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, and for the Alzheimer's Disease Neuroimaging Initiative¹

^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, Genova, Italy*

^k*University Hospitals and University of Geneva, Geneva, Switzerland*

Keywords:

MRI, Image analysis, Longitudinal measure, Alzheimer's disease,

Hippocampus

PACS: 87.61.-c, 87.57.N-, 87.61.Tg

*Corresponding author. Address: INFN, via Dodecaneso 33, I-16146 Genova, Italy. Tel.: +39 010 353 6496; fax: +39 010 313358.

Email address: andrea.chincarini@ge.infn.it (Andrea Chincarini)

¹Data used in preparation of this article were obtained from the Alzheimers 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

Preprint submitted to Neuroimage

September 28, 2015

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 an 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 (1.5T, 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;

1. 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 and Joly, Olivier (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, artefacts,

25 data analysis and algorithmic instabilities) and physiological variability (both
26 intrinsic and due to external conditions such as hydration, lipidic balance,
27 nutrition and hormonal concentration, Duning et al. (2005); Maclaren et al.
28 (2014)). The goal of longitudinal analysis though is to find the long-term
29 trend due to either normal or pathological ageing, neglecting the nuisances
30 of both intrinsic and **extrinsic** variabilities.

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

41 **2. Materials and methods**

42 *2.1. Dataset*

43 MRI scans (1.5T, T1-weighted) were selected from the ADNI database ²
44 and downloaded in the original format (DICOM). The subjects id list is
45 provided in supplemental table S1.

²The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. For up-to-date information, see www.adni-info.org.

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

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

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

59 On average, time to AD occurred after 48 (24 – 84) months (90% confi-
60 dence bounds) from the baseline.

61 2.2. Image processing

62 Image processing closely follows the method detailed in Chincarini et al.
63 (2011), save for two procedural differences. We summarize here the main
64 steps applied to each MR image up to the extraction of its Volumes of Interest
65 (VOI), which were used as starting points of the segmentation algorithm.

66 MR images underwent a series of filters designed for bias B-field reduc-
67 tion, volume normalization, anatomical structure registration and gray level
68 intensity equalization. The two novelties with respect to Chincarini et al.
69 (2011) are the lack of the pyramidal noise filter and the addition of the B-

70 field bias reduction, the latter implemented with the BET algorithm (Smith,
71 2002). The noise filtering step was avoided to keep the intensity contrast
72 between the hippocampus structure and the adjacent structures (amigdala
73 mainly), which could be impaired by the pyramidal filter. Similarly, the B-
74 field bias correction was introduced to improve on the deformable registration
75 cost function used in the segmentation process.

76 As result of the pre-processing steps, images were aligned with a 12-pa-
77 rameters affine transformation to the Montreal Neurological Institute (MNI,
78 mazziotta, Toga, Evans, Fox (1995)) space and the mean gray level intensi-
79 ties of the three major brain constituents - cerebro-spinal fluid (CSF), gray
80 matter (GM) and white matter (WM) - were matched to reference values.
81 In addition, aligned images are spatially sampled as the MNI template, that
82 is with isotropic voxels of $1mm$.

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

87 Finally, a finer intra-cranial volume correction (icv) is computed by non-
88 linear mapping of the segmented MNI brain mask (provided with the tem-
89 plate) onto the affine-registered image and the mask volume is weighted by
90 the affine transformation jacobian. This number is a minor factor (of the
91 order of the unity) and it does not correct for the native volume versus the
92 MNI-space one, as the spatial normalization already compensated for it. It
93 is rather used to adjust for the possible deviations that escape the affine reg-
94 istration. This non-linear-based intra-cranial volume adjustment is used as

95 a hippocampal volume correction factor after the segmentation process.

96 2.3. Segmentation algorithm

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

100 The procedure (referred in the following as *GDIseg*) requires only the
101 hippocampal VOI in input and it is not too dissimilar from that proposed
102 by Wolz et al. (2010a), save for some details. It was developed on a MR
103 set consisting of 100 T1-weighted MR images and tracings (Frisoni et al.
104 (2014), preliminary release) from the “harmonized protocol for hippocampal
105 volumetry” project (HarP, www.hippocampal-protocol.net), subjects not
106 included in the dataset presented in this investigation.

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

112 2.4. Implementations

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

118 All descriptions regarding the hippocampal VOIs have no explicit later-
119 ality labels but it is intended that they are run on the left and right VOI
120 separately.

121 *2.4.1. A: independent processing and segmentation*

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

126 *2.4.2. B: unified image processing*

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

130 The within-subject template is constructed by generating an average in-
131 termediate image H from the 4 scans (baseline, repeat, month 12 and month
132 24) using robust, inverse consistent registration (Reuter et al., 2012a). The
133 intermediate within-subject template is processed up to the extraction of
134 the hippocampal VOIs according to section 2.2. The relevant parameters
135 (registration onto the MNI reference, VOI positions and intensity normaliza-
136 tion) are passed back to the original scans so that the actual VOIs can be
137 extracted.

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

141 2.4.3. *C: atlas matrix re-normalization*

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

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

151 We can define the 4×4 matrix $\mathbf{f}$ whose elements are the f_{ij} and which
 152 contains the identity transformation I on the diagonal, with the requirement
 153 that $f_{ij} + f_{ji} = I$. Similarly, we can define a matrix $\mathbf{a}$ of probabilistic maps
 154 whose elements are $a_{ij} = f_{ij}(A_i)$, i.e. the application of the field f_{ij} to A_i .
 155 By definition, the diagonal elements are $a_{ii} = A_i$. Addition, subtraction and
 156 multiplication by a constant on the deformation field f are intended to be
 157 applied voxel-by-voxel to the displacement vector components. The identity
 158 operator I components are by definition all zero.

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

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

163 A possible choice for α_{ij} could be

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

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

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

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

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

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

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

176 2.4.4. *D: weighted integration*

177 In this last implementation images are preprocessed as in “B” and seg-
 178 mentation undergoes a post-processing step, this time though we drop the
 179 requirement of an actual binary mask per VOI, in favour of the volume in-
 180 formation alone (figure 2 D).

