

Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project

Martina Bocchetta^{a,b}, Marina Boccardi^a, Rossana Ganzola^c, Liana G. Apostolova^d, Gregory Preboske^e, Dominik Wolf^f, Clarissa Ferrari^g, Patrizio Pasqualetti^{h,i}, Nicolas Robitaille^c, Simon Duchesne^c, Clifford R. Jack, Jr.,^e Giovanni B. Frisoni^{a,j,*}, for the EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative¹

^aLENITEM (Laboratory of Epidemiology, Neuroimaging and Telemedicine), IRCCS-Centro S. Giovanni di Dio-Fatebenefratelli Brescia, Brescia, Italy

^bDepartment of Molecular and Translational Medicine, University of Brescia, Brescia, Italy

^cDepartment of Radiology, Université Laval and Centre de Recherche de l'Institut universitaire de santé mentale de Québec, Québec City, Canada

^dMary S. Easton Center for Alzheimer's Disease Research and Laboratory of Neuroimaging, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

^eDepartment of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN, USA

^fKlinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität Mainz, Mainz, Germany

^gIRCCS-Centro S. Giovanni di Dio-Fatebenefratelli Brescia, Brescia, Italy

^hSeSMIT (Service for Medical Statistics and Information Technology), AFaR (Fatebenefratelli Association for Research), Fatebenefratelli Hospital, Rome, Italy

ⁱUnit of Clinical and Molecular Epidemiology, IRCCS "San Raffaele Pisana," Rome, Italy

^jMemory Clinic and LANVIE - Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland

Abstract

Background: A globally harmonized protocol (HarP) for manual hippocampal segmentation based on magnetic resonance has been recently developed by a task force from European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI). Our aim was to produce benchmark labels based on the HarP for manual segmentation.

Methods: Five experts of manual hippocampal segmentation underwent specific training on the HarP and segmented 40 right and left hippocampi from 10 ADNI subjects on both 1.5 T and 3 T scans. An independent expert visually checked segmentations for compliance with the HarP. Descriptive

Project collaborators include: George Bartzokis (Department of Psychiatry, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA), Charles DeCarli (Department of Neurology, University of California, Davis, Sacramento, CA, USA), Leyla de Toledo-Morrell (Department of Neurological Sciences, Rush University, Chicago, IL, USA), Andreas Fellgiebel (Klinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität Mainz, Mainz, Germany), Michael Firbank (Institute for Ageing and Health, Newcastle University, Wolfson Research Centre, Newcastle, UK), Lotte Gerritsen (Karolinska Institute, Stockholm, Sweden), Wouter Henneman (Department of Radiology and Nuclear Medicine; Image Analysis Center, VU University Medical Center, Amsterdam, The Netherlands), Ronald J. Killiany (Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA, USA), Nikolai Malykhin (Department of Biomedical Engineering, Centre for Neuroscience, University of Alberta, Edmonton, Alberta, Canada), Jens C. Pruessner (McGill Centre for Studies in Aging, Department of Psychiatry, McGill University, Montreal, Quebec, Canada), Hilkkä Soininen (Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland), Lei Wang (Department of Psychiatry and Behavioral Sciences, Northwestern Univer-

sity Feinberg School of Medicine, Chicago, IL, USA), Craig Watson (Department of Neurology, University of California, Davis, Sacramento, CA, USA), Henrike Wolf (Department of Psychiatry Research and Geriatric Psychiatry, Psychiatric University Hospitals, University of Zurich, Zurich, Switzerland; German Center for Neurodegenerative Diseases [DZNE], Bonn, Germany).

This manuscript was approved by the ADNI Data and Publication Committee on June 7, 2013.

¹Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*Corresponding author. Tel.: +39-030-3501361; Fax: +39-030-3501592.

E-mail address: gfrisoni@fatebenefratelli.it

measures of agreement between tracers were intraclass correlation coefficients (ICCs) of crude volumes and similarity coefficients of three-dimensional volumes.

Results: Two hundred labels have been provided for the 20 magnetic resonance images. Intra- and interrater ICCs were >0.94 , and mean similarity coefficients were 1.5 T, 0.73 (95% confidence interval [CI], 0.71–0.75); 3 T, 0.75 (95% CI, 0.74–0.76).

Conclusion: Certified benchmark labels have been produced based on the HarP to be used for tracers' training and qualification.

© 2015 The Alzheimer's Association. Published by Elsevier Inc. All rights reserved.

Keywords:

Hippocampus; Hippocampal atrophy; Hippocampal volume; Harmonized protocol; Harmonization; Anatomic landmark; Alzheimer's disease; Manual tracing; Medial temporal lobes; MRI; Neuroimaging; ADNI; Standard operating procedures

1. Introduction

The revised diagnostic criteria for Alzheimer's disease (AD) proposed by the International Working Group [1,2] and National Institute on Aging-Alzheimer's Association groups [3–5] include hippocampal volumetry as a diagnostic marker. The European Medicine Agency has also recently qualified hippocampal volumetry for clinical trial enrichment in mild cognitive impairment (MCI) [6]. To date, manual segmentation on T1-weighted high-resolution magnetic resonance images (MRIs) is the widely accepted in vivo gold standard for hippocampal volumetry [7]. Nonetheless, the availability and widespread use of a large number of different protocols for manual segmentation across laboratories [8,9] has resulted in highly heterogeneous hippocampal volumetric estimates [8,10], preventing comparisons among different studies and having a significant negative impact on the qualification of hippocampal volumetry as a reliable diagnostic biomarker and surrogate marker for clinical trials.

