.

Table 3
Performance (AUC).

Feat.	Impl.	CTRL/MCI-*nc*	CTRL/MCI-*co*	CTRL/AD
$\hat{v}_R$	A	0.71 (0.65–0.74)	0.79 (0.75–0.83)	0.86 (0.81–0.89)
	B	0.71 (0.65–0.75)	0.79 (0.75–0.83)	0.85 (0.81–0.88)
	C	0.71 (0.66–0.77)	0.82 (0.77–0.85)	0.87 (0.83–0.90)
	D	0.71 (0.66–0.76)	0.82 (0.78–0.85)	0.87 (0.83–0.90)
$\hat{v}_L$	A	0.72 (0.67–0.78)	0.82 (0.79–0.86)	0.88 (0.85–0.91)
	B	0.72 (0.68–0.77)	0.83 (0.78–0.86)	0.88 (0.83–0.91)
	C	0.73 (0.68–0.78)	0.84 (0.80–0.87)	0.89 (0.85–0.92)
	D	0.73 (0.67–0.77)	0.84 (0.80–0.87)	0.89 (0.85–0.92)
Λ_R	A	0.52 (0.46–0.57)	0.69 (0.64–0.73)*	0.69 (0.63–0.73)*
	B	0.60 (0.55–0.66)	0.71 (0.66–0.75)*	0.73 (0.68–0.78)*
	C	0.64 (0.58–0.69)	0.78 (0.73–0.82)	0.84 (0.80–0.88)*
	D	0.63 (0.57–0.69)	0.83 (0.79–0.87)	0.89 (0.85–0.92)
Λ_L	A	0.55 (0.49–0.60)*	0.68 (0.63–0.73)*	0.66 (0.60–0.71)*
	B	0.54 (0.47–0.59)*	0.68 (0.63–0.73)*	0.67 (0.62–0.73)*
	C	0.56 (0.50–0.61)	0.77 (0.72–0.80)	0.84 (0.79–0.87)
	D	0.60 (0.55–0.67)	0.82 (0.77–0.86)	0.88 (0.84–0.91)
$(\hat{v},\Lambda)_R$	A	0.68 (0.62–0.74)	0.83 (0.79–0.87)*	0.89 (0.85–0.91)
	B	0.71 (0.66–0.77)	0.83 (0.78–0.86)*	0.89 (0.85–0.91)
	C	0.71 (0.66–0.76)	0.85 (0.81–0.88)	0.90 (0.86–0.93)
	D	0.70 (0.64–0.76)	0.87 (0.84–0.90)	0.92 (0.88–0.94)
$(\hat{v},\Lambda)_L$	A	0.72 (0.66–0.76)	0.85 (0.81–0.88)	0.89 (0.86–0.92)*
	B	0.69 (0.64–0.75)	0.84 (0.81–0.88)	0.88 (0.84–0.91)*
	C	0.70 (0.65–0.76)	0.85 (0.82–0.88)	0.91 (0.87–0.93)
	D	0.71 (0.66–0.75)	0.88 (0.84–0.90)	0.93 (0.90–0.95)

Area under the ROC curve. Numbers within parentheses are the 95% confidence interval. The '*' indicates significant difference ($p < 0.001$) between implementation D and A, B or C for each respective feature and cohort comparison.

Nonetheless, the use of different models in the longitudinal acquisition could show up in the linear fit residuals ξ (cfr. Longitudinal trend section). Indeed, testing the ξ distributions revealed a significant alteration in implementation A only ($p < 0.001$), which would suggest that the adoption of an intra-subject template (used in B, C and D) is sufficient to tame the inter-scanner reproducibility uncertainty. This finding agrees with Jovicich et al. (2013), where the introduction of longitudinal methods for volumes extraction provides a lower and more homogeneous reproducibility error across different scanners.

Another point is the role of laterality. In this study we treated left and right hippocampi equally and separately to avoid any laterality bias in the results.

The significance of a performance superiority of the left side was investigated by comparing the R and L AUC values with a *t*-test, regardless of the implementation and cohort comparison, grouping only by feature ($\hat{v}$, Λ and $(\hat{v},\Lambda)$). For instance, we tested the pooled set of AUC values for $\hat{v}_R$ vs. $\hat{v}_L$ taking all implementations (A–D) and cohort comparison shown in Table 3 (i.e., 12 values). The one-sample *t*-test was used to assess whether the mean of the difference $AUC_L - AUC_R$ was compatible with zero.

Results indicated that the R/L AUC difference was significant for $\hat{v}_L > \hat{v}_R$, ($p < 0.001$), moderately significant for $\Lambda_R > \Lambda_L$ ($p < 0.01$) and not significant for $(\hat{v},\Lambda)$.

The left hippocampus is usually smaller but AD prediction accuracy is less clearly tied to laterality, even though the left side seems to have a prominent role as discussed in Apostolova et al. (2010), Okonkwo et al. (2012). Our findings are in keeping with a meta-analysis pooling together data from several studies, showing that left hippocampal atrophy is usually more severe than the right one (Shi et al., 2009) and with Frankó et al. (2013), where the volume loss in MCI and AD was significantly lower in the left hemisphere than in the right one.

Speculation on the weight of laterality in AD prediction is outside the scope of this study. There are though important physiological findings linking the hippocampal laterality to potential mechanisms of neurodegeneration. In a series of elderly subjects with cognitive disturbance of increasing degrees of severity, a serum marker of oxidative stress was shown to directly correlate with glucose metabolism of the left temporal lobe – including medial structures – but not of the right one (Picco et al., 2014). Also, the multifunctional mitochondrial enzyme 17 β-hydroxysteroid dehydrogenase type 10, with high-affinity binding to amyloid-beta peptides, is more expressed in the left than in the right hippocampus in patients with AD but not in patients with vascular dementia (Hovorkova et al., 2008).