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

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

182

$$A_m = \max_j A_j$$

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

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

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

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

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

190 In short, this formulation modulates the intensities in the bigger map
 191 (A_m , which includes the hippocampal boundary) with the inner intensity

192 values (A_p , where all segmentations agree).

193 2.5. Performance metrics

194 We checked the performance of all described procedures with four metrics.
195 The first one (reliability) is simply a quality control to assess the robustness
196 of *GDIseg* on a large number of images. Then we looked at the test/re-
197 test performance (reproducibility) and at the longitudinal trend. Finally we
198 checked whether the longitudinal information can improve on the accuracy
199 when used as combined biomarker together with the volume.

200 2.5.1. Reliability

201 The segmentation procedure was applied without human intervention to
202 1840 images from the ADNI database. A quality control test checks whether
203 and on how many images the procedure crudely fails. This control does
204 not imply a “correct” hippocampus segmentation - in terms of harmonized
205 protocol - it only points out possible failures in the pre-processing and in
206 the segmentation procedure. To perform this test we construct two identical
207 statistics Re_{voi} and Re_{mask} :

$$Re_{voi} = \min_{t,L,R} \{ \max_i [r(VOI, TB_i)] \}$$

208

$$Re_{mask} = \min_{t,L,R} \{ \max_j [r(mask, TM_j)] \}$$

209 where r is the Pearson correlation coefficient, the ‘max’ is taken on the tem-
210 plates and the ‘min’ is taken among scans (t) and laterality (L, R). Template
211 Boxes (TB) and Template Masks (TM) are the hippocampal VOIs and man-
212 ual tracings on the HarP image dataset (see Appendix A).

213 This test computes the best correlation coefficient among the VOI inten-
 214 sities and each TB , as well as among the segmented mask and each TM ,
 215 then keeping the lowest among these values with respect to the number of
 216 scans and laterality. In other words, from each subject we get 8 VOIs and
 217 8 hippocampal tracings (bilateral regions on 4 scans). If either one or more
 218 are too distant from its nearest template (in terms of correlation coefficient),
 219 the subject is flagged for visual inspection. This formulation assumes that
 220 the HarP subjects are sampled as to represent all relevant physiological vari-
 221 ability.

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

228 *2.5.2. Reproducibility*

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

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

235 where v_b and v_r are the baseline and repeat hippocampal volumes respec-
236 tively.

237 2.5.3. Longitudinal trend

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

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

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

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

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

250 2.5.4. Combined markers

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

254 ROC analysis on the combined volume and trend indexes was computed
255 with a linear discriminant. We used a support vector machine (SVM) classi-
256 fier on the feature set $(\hat{v}, \Lambda)$ and we considered the distance from the sepa-

rating plane as the new marker. Its performance was compared to that of $\hat{v}$ and Λ alone.

2.6. 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 i th 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.

279 Confidence intervals on AUC values in table 3 were computed by boot-
 280 strapping (1000 times) and by using the bias-corrected percentile method
 281 (Martinez, 2011). Statistical significance in table 4 versus the null AUC and
 282 among different markers was carried on according to Hanley and McNeil
 283 (1982, 1983).

284 The estimation of confidence intervals on the AUC can be carried out with
 285 several methods, each delivering slightly different values. Hence the compari-
 286 son and compatibility among tests in table 3 and 4 should take into consider-
 287 ation that confidence intervals are method-dependent estimates. We consid-
 288 ered seven methods, parametric and non-parametric: Hanley-McNeil (para-
 289 metric); Mann-Whitney, Logit and Bootstrap (non-parametric, Gengsheng
 290 Qin and Hotilovac (2007)); Maximum variance (non-parametric, Cortes and
 291 Mohri (2004)); Wald, Wald/continuity-corrected (non-parametric, Kottas
 292 et al. (2014)).

293 For instance, the width of the confidence interval on $\hat{v}_L$ for the CTRL/AD
 294 cohorts (implementation D, AUC=0.89 in table 3) ranges from 0.06 (Hanley-
 295 McNeil) to 0.09 (Mann-Whitney); in numbers 0.86–0.92 and 0.84–0.93. An-
 296 other example with Λ_R , implementation C and CTRL/MCI-*co* (AUC=0.78)
 297 shows a substantially similar interval width of all methods (0.74 – 0.82 Han-
 298 ley, Mann-Whitney; 0.73 – 0.84 Maximum Variance). The bias-corrected
 299 percentile bootstrap was regarded as a safe estimate as it did not require any
 300 assumption about the normality of the log-transformed AUC (Ahn and Yim,
 301 2009).

302 3. Results

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

306 3.1. Quality control

307 Figure 3 shows the distribution of Re_{voi} and Re_{mask} for all 460 subjects.
308 There are three distinctive outliers which are excluded from subsequent anal-
309 yses and whose inconsistent VOIs and tracings are shown aside (fig. 3a, b
310 and c). Potential outliers - placed in the low value regions of the Re_{voi} /
311 Re_{mask} scatter plot - are visually screened to ensure that they are correctly
312 classified as proper VOI and hippocampal tracings.

313 One of the outliers (figure 3a) stems from a blind injection: a null image
314 (white noise only) was placed in the analysis pipeline on purpose, in order
315 to test the reliability of the whole analysis procedure. Another outlier (fig.
316 3b) is due to incorrect brain spatial registration, causing the VOIs to be
317 misplaced. The third one (fig. 3c) is due to the peculiar atrophy conditions,
318 which has no related template in the HarP subject selection.

319 3.2. Reproducibility

320 The relative volume variation over baseline and repeat scan is given for
321 the A, B, C and D implementations in percent units (%), mean and standard
322 deviation: $\Delta_A = -0.1 \pm 3.5$, $\Delta_B = -0.1 \pm 2.7$, $\Delta_C = 0.0 \pm 0.1$ and $\Delta_D =$
323 0.1 ± 1.2 . The absolute value of the standard deviation σ_v over the quantity
324 $v_r - v_b$ is: $\sigma_v^A = 156$, $\sigma_v^B = 128$, $\sigma_v^C = 5$ and $\sigma_v^D = 68$ (units in mm^3).

3.3. 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, figure 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 $\rightarrow$ 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

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

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 figure 6.

3.4. 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 figure 5. In each comparison, we marked whether the combined information fared significantly better than either factors. Considering a total of 3 (group comparisons) $\times$ 4 (implementations) $\times$ 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.

3.5. 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

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

we chose $\delta = 0.25\bar{\Lambda}$ where $\bar{\Lambda} = (\Lambda_R + \Lambda_L)/2$ 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.

4. 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

394 check.

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

402 To further reduce the atrophy rate uncertainties we could have used sev-
403 eral more time points. This however would be an impractical protocol to
404 implement outside clinical trials. Similarly, using two time points only (i.e.
405 0 – 12m) would result in a larger error and a lower discrimination power
406 (Wolz et al., 2010b).

407 4.1. *Quality control*

408 All procedures need a stable segmentation, which in turns depends on
409 an accurate VOI placing. Segmentation accuracy with respect to the expert
410 tracing is comparable to results in literature: the LEAP method (Wolz et al.,
411 2010a) DICE index $\simeq 0.85$; adaboost, ada-SVM and Freesurfer (Morra et al.,
412 2010) Precision $\simeq 0.71 - 0.84$, Recall $\simeq 0.73 - 0.87$; and in Lötjönen et al.
413 (2011) DICE index $\simeq 0.87$.

414 In this study the supplemental Re_{voi} and Re_{mask} statistics are used as
415 warning indicators of outliers as they compare a new VOI and related seg-
416 mentation with the reference templates. If the templates do not sample the
417 population extensively enough we may incur in extreme statistic values. In

the particular example shown in figure 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 favourably impact the finding of true outliers.

In the case of the purely noisy image (blank test) of figure 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).

4.2. 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} .

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

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

449 Comparison to literature shows that results similar to the basic imple-
450 mentations A and B are obtained in Maclaren et al. (2014) (with a total
451 coefficient of variation of $\simeq 3\%$ on the hippocampus and using Freesurfer).

452

453 *4.3. Further methodological considerations*

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

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

466 ability, particularly in single-modality (T1) segmentations.
 467 We looked for subjects whose record showed the use of different MRI ma-
 468 chines. A survey of the CTRL cohort indicated that 42 out of 148 subjects
 469 ($\simeq 28\%$) were acquired with different scanner models at some follow-up visit
 470 (with respect to the MRI system used at baseline).
 471 The potential added variability was gauged with a direct comparison of the
 472 statistics using the non-parametric Kolmogorov-Smirnov test. The applica-
 473 tion to the sample of 106 CTRL (same scanner model across longitudinal
 474 measures but different cross-sectionally) and 42 CTRL subjects (different
 475 scanner model both in longitudinal measures and cross-sectionally) found no
 476 significant difference the Λ statistics, regardless of the implementation.
 477 Nonetheless, the use of different models in the longitudinal acquisition could
 478 show up in the linear fit residuals ξ (cfr. section 2.5.3). Indeed, testing
 479 the ξ distributions revealed a significant alteration in implementation A only
 480 ($p < 0.001$), which would suggest that the adoption of an intra-subject tem-
 481 plate (used in B, C and D) is sufficient to tame the inter-scanner repro-
 482 ducibility uncertainty. This finding agrees with Jovicich et al. (2013), where
 483 the introduction of longitudinal methods for volumes extraction provides a
 484 lower and more homogeneous reproducibility error across different scanners.
 485 Another point is the role of laterality. In this study we treated left and
 486 right hippocampi equally and separately to avoid any laterality bias in the
 487 results.
 488 The significance of a performance superiority of the left side was investigated
 489 by comparing the R and L AUC values with a t-test, regardless of the im-

plementation 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 $\text{AUC}_L - \text{AUC}_R$ 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 signifi-
cant 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 promi-
nent 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 se-
vere than the right one (Shi et al., 2009) and with Frankó and Joly, Olivier
(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 corre-
late with glucose metabolism of the left temporal lobe – including medial
structures – but not of the right one (Picco et al., 2014). Also, the multi-
functional mitochondrial enzyme 17β -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 $\hat{v}$ 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 figure 6.

4.4. 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 ageing 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

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

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

552 To be clinically relevant, the use of repeated scans should improve on
553 clinical group discrimination, and with respect to the baseline volume infor-
554 mation.

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

563 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 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

588 approaches. Restricting the investigation to the time-varying hippocampal
589 volume, it would be interesting to know whether this measure (on 2-y pe-
590 riod and with 1.5T images) has reached an upper limit in terms of added
591 value. This could perhaps be challenged by a longitudinal extension to the
592 harmonized hippocampal segmentation study.

593 *4.5. Study limitations*

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

602 In addition, the use of a preliminary release (100 out of the now available
603 135 labels) of the cross-sectional gold-standard tracings – without a longitu-
604 dinal benchmark – did not provide a hint to the longitudinal performance
605 achievable by a given algorithm. Perhaps a further evolution of the hip-
606 pocampal protocol study could help in assessing new methods cross-sectional
607 as well as longitudinal performance.

608 Another point arises from the use of the hippocampal volume and its
609 derivative marker Λ , as they do not necessarily implement the most sensi-
610 tive measure of early AD. For instance, more sophisticated approaches based
611 on local geometry measures could be more informative (see Frankó and Joly,

612 Olivier (2013)). Still, the volume is a rather straightforward and robust mea-
613 sure which more easily serves the purpose of confrontation among algorithms
614 and studies. In addition, the hippocampal volume is now a widely accepted
615 marker among clinicians.

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

622 **5. Disclosure statement**

623 All authors disclose any actual or potential conflicts of interest including
624 any financial, personal or other relationships with other people or organiza-
625 tions that could inappropriately influence their work.

626 All experiments were performed with the informed consent of each par-
627 ticipant or caregiver, in line with the Code of Ethics of the World Medical
628 Association (Declaration of Helsinki). Local institutional ethics committees
629 approved the study.