That said, the bilateral average usually offers a more robust estimator. In all implementations the standard deviation of the bilateral average (σ_{RL}) is smaller than the mono-lateral counterparts. The relative measure $2\sigma_{RL}/(\sigma_R + \sigma_L)$ ranges in 92%–96% for $\dot{v}$ and 80%–90% for Λ. This suggests that informed clinical use of atrophy rate should take into account both hippocampi, as we did in Table 5 and in Fig. 6.

Longitudinal trend and combined markers

The annualized volume loss (atrophy rate) is in par with literature results (Barnes et al., 2009; Leung et al., 2010). Although other authors report different average values (Morra et al., 2009; Wolz et al., 2010b; Schuff et al., 2009), these values do not contrast with our findings due to the relatively large reported confidence intervals and possibly because of a potential difference in region definition, subjects selection and methodology, as also discussed in the Barnes et al. (2009) meta-analysis.

Compared to a more recent meta-analysis by Fraser et al. (2015), the present annual atrophy rate in elderly controls (1.6%) is higher than the mean estimate reported in the meta-analysis (1.12%). However, there is a rather high variability among studies, mainly accounted for by aging and segmentation methods since 42% of reviewed papers reported values close or even higher than ours. In Fraser et al. (2015) atrophy rate in the elderly was lower with automatic segmentation methods than with the reference manual segmentation. This bias is attributed from the authors to inconsistencies of automatic methods allowing the

Table 4
Performance comparison.

Impl.		CTRL/MCI-*co*			CTRL/AD			MCI-*nc*/MCI-*co*		
		$\hat{v}$	Λ	$(\hat{v},\Lambda)$	$\hat{v}$	Λ	$(\hat{v},\Lambda)$	$\hat{v}$	Λ	$(\hat{v},\Lambda)$
R	A	0.79	0.69	$0.83^{*\dagger}$	0.86	0.69	$0.89^{\dagger}$	$0.58^{\ddagger}$	0.67	0.66^{*}
	B	0.79	0.71	$0.83^{*\dagger}$	0.85	0.73	$0.88^{\dagger}$	$0.58^{\ddagger}$	0.63	0.64
	C	0.82	0.78	$0.85^{\dagger}$	0.87	0.84	$0.90^{\dagger}$	0.62	0.64	0.66
	D	0.82	0.83	$0.87^{*\dagger}$	0.87	0.89	0.92^{*}	0.62	0.71	0.72^{*}
L	A	0.82	0.68	$0.85^{\dagger}$	0.88	0.66	$0.90^{\dagger}$	0.61	0.62	0.66
	B	0.83	0.68	$0.84^{\dagger}$	0.88	0.67	$0.88^{\dagger}$	0.61	0.63	0.67^{*}
	C	0.84	0.77	$0.85^{\dagger}$	0.89	0.84	$0.91^{*\dagger}$	0.64	0.71	0.71^{*}
	D	0.84	0.8	$0.88^{*\dagger}$	0.89	0.88	$0.93^{*\dagger}$	0.64	0.72	0.73^{*}

Performance (AUC) comparison for $\dot{v}$, Λ and the combined marker. Significant changes ($p < 0.001$) are marked as '*' for the test $(\hat{v},\Lambda)$ vs. $\dot{v}$; '†' for the test $(\hat{v},\Lambda)$ vs. Λ. '‡' shows the AUC which are not significantly different from 0.5.