630 **6. Acknowledgements**

631 This research was supported by Istituto Nazionale di Fisica Nucleare
632 (INFN), Italy. This research was also directly supported by grants to FS
633 from INFN and to LR from Università degli Studi Di Genova.

634 Data collection and sharing for this project was funded by the Alzheimer's
635 Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant
636 U01 AG024904). ADNI is funded by the National Institute on Aging, the
637 National Institute of Biomedical Imaging and Bioengineering, and through
638 generous contributions from the following: Abbott; Alzheimer's Associa-
639 tion; Alzheimers Drug Discovery Foundation; Amorfix Life Sciences Ltd.;
640 AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-
641 Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and
642 Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech,
643 Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer
644 Immunotherapy Research & Development, LLC.; Johnson & Johnson Phar-
645 maceutical Research & Development LLC.; Medpace, Inc.; Merck & Co.,
646 Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation;
647 Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company.

648 The Canadian Institutes of Health Research is providing funds to support
649 ADNI clinical sites in Canada. Private sector contributions are facilitated
650 by the Foundation for the National Institutes of Health (www.fnih.org).
651 The grantee organization is the Northern California Institute for Research
652 and Education, and the study is coordinated by the Alzheimer's Disease
653 Cooperative Study at the University of California, San Diego.

654 ADNI data are disseminated by the Laboratory for Neuro Imaging at the
655 University of California, Los Angeles. This research was also supported by
656 NIH grants P30 AG010129 and K01 AG030514.

657 References

- 658 Ahn, B.J., Yim, D.S., 2009. Comparison of Parametric and Bootstrap
659 Method in Bioequivalence Test. *The Korean Journal of Physiology and*
660 *Pharmacology* 13, 367.
- 661 Apostolova, L.G., Thompson, P.M., Green, A.E., Hwang, K.S., Zoumalan,
662 C., Jack, C.R., Harvey, D.J., Petersen, R.C., Thal, L.J., Aisen, P.S., Toga,
663 A.W., Cummings, J.L., DeCarli, C.S., 2010. 3D comparison of low, in-
664 termediate, and advanced hippocampal atrophy in MCI. *Human Brain*
665 *Mapping* 31, 786–797.
- 666 Apostolova, L.G., Zarow, C., Biado, K., Hurtz, S., Boccardi, M., Somme, J.,
667 Honarpisheh, H., Blanken, A.E., Brook, J., Tung, S., Lo, D., Ng, D., Al-
668 ger, J.R., Vinters, H.V., Bocchetta, M., Duvernoy, H., Jack, C.R., Frisoni,
669 G.B., 2015. Relationship between hippocampal atrophy and neuropathol-
670 ogy markers: a 7T MRI validation study of the EADC-ADNI Harmonized
671 Hippocampal Segmentation Protocol. *Alzheimer’s & dementia : the jour-*
672 *nal of the Alzheimer’s Association* 11, 139–50.
- 673 Barnes, J., Bartlett, J.W., van de Pol, L.a., Loy, C.T., Scahill, R.I., Frost, C.,
674 Thompson, P., Fox, N.C., 2009. A meta-analysis of hippocampal atrophy
675 rates in Alzheimer’s disease. *Neurobiology of Aging* 30, 1711–1723.
- 676 Bateman, R.J., Xiong, C., Benzinger, T.L.S., Fagan, A.M., Goate, A., Fox,
677 N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., Holtzman, D.M.,
678 Santacruz, A., Buckles, V., Oliver, A., Moulder, K., Aisen, P.S., Ghetti,
679 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. *The New England journal of medicine* 367, 795–804.

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.

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. *Alzheimer’s & Dementia*
 11, 175–183.

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 hippocam-
 pus on magnetic resonance: The EADC-ADNI project. *Alzheimer’s &
 Dementia* .

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.

Chow, N., Hwang, K.S., Hurtz, S., Green, A.E., Somme, J.H., Thompson,

703 P.M., Elashoff, D.A., Jack, C.R., Weiner, M., Apostolova, L.G., 2015.
704 Comparing 3T and 1.5T MRI for Mapping Hippocampal Atrophy in the
705 Alzheimer’s Disease Neuroimaging Initiative. American Journal of Neuro-
706 radiology .

707 Cortes, C., Mohri, M., 2004. Confidence Intervals for the Area Under the
708 ROC Curve., in: Nips.

709 Duning, T., Kloska, S., Steinsträter, O., Kugel, H., Heindel, W., Knecht, S.,
710 2005. Dehydration confounds the assessment of brain atrophy. Neurology
711 64, 548–550.

712 Fennema-Notestine, C., Hagler, D.J., McEvoy, L.K., Fleisher, A.S., Wu,
713 E.H., Karow, D.S., Dale, A.M., 2009. Structural MRI biomarkers for
714 preclinical and mild Alzheimer’s disease. Human Brain Mapping 30, 3238–
715 3253.

716 Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to
717 Alzheimer’s disease. Lancet 363, 392–4.

718 Frankó, E., Joly, Olivier, A.D.N.I., 2013. Evaluating Alzheimer’s disease
719 progression using rate of regional hippocampal atrophy. PloS one 8, e71354.

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

723 Frisoni, G.B., Jack, C.R., Bocchetta, M., Bauer, C., Frederiksen, K.S., Liu,
724 Y., Preboske, G., Swihart, T., Blair, M., Cavedo, E., Grothe, M.J., Lan-

fredi, 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., Lehericy, 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.
Alzheimer's & Dementia .

Gengsheng Qin, Hotilovac, L., 2007. Comparison of non-parametric confi-
dence intervals for the area under the ROC curve of a continuous-scale
diagnostic test. Statistical Methods in Medical Research 17, 207–221.

Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim,
H.S., Niethammer, M., Dubois, B., Lehericy, 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.

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.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas un-

747 der receiver operating characteristic curves derived from the same cases.
 748 Radiology 148, 839–43.

749 Hogan, R.E., Wang, L., Bertrand, M.E., Willmore, L.J., Bucholz, R.D.,
 750 Nassif, A.S., Csernansky, J.G., 2004. MRI-based high-dimensional hip-
 751 pocampal mapping in mesial temporal lobe epilepsy. *Brain: A journal of*
 752 *neurology* 127, 1731–40.

753 Hovorkova, P., Kristofikova, Z., Horinek, A., Ripova, D., Majer, E., Zach,
 754 P., Sellinger, P., Rícný, J., 2008. Lateralization of 17beta-hydroxysteroid
 755 dehydrogenase type 10 in hippocampi of demented and psychotic people.
 756 *Dementia and geriatric cognitive disorders* 26, 193–8.

757 Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Har-
 758 vey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M.,
 759 Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-
 760 Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward,
 761 H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Deb-
 762 bins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G.,
 763 Mugler, J., Weiner, M.W., 2008. The Alzheimer’s Disease Neuroimaging
 764 Initiative (ADNI): MRI methods. *Journal of magnetic resonance imaging*
 765 : *JMRI* 27, 685–91.

766 Jovicich, J., Marizzoni, M., Sala-Llloch, R., Bosch, B., Bartrés-Faz, D.,
 767 Arnold, J., Benninghoff, J., Wiltfang, J., Roccatagliata, L., Nobili, F.,
 768 Hensch, T., Tränkner, A., Schönknecht, P., Leroy, M., Lopes, R., Bordet,
 769 R., Chanoine, V., Ranjeva, J.P., Didic, M., Gros-Dagnac, H., Payoux, P.,

770 Zoccatelli, G., Alessandrini, F., Beltramello, A., Bargalló, N., Blin, O.,
771 Frisoni, G.B., 2013. Brain morphometry reproducibility in multi-center
772 3T MRI studies: A comparison of cross-sectional and longitudinal seg-
773 mentations. *NeuroImage* 83, 472–484.

774 Kottas, M., Kuss, O., Zapf, A., 2014. A modified Wald interval for the
775 area under the ROC curve (AUC) in diagnostic case-control studies. *BMC*
776 *Medical Research Methodology* 14, 26.

777 Leung, K.K.K., Barnes, J., Ridgway, G.G.R.G., Bartlett, J.W., Clarkson,
778 M.M.J., Macdonald, K., Schuff, N., Fox, N.N.C.N., Ourselin, S., Barlett,
779 J., 2010. Automated cross-sectional and longitudinal hippocampal vol-
780 ume measurement in mild cognitive impairment and Alzheimer’s disease.
781 *Neuroimage* 51, 1345–1359.

782 Lobanova, I., Sohail, A., Adil, M., Saed, A., Qureshi, A., 2014. Progressive
783 Hippocampus Atrophy in Persons with Alzheimer’s Disease: A Longitudi-
784 nal MRI Study. *Neurology* 82, P6.333–.

785 Lötjönen, J., Wolz, R., Koikkalainen, J., Julkunen, V., Thurfjell, L.,
786 Lundqvist, R., Waldemar, G., Soininen, H., Rueckert, D., 2011. Fast
787 and robust extraction of hippocampus from MR images for diagnostics of
788 Alzheimer’s disease. *NeuroImage* 56, 185–196.

789 Macdonald, K.E., Leung, K.K., Bartlett, J.W., Blair, M., Malone, I.B.,
790 Barnes, J., Ourselin, S., Fox, N.C., 2014. Automated template-based hip-
791 pocampal segmentations from MRI: the effects of 1.5T or 3T field strength
792 on accuracy. *Neuroinformatics* 12, 405–12.

793 Maclaren, J., Han, Z., Vos, S.B., Fischbein, N., Bammer, R., 2014. Reliability
794 of brain volume measurements : A test-retest dataset. *Nature - scientific*
795 *data* , 1–9.

796 Martinez, W.L., 2011. *Computational Statistics in MATLAB*. Wiley Inter-
797 disciplinary Reviews: Computational Statistics 3, 69–74.

798 mazziotta, Toga, Evans, Fox, L., 1995. A probabilistic atlas of the human
799 brain: theory and rationale for its developement. *Neuroimage* 2, 89–101.

800 McEvoy, L.K., Holland, D., Hagler, D.J., Fennema-Notestine, C., Brewer,
801 J.B., Dale, A.M., 2011. Mild cognitive impairment: baseline and lon-
802 gitudinal structural MR imaging measures improve predictive prognosis.
803 *Radiology* 259, 834–43.

804 Morra, J., Tu, Z., Apostolova, L., Green, A., Avedissian, C., Madsen, S.,
805 Parikshak, N., Toga, A., Jack, C., Schuff, N., 2009. Automated Hippocam-
806 pal Segmentation and Mapping Reveals Genetically Accelerated Tissue
807 Loss in 1-year Repeat MRI data from 490 Alzheimer’s Disease, MCI, and
808 Control Subjects. *NeuroImage* 47, S122–S122.

809 Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Avedissian, C., Mad-
810 sen, S.K., Parikshak, N., Hua, X., Toga, A.W., Jack, C.R., Weiner, M.W.,
811 Thompson, P.M., 2008. Validation of a fully automated 3D hippocampal
812 segmentation method using subjects with Alzheimer’s disease mild cogni-
813 tive impairment, and elderly controls. *NeuroImage* 43, 59–68.

814 Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson,
815 P.M., 2010. Comparison of adaboost and support vector machines for de-

816 tecting alzheimer’s disease through automated hippocampal segmentation.
 817 IEEE Transactions on Medical Imaging 29, 30–43.

818 Okonkwo, O.C., Xu, G., Dowling, N.M., Bendlin, B.B., LaRue, A., Hermann,
 819 B.P., Kosciak, R., Jonaitis, E., Rowley, H.a., Carlsson, C.M., Asthana, S.,
 820 Sager, M.a., Johnson, S.C., 2012. Family history of Alzheimer disease
 821 predicts hippocampal atrophy in healthy middle-aged adults. Neurology
 822 78, 1769–1776.

823 Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian
 824 model of shape and appearance for subcortical brain segmentation. Neuro-
 825 Image 56, 907–22.