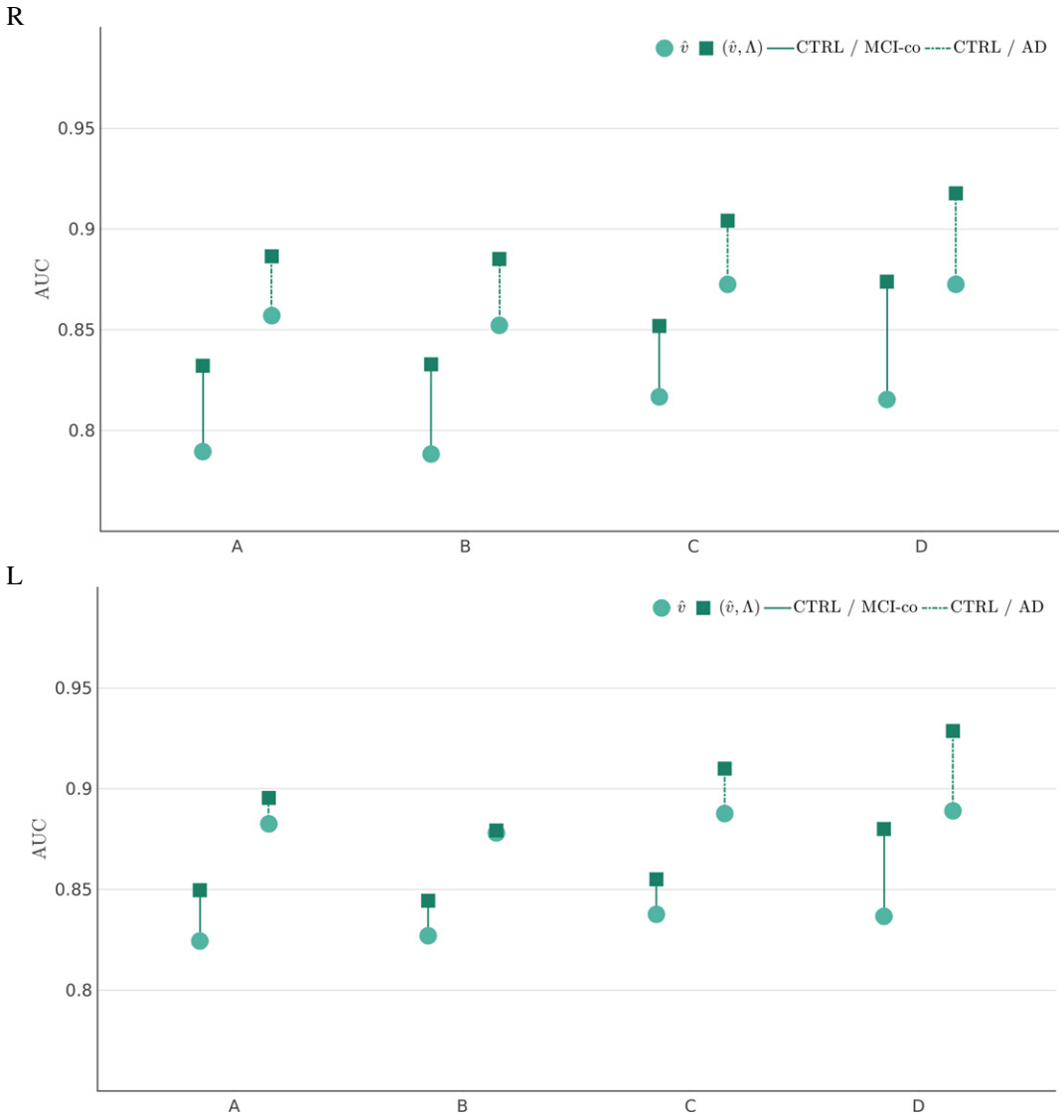

Fig. 5. Baseline volume $\hat{v}$ and combined markers $(\hat{v}, \Lambda)$ performance comparison and implementation dependence. Area under the ROC curve (AUC) is shown for CTRL vs. MCI-*co* (full line) and CTRL vs. AD subjects (dotted line).

analysis of regions close to the hippocampus but with a lower atrophy rate than the hippocampus itself. This explanation seems consistent with the findings by Mulder et al. (2014) showing that manual segmentation produced higher atrophy rates than Freesurfer (Reuter et al., 2012b) or FIRST (Patenaude et al., 2011). In this regard, our method seems closer to the reference manual segmentation than other automatic methods and, as such, potentially more accurate.

In terms of discrimination power among groups, raw performance of volume is comparable to Lötjönen et al. (2011) (CTRL / AD AUC = 0.89) and atrophy rate relates to those in Wolz et al. (2010b) where their method delivers AUC = 0.88–0.92 for CTRL vs. AD, AUC = 0.83–0.86 for CTRL vs. MCI-*co*, and AUC = 0.71–0.72 for MCI-*nc* vs MCI-*co*; numbers that agree with our integrated implementation D within the CL.

To be clinically relevant, the use of repeated scans should improve on clinical group discrimination, and with respect to the baseline volume information.

Results indicate that we can get substantially more insight only using implementation D, which comes at the expense of a partial segmentation, that is one that does not deliver a tracing around the anatomical structure. This can be understood if we consider that in hippocampal segmentation literature near-boundary voxels are those who carry the burden of uncertainty (in our study, the threshold applied to the probabilistic map is the major source of error). Giving up the tracing we (re-)discover that the probabilistic map does carry a significant information.

If we compare the effect of the implementation on the longitudinal and baseline values while fixing the cohort comparison and feature (Table 3), we find evidence that the use of an intra-subject template (impl. B) is not enough to make the difference. The decisive approach is the unified segmentation, in either variant (C and D).

In clinical practice physicians are used to evaluate basal information on patient status, generate diagnostic hypothesis, plan treatment and then evaluate response in the longitudinal assessment. Moreover the trend observed in longitudinal assessment adds value to confirm or

Table 5
Sample size calculation.

Impl.	CTRL	MCI-*nc*	MCI-*co*	AD
A	267 (210–357)	268 (211–359)	88 (69–117)	85 (67–114)
B	153 (120–204)	169 (133–227)	77 (61–104)	101 (79–135)
C	91 (72–122)	103 (81–138)	58 (46–78)	42 (33–57)
D	25 (20–33)	45 (35–60)	33 (26–44)	33 (26–44)

Estimated sample sizes for both arms that would be needed to detect a 25% reduction in atrophy in all clinical cohorts and implementations. Numbers are given at fixed $\alpha = 0.05$ and for power $1 - \beta = 0.8$ (0.7 − 0.9).

put in discussion the original assumption. Theoretically, this longitudinal evaluation sounds more robust because intra-subject variance due to confounders is smaller than between-subject variance in cross-sectional data. Hence a longitudinal measure of hippocampal atrophy could in principle be more informative than a spot measure whenever taken during the patient history.

Translated into practice this would be similar to the advantage to have – for instance – serial MMSE scores during patient follow-up as a measure of disease worsening, but based on a solid neurodegeneration marker. The pathological basis of our assumption is the ongoing neurodegeneration process in MTL structures during the early stages of the disease leading to progressive atrophy that can be precisely detected by adequate MRI measures.

As closing remark, the shorter the follow-up time, the higher the need for sophisticated analysis tools. Probably a longer (say 5 years) period would allow simpler methods to detect significant changes, although that would void their need as the information would overlap with more direct and simpler approaches. Restricting the investigation to the time-varying hippocampal volume, it would be interesting to know whether this measure (on 2-y period and with 1.5 T images) has reached an upper limit in terms of added value. This could perhaps be challenged by a longitudinal extension to the harmonized hippocampal segmentation study.