826 Picco, A., Polidori, M.C., Ferrara, M., Cecchetti, R., Arnaldi, D., Baglioni,
 827 M., Morbelli, S., Bastiani, P., Bossert, I., Fiorucci, G., Brugnolo, A., Dot-
 828 torini, M.E., Nobili, F., Mecocci, P., 2014. Plasma antioxidants and brain
 829 glucose metabolism in elderly subjects with cognitive complaints. Euro-
 830 pean journal of nuclear medicine and molecular imaging 41, 764–75.

831 Pruessner, J.C., Li, L.M., Serles, W., Pruessner, M., Collins, D.L., Kabani,
 832 N., Lupien, S., Evans, A.C., 2000. Volumetry of hippocampus and amyg-
 833 dala with high-resolution MRI and three-dimensional analysis software:
 834 minimizing the discrepancies between laboratories. Cerebral cortex 10,
 835 433–42.

836 Reig, S., Sánchez-González, J., Arango, C., Castro, J., González-Pinto, A.,
 837 Ortuño, F., Crespo-Facorro, B., Bargalló, N., Desco, M., 2009. Assessment

838 of the increase in variability when combining volumetric data from different
839 scanners. *Human brain mapping* 30, 355–68.

840 Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012a. Within-subject
841 template estimation for unbiased longitudinal image analysis. *NeuroImage*
842 61, 1402–18.

843 Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B., 2012b. Within-subject
844 template estimation for unbiased longitudinal image analysis. *NeuroImage*
845 61, 1402–18.

846 Rusinek, H., De Santi, S., Frid, D., Tsui, W.H., Tarshish, C.Y., Convit, A.,
847 de Leon, M.J., 2003. Regional brain atrophy rate predicts future cognitive
848 decline: 6-year longitudinal MR imaging study of normal aging. *Radiology*
849 229, 691–696.

850 Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L.M., Trojanowski,
851 J.Q., Thompson, P.M., Jack, C.R., Weiner, M.W., 2009. MRI of hip-
852 pocampal volume loss in early Alzheimer’s disease in relation to ApoE
853 genotype and biomarkers. *Brain* 132, 1067–1077.

854 Shen, D., Moffat, S., Resnick, S.M., Davatzikos, C., 2002. Measuring size
855 and shape of the hippocampus in MR images using a deformable shape
856 model. *NeuroImage* 15, 422–34.

857 Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and
858 asymmetry in mild cognitive impairment and Alzheimer’s disease: Meta-
859 analyses of MRI studies. *Hippocampus* 19, 1055–64.

860 Smith, S.M., 2002. Fast robust automated brain extraction. Human brain
861 mapping 17, 143–55.

862 Spulber, G., Simmons, A., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki,
863 M., Kloszewska, I., Soininen, H., Spenger, C., Lovestone, S., Wahlund,
864 L.O., Westman, E., 2013. An MRI-based index to measure the severity of
865 Alzheimer’s disease-like structural pattern in subjects with mild cognitive
866 impairment. Journal of Internal Medicine 273, 396–409.

867 Thirion, J.P., 1998. Image matching as a diffusion process: an analogy with
868 Maxwell’s demons. Medical image analysis 2, 243–60.

869 Thompson, W.K., Holland, D., 2011. Bias in tensor based morphometry
870 Stat-ROI measures may result in unrealistic power estimates. NeuroImage
871 57, 1–4.

872 Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., 2010a.
873 LEAP: Learning embeddings for atlas propagation. NeuroImage 49, 1316–
874 1325.

875 Wolz, R., Heckemann, R.a., Aljabar, P., Hajnal, J.V., Hammers, A.,
876 Lötjönen, J., Rueckert, D., 2010b. Measurement of hippocampal atro-
877 phy using 4D graph-cut segmentation: application to ADNI. NeuroImage
878 52, 109–18.

879 Wolz, R., Schwarz, A.J., Yu, P., Cole, P.E., Rueckert, D., Jack, C.R., Raunig,
880 D., Hill, D., 2014. Robustness of automated hippocampal volumetry across
881 magnetic resonance field strengths and repeat images. Alzheimer’s and
882 Dementia 10, 430–438.

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- <i>nc</i>	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- <i>co</i>	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)

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

Table 2: Mean Λ values.

	CTRL	MCI- <i>nc</i>	MCI- <i>co</i>	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).

Table 3: Performance (AUC).

Feat.	Impl.	CTRL/MCI- <i>nc</i>	CTRL/MCI- <i>co</i>	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.

Table 4: Performance comparison.

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

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

Table 5: Sample size calculation.

Impl.	CTRL	MCI- <i>nc</i>	MCI- <i>co</i>	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).

Figure 1: Positioning and content of a sample hippocampal VOI.

Figure 2: Schematic flowchart of the four implementations. The four MRI drawings represents the baseline, repeat, month 12 and month 24 scans. In implementation A (section 2.4.1) all four images follow a separate preprocessing and segmentation path. In implementation B (section 2.4.2) 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 (section 2.4.3) the VOIs extracted with the B procedure are segmented together with atlas re-normalization. Implementation D (section 2.4.4) avoids the shape segmentation and delivers an equivalent volume only.

Figure 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 *GDIseq* hippocampal tracing.

R

L

Figure 4: Distribution of Λ for the right hippocampus (top) and left hippocampus (bottom) on Controls (CTRL), Mild Cognitive Impairment non-converters / converters (MCIInc/MCICo) 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

Figure 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).

Figure 6: Boxplot 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.

883 Appendix A. Segmentation algorithm

884 The *GDIseg* algorithm is based on a set of manually traced segmentations
885 by expert and certified readers from the HarP project. At the time of this
886 writing, 100 manual tracings were made available (58 1.5T and 42 3T)

887 Reference HarP images were processed as in section 2.2. In addition we
888 extracted the VOIs from the manually segmented masks, using the same
889 coordinates found for extracting the VOIs from the MRI.