Study limitations

We considered 1.5 T images only. Surely 3 T images could provide better contrast and potentially a more reliable segmentation (Chow et al., 2015). In practice though, this and other studies (Lötjönen et al., 2011; Macdonald et al., 2014) show that the advantages of 3 T images do not necessarily translate into a decidedly smaller variance in test/re-test conditions. Besides, clinical practice and still many trials must cope with 1.5 T scanners. These reasons would qualify the present study as delivering a lower bound, on which the use of better scanners and acquisition protocols should only improve.

In addition, the use of a preliminary release (100 out of the now available 135 labels) of the cross-sectional gold-standard tracings – without a longitudinal benchmark – did not provide a hint to the longitudinal performance achievable by a given algorithm. Perhaps a further evolution of the hippocampal protocol study could help in assessing new methods cross-sectional as well as longitudinal performance.

Another point arises from the use of the hippocampal volume and its derivative marker Λ, as they do not necessarily implement the most sensitive measure of early AD. For instance, more sophisticated approaches based on local geometry measures could be more informative (see Frankó et al. (2013)). Still, the volume is a rather straightforward and robust measure which more easily serves the purpose of confrontation among algorithms and studies. In addition, the hippocampal volume is now a widely accepted marker among clinicians.

We must also consider that the cohorts in this study consist of rather elderly subjects. It is conceivable that younger subjects (i.e., 40–60 y) exhibit smaller longitudinal variability than their elderly counterparts. In this case, the distinction between healthy controls and a population at risk could be made more substantial and a longitudinal marker would be instrumental. Further studies are needed on relatively young subjects.

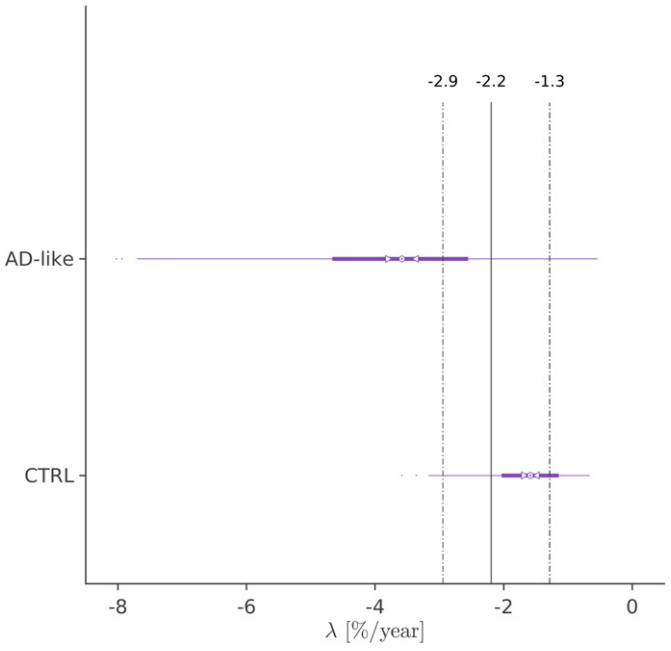

Fig. 6. Box plot of the bilateral average of the relative annualized loss λ on the CTRL and the 'AD-like' (AD + MCI-*co*) cohorts. Vertical lines shows three possible cut-off values: maximum accuracy (solid line), 95% sensitivity and 95% specificity (dashed lines). The median and its 95% conf. interval are marked with a dot and triangles on each bar.

Disclosure statement

All authors disclose any actual or potential conflicts of interest including any financial, personal or other relationships with other people or organizations that could inappropriately influence their work.

All experiments were performed with the informed consent of each participant or caregiver, in line with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Local institutional ethics committees approved the study.

Acknowledgments

This research was supported by Istituto Nazionale di Fisica Nucleare (INFN), Italy. This research was also directly supported by grants to FS from INFN (D.R.N. 1229/20.12.2013) and to LR from Università degli Studi Di Genova (n. 7801/29.04.2011).

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

Appendix A. Segmentation algorithm

The *GDIseg* algorithm is based on a set of manually traced segmentations by expert and certified readers from the HarP project. At the time

of this writing, 100 manual tracings were made available (58 1.5 T and 42 3 T)

Reference HarP images were processed as in Image processing section. In addition we extracted the VOIs from the manually segmented masks, using the same coordinates found for extracting the VOIs from the MRI.

We refer to the set of VOIs from the HarP MR images as Template Boxes (*TB*s) and the set of the corresponding segmented masks as Template Masks (*TM*s), both naturally coming with the right (*R*) and left (*L*) label. A pictorial overview of the segmentation process is shown in supplemental Fig. S1.

For each new segmentation, the MRI goes through the pre-process steps up to the extraction of both hippocampal VOIs (target VOIs). Subsequently, each *TB* is mapped onto the target VOI with a deformable registration transform, implemented in ITK with the "Diffeomorphic Demons" algorithm (Thirion (1998) and http://hdl.handle.net/1926/510). The resulting deformation field – one for each *TB* – is applied to the corrensponding *TM*.

At this point of the procedure, we have 100 deformed *TB*s (δTBs) and *TM*s (δTMs) to map the target VOI (*L* and *R* VOIs are run separately). Naturally, the more similar the original *TB* is to the target VOI, the lesser deformation it experiences and the more it ideally maps onto the target VOI.

A probabilistic atlas *A* is generated by weighted average of all deformed *TM*s, followed by a normalization. All VOIs, *TB*s, *TM*s and their deformed counterparts (δTBs, δTMs) have the same dimensions and number of voxels, so that we can write

$$A = \sum_{i=1}^{N_t} w_i \, \delta TM_i$$

where N_t is the number of templates.