890 We refer to the set of VOIs from the HarP MR images as Template
891 Boxes (*TBs*) and the set of the corresponding segmented masks as Template
892 Masks (*TMs*), both naturally coming with the right (*R*) and left (*L*) label.
893 A pictorial overview of the segmentation process is shown in supplemental
894 figure S1.

895 For each new segmentation, the MRI goes through the pre-process steps
896 up to the extraction of both hippocampal VOIs (target VOIs). Subsequently,
897 each *TB* is mapped onto the target VOI with a deformable registration
898 transform, implemented in ITK with the “Diffeomorphic Demons” algorithm
899 (Thirion (1998) and <http://hdl.handle.net/1926/510>). The resulting de-
900 formation field - one for each *TB* - is applied to the corresponding *TM*.

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

905 A probabilistic atlas *A* is generated by weighted average of all deformed
906 *TMs*, followed by a normalization. All VOIs, *TBs*, *TMs* and their deformed

907 counterparts (δTBs , δTMs) have the same dimensions and number of voxels,
 908 so that we can write

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

909 where N_t is the number of templates.

910 In order to find the weights w_i , the TBs are ranked according to the Pear-
 911 son correlation coefficient r with the target VOI. The correlation coefficient
 912 is not computed over the whole volume of the VOI, but on a subset of voxels
 913 corresponding to the volume surrounding the TM . The detailed procedure
 914 consists in three steps: a) dilation of the the binary TM (distance of $3mm$),
 915 b) mapping of the the dilated TM onto both the target VOI and the TB
 916 (voxel selection), c) computation of the correlation coefficient r between the
 917 intensities of both volumes over the selected voxels.

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

924 The correlation rank is used to compute the weights in the TMs average,
 925 under the hypothesis that it contains information on the “true segmentation”.
 926 In this sense, correlation values are used as proxies for the segmentation
 927 similarity.

928 Since we do not know the target VOI true segmentation, we use a sur-
 929rogate target δTB^* - that is the deformed TB with the best rank - in place

930 of the target VOI, with the benefit that the true segmentation δTM^* is now
 931 available.

932 Weights are thought to be a simple exponential functions of the correla-
 933 tion coefficient, they are computed by minimizing the distance m over the
 934 free parameter s ($s \geq 0$)

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

935

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

936 where N is the number of templates, i^* is the index of the surrogate tar-
 937 get δTB^* and r_i are the correlation coefficients now computed between the
 938 surrogate target δTB^* and the TBs .

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

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

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

950 to the probabilistic atlas.

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

953 The optimal threshold is defined as

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

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

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

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

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

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

969 Since the 100 images from the HarP database consisted in 58 1.5T and 42

970 3.0T MRI, we show the performance by field strength, demonstrating that
971 the segmentation algorithm is not affected by the B-field intensity.

Table S1: ADNI subjects id.