In order to find the weights w_i, the *TB*s are ranked according to the Pearson correlation coefficient *r* with the target VOI. The correlation coefficient is not computed over the whole volume of the VOI, but on a subset of voxels corresponding to the volume surrounding the *TM*. The detailed procedure consists in three steps: a) dilation of the binary *TM* (distance of 3 *mm*), b) mapping of the dilated *TM* onto both the target VOI and the *TB* (voxel selection), c) computation of the correlation coefficient *r* between the intensities of both volumes over the selected voxels.

This procedure is applied to each *TB* using the related *TM* as initial mask to dilate. The dilation step is instrumental to capture the intensity gradient of the hippocampal borders, thereby ranking *TB*s according to their similarity to the target VOI more effectively. If we had used the whole VOI volume, the correlation coefficient would have been swayed by intensities coming from tissues unrelated to the hippocampus.

The correlation rank is used to compute the weights in the *TM*s average, under the hypothesis that it contains information on the "true segmentation". In this sense, correlation values are used as proxies for the segmentation similarity.

Since we do not know the target VOI true segmentation, we use a surrogate target δTB^* – that is the deformed *TB* with the best rank – in place of the target VOI, with the benefit that the true segmentation δTM^* is now available.

Weights are thought to be a simple exponential functions of the correlation coefficient, they are computed by minimizing the distance *m* over the free parameter *s* ($s \geq 0$)

$$m = \sum_{all\ voxels} \left(\delta TM^* - \frac{\sum_{i=1, i\neq i^*}^{N} w_i \, \delta TM_i}{\sum_{i=1, i\neq i^*}^{N} w_i} \right)^2$$

$$w_i = \left(\frac{r_i}{\max_i(r_i)} \right)^s$$

where *N* is the number of templates, i^* is the index of the surrogate target δTB^* and r_i are the correlation coefficients now computed between the surrogate target δTB^* and the *TB*s.

Once we find the optimal value of the parameter, we have a relationship between the correlation coefficients and the weights, which is then used to construct the probabilistic atlas.

The weight function optimizes the atlas generation by selecting *TB*s with a non-linear proportionality relationship. This step is necessary to the algorithm accuracy as a simple average (equal weights, $s = 0$) of the deformed masks typically results in smeared out atlas, not always able to capture the subtle anatomical and intensity differences in the target VOI.

The optimization is carried out for each target VOI, so that parameter values are adapted to the target. We found that the weight function w_i is usually rather steep ($s \gg 1$), that is only a small number of δTMs contribute to the probabilistic atlas.

The last step takes the probabilistic atlas *A* and applies a threshold *t* on its intensity values to convert it to a binary mask: $A_{(t)} = \{x_i \; such \; as \; A(x_i) \geq t\}$.

The optimal threshold is defined as

$$t^* = \max_j \left\{ \frac{1}{n} \sum_{x_i \in \partial A_{(t)}} [\nabla A(x_i)]^2 \right\}$$

where ∇A is the 3D-gradient of the atlas *A*, x_i is the $i - th$ voxel, $\partial A_{(t)}$ is the boundary of the thresholded atlas, *n* is the number of voxels x_i belonging to $\partial A_{(t)}$. That is, the optimal threshold is the intensity value t^* that maximizes the overlap of the thresholded atlas boundary onto the atlas squared gradient.

We have found that the maximization over the gradient gives superior performance – in terms of DICE index – compared to the simple intensity rule

$$t^* = \frac{1}{2} \max_{x_i} A(x_i)$$

The thresholded atlas naturally yields the hippocampal volume *v* which is used as base measure in this study.

The performance of the *GDIseg* procedure was tested on the same HarP dataset using a 20-fold cross-validation method (kfcv) and it was evaluated by three standard indexes: DICE (*Dc*, or F_1-score), Recall (*Rc*, or sensitivity) and Precision (*Pr*, or positive predictive value). Results are shown in supplemental table S2.

Since the 100 images from the HarP database consisted in 58 1.5 T and 42 3.0 T MRI, we show the performance by field strength, demonstrating that the segmentation algorithm is not affected by the B-field intensity.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.10.065.

References

Ahn, B.J., Yim, D.S., 2009. Comparison of parametric and bootstrap method in bioequivalence test. Korean J. Physiol. Pharmacol. 13, 367. http://dx.doi.org/10.4196/kjpp.2009.13.5.367 (URL: http://synapse.koreamed.org/DOIx.php?id=10.4196/kjpp.2009.13.5.367).

Apostolova, L.G., Thompson, P.M., Green, A.E., Hwang, K.S., Zoumalan, C., Jack, C.R., Harvey, D.J., Petersen, R.C., Thal, L.J., Aisen, P.S., Toga, A.W., Cummings, J.L., DeCarli, C.S., 2010. 3D comparison of low, intermediate, and advanced hippocampal atrophy in MCI. Hum. Brain Mapp. 31, 786–797. http://dx.doi.org/10.1002/hbm.20905 (URL: http://doi.wiley.com/10.1002/hbm.20905).

Apostolova, L.G., Zarow, C., Biado, K., Hurtz, S., Boccardi, M., Somme, J., Honarpisheh, H., Blanken, A.E., Brook, J., Tung, S., Lo, D., Ng, D., Alger, J.R., Vinters, H.V., Bocchetta, M., Duvernoy, H., Jack, C.R., Frisoni, G.B., 2015. Relationship between hippocampal atrophy and neuropathology markers: a 7 T MRI validation study of the EADC-ADNI Harmonized Hippocampal Segmentation Protocol. Alzheimers Dement. 11, 139–150 (URL: http://www.ncbi.nlm.nih.gov/pubmed/25620800).

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, 1711–1723. http://dx.doi.org/10.1016/j.neurobiolaging.2008.01.010 (URL: http://linkinghub.elsevier.com/retrieve/pii/S0197458008000262).