002.S.0295	002.S.0413	002.S.0559	002.S.0685	002.S.0729	002.S.0782	002.S.0838	002.S.1018	002.S.1070	002.S.1155	002.S.1261	002.S.1268	002.S.1280	003.S.0981	003.S.1057	003.S.1074
003.S.1122	005.S.0222	005.S.0223	005.S.0324	005.S.0448	005.S.0546	005.S.0553	005.S.0602	005.S.0610	005.S.0814	005.S.1224	005.S.1341	006.S.0498	006.S.0547	006.S.0675	006.S.0681
006.S.0731	006.S.1130	007.S.0041	007.S.0068	007.S.0101	007.S.0128	007.S.0249	007.S.0283	007.S.0316	007.S.0344	007.S.0414	007.S.0698	007.S.1206	007.S.1222	007.S.1339	010.S.0067
010.S.0419	010.S.0786	011.S.0003	011.S.0005	011.S.0010	011.S.0016	011.S.0021	011.S.0022	011.S.0023	011.S.0053	011.S.0183	011.S.0241	011.S.0326	011.S.0362	011.S.0861	011.S.1080
011.S.1282	012.S.0634	012.S.0637	012.S.0689	013.S.0240	013.S.0325	013.S.0502	013.S.0575	013.S.1035	013.S.1186	013.S.1205	014.S.0169	014.S.0328	014.S.0519	014.S.0520	014.S.0548
014.S.0557	014.S.0558	014.S.0563	014.S.0658	016.S.0359	016.S.0702	016.S.0991	016.S.1028	016.S.1121	016.S.1326	018.S.0057	018.S.0142	018.S.0155	018.S.0286	018.S.0335	018.S.0369
018.S.0406	018.S.0450	018.S.0633	018.S.0682	020.S.0097	020.S.0899	020.S.1288	021.S.0141	021.S.0159	021.S.0231	021.S.0273	021.S.0276	021.S.0337	021.S.0343	021.S.0424	021.S.0626
021.S.0647	021.S.0753	021.S.0984	021.S.1109	022.S.0014	022.S.0066	022.S.0096	022.S.0129	022.S.0130	022.S.0543	022.S.0544	022.S.0750	022.S.0961	022.S.1097	022.S.1351	022.S.1394
023.S.0030	023.S.0031	023.S.0042	023.S.0058	023.S.0061	023.S.0081	023.S.0083	023.S.0084	023.S.0093	023.S.0126	023.S.0139	023.S.0217	023.S.0331	023.S.0376	023.S.0388	023.S.0604
023.S.0625	023.S.0887	023.S.0916	023.S.0926	023.S.0963	023.S.1046	023.S.1126	023.S.1190	023.S.1247	023.S.1262	024.S.0985	024.S.1063	024.S.1171	024.S.1307	027.S.0074	027.S.0116
027.S.0118	027.S.0120	027.S.0179	027.S.0256	027.S.0307	027.S.0403	027.S.0404	027.S.0408	027.S.0644	027.S.0835	027.S.0850	027.S.1045	027.S.1081	027.S.1082	027.S.1213	027.S.1254
027.S.1385	027.S.1387	029.S.0824	029.S.0843	029.S.0845	029.S.0866	029.S.0878	029.S.0914	029.S.1056	029.S.1073	029.S.1215	029.S.1218	029.S.1318	029.S.1384	031.S.0351	031.S.0554
031.S.0568	031.S.0618	031.S.0830	031.S.0867	031.S.1066	031.S.1209	032.S.0147	032.S.0187	032.S.0214	032.S.0400	032.S.0479	032.S.0677	032.S.0718	032.S.1169	033.S.0511	033.S.0513
033.S.0514	033.S.0516	033.S.0567	033.S.0723	033.S.0724	033.S.0725	033.S.0733	033.S.0739	033.S.0741	033.S.0889	033.S.0906	033.S.0920	033.S.0922	033.S.0923	033.S.1016	033.S.1086
033.S.1098	033.S.1279	033.S.1281	033.S.1283	033.S.1285	033.S.1308	033.S.1309	035.S.0033	035.S.0048	035.S.0156	035.S.0204	035.S.0292	035.S.0341	035.S.0555	035.S.0997	036.S.0576
036.S.0577	036.S.0656	036.S.0673	036.S.0748	036.S.0759	036.S.0760	036.S.0813	036.S.0869	036.S.0945	036.S.1023	036.S.1135	036.S.1240	037.S.0150	037.S.0303	037.S.0327	037.S.0377
037.S.0454	037.S.0467	037.S.0539	037.S.0552	037.S.0566	037.S.0588	037.S.1078	041.S.0125	041.S.0262	041.S.0314	041.S.0679	041.S.0898	041.S.1002	041.S.1010	041.S.1260	041.S.1368
041.S.1418	041.S.1425	051.S.1123	051.S.1131	051.S.1296	052.S.0671	052.S.0951	052.S.1054	052.S.1250	052.S.1251	052.S.1346	053.S.0389	053.S.0507	053.S.0621	053.S.0919	057.S.0464
057.S.0474	057.S.0643	057.S.0779	057.S.0818	057.S.0839	057.S.0934	057.S.0941	057.S.1007	057.S.1217	057.S.1265	057.S.1269	057.S.1371	057.S.1373	057.S.1379	062.S.0535	062.S.0578
062.S.0690	062.S.0730	062.S.0768	062.S.0793	062.S.1182	062.S.1299	067.S.0076	067.S.0077	067.S.0176	067.S.0177	067.S.0257	067.S.0290	067.S.0336	067.S.0607	068.S.0109	068.S.0127
068.S.0210	068.S.0473	073.S.0089	073.S.0311	073.S.0386	073.S.0518	073.S.0565	073.S.0746	073.S.0909	082.S.0832	094.S.0434	094.S.0526	094.S.0692	094.S.0711	094.S.0921	094.S.1164
094.S.1188	094.S.1241	094.S.1267	094.S.1314	098.S.0149	098.S.0160	098.S.0171	098.S.0172	098.S.0269	098.S.0667	098.S.0896	099.S.0040	099.S.0051	099.S.0054	099.S.0090	099.S.0291
099.S.0352	099.S.0372	099.S.0470	099.S.0533	099.S.0534	099.S.0551	099.S.1034	099.S.1144	100.S.0995	109.S.0950	109.S.0967	109.S.1114	109.S.1157	109.S.1183	114.S.0166	114.S.0173
114.S.0374	114.S.0378	114.S.0416	114.S.0601	114.S.0679	114.S.1106	114.S.1118	116.S.0361	116.S.0370	116.S.0382	116.S.0392	116.S.0487	116.S.0648	116.S.0649	116.S.0657	116.S.0752
116.S.0834	116.S.1232	116.S.1249	116.S.1271	116.S.1315	123.S.0050	123.S.0072	123.S.0091	123.S.0094	123.S.0106	123.S.0108	123.S.0113	123.S.0162	123.S.0298	123.S.0390	123.S.1300
126.S.0605	126.S.0680	126.S.0708	126.S.0784	126.S.0865	126.S.0891	126.S.1187	126.S.1221	127.S.0259	127.S.0260	127.S.0394	127.S.0431	127.S.0622	127.S.0754	127.S.0925	127.S.1032
127.S.1140	127.S.1382	127.S.1419	127.S.1427	128.S.0167	128.S.0188	128.S.0200	128.S.0205	128.S.0216	128.S.0225	128.S.0227	128.S.0230	128.S.0258	128.S.0266	128.S.0500	128.S.0522
128.S.0545	128.S.0608	129.S.0778	129.S.1246	130.S.0102	130.S.0285	130.S.0289	130.S.0783	131.S.0384	131.S.1389	133.S.0727	133.S.0912	133.S.1031	136.S.0107	136.S.0186	136.S.0196
136.S.0300	136.S.0426	136.S.0429	137.S.0972	137.S.1414	141.S.0767	141.S.0810	941.S.1197	941.S.1311	123.S.1300	016.S.0702	036.S.0656				

Table S2: Cross-validation performance.

Metric	1.5T+3.0T	1.5T	3.0T
Dc	0.85 (0.82 – 0.88)	0.85 (0.83 – 0.87)	0.86 (0.81 – 0.89)
Pr	0.87 (0.80 – 0.92)	0.87 (0.80 – 0.91)	0.87 (0.76 – 0.93)
Rc	0.85 (0.79 – 0.90)	0.84 (0.79 – 0.89)	0.85 (0.76 – 0.91)

Dice (Dc), Recall (Rc) and Precision (Pr) measured with a k-fold cross-validation method on the 100 HarP manual tracings. Statistics are calculated on the right and left hippocampi together, for a total of 200 (1.5T+3.0T), 116 (1.5T) and 94 (3.0T) segmentations. Within parentheses are the 5% and 95% confidence level values.

Figure S1: Segmentation algorithm flowchart. TB set: reference VOIs (template boxes); TM set: manually traced reference segmentations (template masks); δTB , δTM : reference boxes and labels after the deformable registration; δTB^* , δTM^* : surrogate box and mask, i.e. the transformed template box and mask which has the highest correlation rank with the VOI.

Figure S2: Normal probability plot of the reproducibility error Δ . Dotted lines show the best gaussian distribution fitted over the experimental data. Deviation from the straight line indicates non-gaussian behaviour.