Bateman, R.J., Xiong, C., Benzinger, T.L.S., Fagan, A.M., Goate, A., Fox, N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., Holtzman, D.M., Santacruz, A., Buckles, V., Oliver, A., Moulder, K., Aisen, P.S., Ghetti, B., Klunk, W.E., McDade, E., Martins, R.N., Masters, C.L., Mayeux, R., Ringman, J.M., Rossor, M.N., Schofield, P.R., Sperling, R.a., Salloway, S., Morris, J.C., 2012. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N. Engl. J. Med. 367, 795–804. http://dx.doi.org/10.1056/NEJMoa1202753 (URL: http://www.ncbi.nlm.nih.gov/pubmed/22784036).

Bishop, C.A., Jenkinson, M., Andersson, J., Declerck, J., Merhof, D., 2011. Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH): method and validation on clinical data. NeuroImage 55, 1009–1019 (URL: http://www.ncbi.nlm.nih.gov/pubmed/21195778).

Boccardi, M., Bocchetta, M., Morency, F.C., Collins, D.L., Nishikawa, M., Ganzola, R., Grothe, M.J., Wolf, D., Redolfi, A., Pievani, M., Antelmi, L., Fellgiebel, A., Matsuda, H., Teipel, S., Duchesne, S., Jack, C.R., Frisoni, G.B., 2015. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimers Dement. 11, 175–183. http://dx.doi.org/10.1016/j.jalz.2014.12.002 (URL: http://linkinghub.elsevier.com/retrieve/pii/S155252601402891X http://www.ncbi.nlm.nih.gov/pubmed/25616957).

Bocchetta, M., Boccardi, M., Ganzola, R., Apostolova, L.G., Preboske, G., Wolf, D., Ferrari, C., Pasqualetti, P., Robitaille, N., Duchesne, S., Jack, C.R., Frisoni, G.B., 2014. Harmonized benchmark labels of the hippocampus on magnetic resonance: the EADC-ADNI project. Alzheimers Dement. http://dx.doi.org/10.1016/j.jalz.2013.12.019 (URL: http://linkinghub.elsevier.com/retrieve/pii/S1552526014000107).

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., 2011. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. NeuroImage 58, 469–480. http://dx.doi.org/10.1016/j.neuroimage.2011.05.083 (URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-80051784240&partnerID=tZOtx3y1 http://www.ncbi.nlm.nih.gov/pubmed/21718788 http://apps.isiknowledge.com/full_record.do?product=UA&search_mode=Refine&qid=5&SID=Q25xjKGZbUAEd14dpWk&page=1&doc=44).

Chow, N., Hwang, K.S., Hurtz, S., Green, A.E., Somme, J.H., Thompson, P.M., Elashoff, D.A., Jack, C.R., Weiner, M., Apostolova, L.G., 2015. Comparing 3 T and 1.5 T MRI for mapping hippocampal atrophy in the Alzheimer's Disease neuroimaging initiative. Am. J. Neuroradiol. 36 (4), 653–660. http://dx.doi.org/10.3174/ajnr.A4228 (URL: http://www.ncbi.nlm.nih.gov/pubmed/25614473).

Cortes, C., Mohri, M., 2004. Confidence Intervals for the Area Under the ROC Curve. Nips (URL: https://papers.nips.cc/paper/2645-confidence-intervals-for-the-area-under-the-roc-curve.pdf).

Duning, T., Kloska, S., Steinsträter, O., Kugel, H., Heindel, W., Knecht, S., 2005. Dehydration confounds the assessment of brain atrophy. Neurology 64, 548–550. http://dx.doi.org/10.1212/01.WNL.0000150542.16969.CC.

Fennema-Notestine, C., Hagler, D.J., McEvoy, L.K., Fleisher, A.S., Wu, E.H., Karow, D.S., Dale, A.M., 2009. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. Hum. Brain Mapp. 30, 3238–3253 (URL: http://www.ncbi.nlm.nih.gov/pubmed/19277975).

Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. Lancet 363, 392–394 (URL: http://www.ncbi.nlm.nih.gov/pubmed/15074306).

Frankó, E., Joly, Olivier, A.D.N.I., 2013. Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. PLoS One 8, e71354. http://dx.doi.org/10.1371/journal.pone.0071354 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3741167&tool=pmcentrez&rendertype=abstract).

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. http://dx.doi.org/10.1016/j.neuroimage.2015.03.035 (URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811915002177).

Frisoni, G.B., Jack, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu, Y., Preboske, G., Swihart, T., Blair, M., Cavedo, E., Grothe, M.J., Lanfredi, M., Martinez, O., Nishikawa, M., Portegies, M., Stoub, T., Ward, C., Apostolova, L.G., Ganzola, R., Wolf, D., Barkhof, F., Bartzokis, G., DeCarli, C., Csernansky, J.G., DeToledo-Morrell, L., Geerlings, M.I., Kaye, J., Killiany, R.J., Lehéricy, S., Matsuda, H., O'Brien, J., Silbert, L.C., Scheltens, P., Soininen, H., Teipel, S., Waldemar, G., Fellgiebel, A., Barnes, J., Firbank, M., Gerritsen, L., Henneman, W., Malykhin, N., Pruessner, J.C., Wang, L., Watson, C., Wolf, H., DeLeon, M., Pantel, J., Ferrari, C., Bosco, P., Pasqualetti, P., Duchesne, S., Duvernoy, H., Boccardi, M., 2014. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement. http://dx.doi.org/10.1016/j.jalz.2014.05.1756 (URL: http://linkinghub.elsevier.com/retrieve/pii/S1552526014024686).

Gengsheng Qin, Hotilovac, L., 2007. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat. Methods Med. Res. 17, 207–221. http://dx.doi.org/10.1177/0962280207087173 (URL: http://smm.sagepub.com/cgi/doi/10.1177/0962280207087173).

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehéricy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47, 1476–1486. http://dx.doi.org/10.1016/j.neuroimage.2009.05.036 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3001345&tool=pmcentrez&rendertype=abstract).

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36 (URL: http://www.ncbi.nlm.nih.gov/pubmed/7063747).

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843 (URL: http://www.ncbi.nlm.nih.gov/pubmed/6878708).

Hogan, R.E., Wang, L., Bertrand, M.E., Willmore, L.J., Bucholz, R.D., Nassif, A.S., Csernansky, J.G., 2004. MRI-based high-dimensional hippocampal mapping in mesial temporal lobe epilepsy. Brain J. Neurol. 127, 1731–1740. http://dx.doi.org/10.1093/brain/awh197 (URL: http://www.ncbi.nlm.nih.gov/pubmed/15231583).

Hovorkova, P., Kristofikova, Z., Horinek, A., Ripova, D., Majer, E., Zach, P., Sellinger, P., Ricny, J., 2008. Lateralization of 17beta-hydroxysteroid dehydrogenase type 10 in hippocampi of demented and psychotic people. Dement. Geriatr. Cogn. Disord. 26, 193–198 (URL: http://www.ncbi.nlm.nih.gov/pubmed/18765932).

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J.L., Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27, 685–691. http://dx.doi.org/10.1002/jmri.21049 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2544629&tool=pmcentrez&rendertype=abstract).

Jovicich, J., Marizzoni, M., Sala-Llonch, R., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F., Hensch, T., Tränkner, A., Schönknecht, P., Leroy, M., Lopes, R., Bordet, R., Chanoine, V., Ranjeva, J.P., Didic, M., Gros-Dagnac, H., Payoux, P., Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalló, N., Blin, O., Frisoni, G.B., 2013. Brain morphometry reproducibility in multi-center 3 T MRI studies: a comparison of cross-sectional and longitudinal segmentations. NeuroImage 83, 472–484. http://dx.doi.org/10.1016/j.neuroimage.2013.05.007 (URL: http://linkinghub.elsevier.com/retrieve/pii/S1053811913004977).

Kottas, M., Kuss, O., Zapf, A., 2014. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. BMC Med. Res. Methodol. 14, 26. http://dx.doi.org/10.1186/1471-2288-14-26 (URL: http://www.biomedcentral.com/1471-2288/14/26).

Leung, K.K.K., Barnes, J., Ridgway, G.G.R.G., Bartlett, J.W., Clarkson, M.M.J., Macdonald, K., Schuff, N., Fox, N.N.C.N., Ourselin, S., Barlett, J., 2010. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. NeuroImage 51, 1345–1359. http://dx.doi.org/10.1016/j.neuroimage.2010.03.018.Automated.

Lobanova, I., Sohail, A., Adil, M., Saed, A., Qureshi, A., 2014. Progressive hippocampus atrophy in persons with Alzheimer's disease: a longitudinal MRI study. Neurology 82, P6.333 (URL: http://www.neurology.org/content/82/10_Supplement/P6.333).

Lötjönen, J., Wolz, R., Koikkalainen, J., Julkunen, V., Thurfjell, L., Lundqvist, R., Waldemar, G., Soininen, H., Rueckert, D., 2011. Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. NeuroImage 56, 185–196. http://dx.doi.org/10.1016/j.neuroimage.2011.01.062 (URL: http://www.ncbi.nlm.nih.gov/pubmed/21281717).

Macdonald, K.E., Leung, K.K., Bartlett, J.W., Blair, M., Malone, I.B., Barnes, J., Ourselin, S., Fox, N.C., 2014. Automated template-based hippocampal segmentations from MRI: the effects of 1.5 T or 3 T field strength on accuracy. Neuroinformatics 12, 405–412 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4104163&tool=pmcentrez&rendertype=abstract).

Maclaren, J., Han, Z., Vos, S.B., Fischbein, N., Bammer, R., 2014. Reliability of brain volume measurements: a test–retest dataset. Nat. Sci. Data 1–9 http://dx.doi.org/10.1038/sdata.2014.37.

Martinez, W.L., 2011. Computational Statistics in MATLAB®. Wiley Interdiscip. Rev. Comput. Stat. 3, 69–74.

mazziotta, Toga, Evans, Fox, L., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. NeuroImage 2, 89–101.

McEvoy, L.K., Holland, D., Hagler, D.J., Fennema-Notestine, C., Brewer, J.B., Dale, A.M., 2011. Mild cognitive impairment: baseline and longitudinal structural MR imaging measures improve predictive prognosis. Radiology 259, 834–843. http://dx.doi.org/10.1148/radiol.11101975 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3099042&tool=pmcentrez&rendertype=abstract).

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Madsen, S.K., Parikshak, N., Hua, X., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M., 2008. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. NeuroImage 43, 59–68. http://dx.doi.org/10.1016/j.neuroimage.2008.07.003 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2624575&tool=pmcentrez&rendertype=abstract).

Morra, J., Tu, Z., Apostolova, L., Green, A., Avedissian, C., Madsen, S., Parikshak, N., Toga, A., Jack, C., Schuff, N., 2009. Automated hippocampal segmentation and mapping reveals genetically accelerated tissue loss in 1-year repeat MRI data from 490 Alzheimer's disease, MCI, and control subjects. NeuroImage 47. http://dx.doi.org/10.1016/S1053-8119(09)71153-X (S122–S122, URL: http://linkinghub.elsevier.com/retrieve/pii/S105381190971153X).

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M., 2010. Comparison of adaboost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE Trans. Med. Imaging 29, 30–43. http://dx.doi.org/10.1109/TMI.2009.2021941.

Mulder, E.R., de Jong, R.A., Knol, D.L., van Schijndel, R.A., Cover, K.S., Visser, P.J., Barkhof, F., Vrenken, H., Alzheimer's Disease Neuroimaging Initiative, 2014. Hippocampal volume change measurement: quantitative assessment of the reproducibility of expert

manual outlining and the automated methods FreeSurfer and FIRST. Neuroimage 92, 169–181. http://dx.doi.org/10.1016/j.neuroimage.2014.01.058.

Okonkwo, O.C., Xu, G., Dowling, N.M., Bendlin, B.B., LaRue, A., Hermann, B.P., Koscik, R., Jonaitis, E., Rowley, H.a., Carlsson, C.M., Asthana, S., Sager, M.a., Johnson, S.C., 2012. Family history of Alzheimer disease predicts hippocampal atrophy in healthy middle-aged adults. Neurology 78, 1769–1776. http://dx.doi.org/10.1212/WNL.0b013e3182583047 (URL: http://www.neurology.org/cgi/doi/10.1212/WNL.0b013e3182583047).

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56, 907–922 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3417233&tool=pmcentrez&rendertype=abstract).

Picco, A., Polidori, M.C., Ferrara, M., Cecchetti, R., Arnaldi, D., Baglioni, M., Morbelli, S., Bastiani, P., Bossert, I., Fiorucci, G., Brugnolo, A., Dottorini, M.E., Nobili, F., Mecocci, P., 2014. Plasma antioxidants and brain glucose metabolism in elderly subjects with cognitive complaints. Eur. J. Nucl. Med. Mol. Imaging 41, 764–775 (URL: http://www.ncbi.nlm.nih.gov/pubmed/24297504).

Pruessner, J.C., Li, L.M., Serles, W., Pruessner, M., Collins, D.L., Kabani, N., Lupien, S., Evans, A.C., 2000. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb. Cortex 10, 433–442 (URL: http://www.ncbi.nlm.nih.gov/pubmed/10769253).

Reig, S., Sánchez-González, J., Arango, C., Castro, J., González-Pinto, A., Ortuño, F., Crespo-Facorro, B., Bargalló, N., Desco, M., 2009. Assessment of the increase in variability when combining volumetric data from different scanners. Hum. Brain Mapp. 30, 355–368 (URL: http://www.ncbi.nlm.nih.gov/pubmed/18064586).

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012a. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418 (URL: http://www.sciencedirect.com/science/article/pii/S1053811912002765).

Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012b. Within-subject template estimation for unbiased longitudinal image analysis. NeuroImage 61, 1402–1418 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3389460&tool=pmcentrez&rendertype=abstract).

Rusinek, H., De Santi, S., Frid, D., Tsui, W.H., Tarshish, C.Y., Convit, A., de Leon, M.J., 2003. Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging. Radiology 229, 691–696. http://dx.doi.org/10.1148/radiol.2293021299.

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L.M., Trojanowski, J.Q., Thompson, P.M., Jack, C.R., Weiner, M.W., 2009. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain 132, 1067–1077. http://dx.doi.org/10.1093/brain/awp007 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2668943&tool=pmcentrez&rendertype=abstract).

Shen, D., Moffat, S., Resnick, S.M., Davatzikos, C., 2002. Measuring size and shape of the hippocampus in MR images using a deformable shape model. NeuroImage 15, 422–434. http://dx.doi.org/10.1006/nimg.2001.0987 (URL: http://www.ncbi.nlm.nih.gov/pubmed/11798276).

Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: meta-analyses of MRI studies. Hippocampus 19, 1055–1064 (URL: http://www.ncbi.nlm.nih.gov/pubmed/19309039).

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155. http://dx.doi.org/10.1002/hbm.10062 (URL: http://www.ncbi.nlm.nih.gov/pubmed/12391568).

Spulber, G., Simmons, A., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Spenger, C., Lovestone, S., Wahlund, L.O., Westman, E., 2013. An MRI-based index to measure the severity of Alzheimer's disease-like structural pattern in subjects with mild cognitive impairment. J. Intern. Med. 273, 396–409. http://dx.doi.org/10.1111/joim.12028 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605230&tool=pmcentrez&rendertype=abstract).

Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with Maxwell's demons. Med. Image Anal. 2, 243–260 (URL: http://www.ncbi.nlm.nih.gov/pubmed/9873902).

Thompson, W.K., Holland, D., 2011. Bias in tensor based morphometry Stat-ROI measures may result in unrealistic power estimates. NeuroImage 57, 1–4 (URL: http://www.sciencedirect.com/science/article/pii/S1053811911001492).

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., 2010a. LEAP: learning embeddings for atlas propagation. NeuroImage 49, 1316–1325. http://dx.doi.org/10.1016/j.neuroimage.2009.09.069 (URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3068618&tool=pmcentrez&rendertype=abstract).

Wolz, R., Heckemann, R.a., Aljabar, P., Hajnal, J.V., Hammers, A., Lötjönen, J., Rueckert, D., 2010b. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. NeuroImage 52, 109–118. http://dx.doi.org/10.1016/j.neuroimage.2010.04.006 (URL: http://www.ncbi.nlm.nih.gov/pubmed/20382238).

Wolz, R., Schwarz, A.J., Yu, P., Cole, P.E., Rueckert, D., Jack, C.R., Raunig, D., Hill, D., 2014. Robustness of automated hippocampal volumetry across magnetic resonance field strengths and repeat images. Alzheimers Dement. 10, 430–438. http://dx.doi.org/10.1016/j.jalz.2013.09.014 (URL: http://www.ncbi.nlm.nih.gov/pubmed/24985688).