In 2008, the European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) centers, supported by the Alzheimer's Association, initiated the development and validation of a consensus harmonized protocol (HarP) for manual hippocampal segmentation on MRIs [11,12] (www.hippocampal-protocol.net). The previous steps of this project consisted of an exhaustive literature survey for available hippocampal segmentation protocols, careful analysis of the segmentation differences among the 12 most popular protocols, operationalization and quantification of these differences via segmentation units (SUs) in terms of reliability values, their effects on total hippocampal volume estimates, and susceptibility to AD-related atrophy [10,13]. These quantitative data were provided to a Delphi panel composed of experts in hippocampal anatomy in healthy aging and AD. The panelists were asked to make evidence-based decisions and, through iterative rounds, to converge on a consensus definition of HarP [14].

The aim of the present work was to provide benchmark hippocampal segmentations certified for reflecting the landmarks and segmentation modalities defined in the HarP criteria. This step consisted in translating the Delphi panel

decisions and concepts into the concrete gold standard to be used for the training/testing of future tracers.

2. Methods

2.1. Design

Fig. 1 shows the design of the study. The *preliminary and training phase* served the purpose to train and qualify five experts in manual hippocampal segmentations as “master tracers.” Of these, two (M.Bocch. and R.G.) were involved in the development of the HarP from its beginning (*preliminary phase*, Fig. 1 [10,13]). The other three (G.P., L.G.A., and D.W.) were enrolled subsequently, based on their acknowledged contributions to hippocampal segmentations in AD. The *benchmark phase* was aimed at developing benchmark hippocampal segmentations that “embody” the HarP.

2.1.1. Preliminary phase

The preliminary phase, carried out in previous steps of the project, was aimed to operationalize the differences among the segmentation protocols into SUs [13]. SUs [10] correspond to either portions of the hippocampus that were included by some but not other segmentation protocols (Alveus/Fimbria [A/F], Crura and Tail End) or that were variably segmented (i.e., Subiculum - Horizontal, Morphological or Oblique criteria). The hippocampal tissue included by all protocols was coined “Minimum Hippocampus.”

M.Bocch. and R.G. manually segmented SUs on MRI to derive reliability values, SUs effect on total hippocampal volume, and susceptibility to AD-related atrophy. These data were fed to the Delphi panel of experts to facilitate their consensus agreement on the HarP [10]. In this preliminary phase, M.Bocch. and R.G. extensively familiarized with the SU segmentations and had their SU reliability measured.

2.1.2. Training phase

Three additional experts (G.P., L.G.A., and D.W.) in hippocampal anatomy in aging and AD were selected from other centers participating in the HarP project. They were required to complete a detailed training on SUs to make their segmentation background and performance equal to R.G. and M.Bocch's.

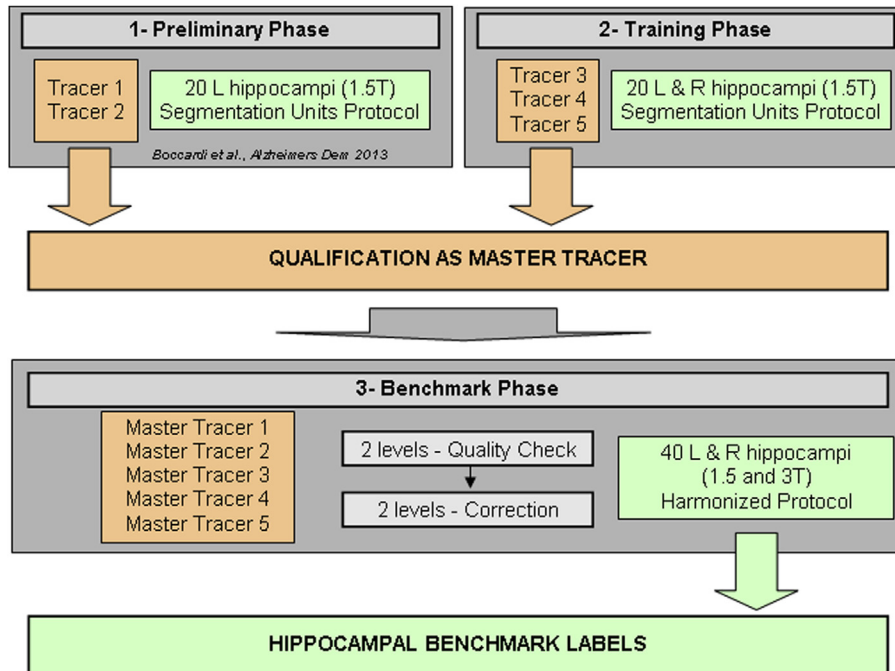


Fig. 1. Study design. In the preliminary phase, Segmentation Units (SUs) summarizing the differences among the most frequently used segmentation protocols were identified and operationalized [10,13]; two tracers took part to this phase and segmented a large number of hippocampi into SUs [13], including those of the so-called preliminary phase [13]. These data were fed to the Delphi panel, which converged on the landmarks of the harmonized protocol (HarP) [14]. In the training phase, three additional tracers were instructed to segment SUs to pair their segmentation background at the other two. In the benchmark phase, all the five tracers performed segmentations on a new set of hippocampi following the HarP, thus producing the benchmark labels.

Teleconferences were carried out with real-time screen sharing and segmentation demonstrations to facilitate the new tracers' learning of segmentation tool usage and SU landmarks [10]. Tracers could practice on a couple of images: segmentations were carefully inspected for SU protocol adherence and feedback was provided. The three tracers were then asked to trace and retrace all SUs on a new set of 10 ADNI images (for a total of 20 hippocampi) to compute reliability values. Reliability measurements for each SU were computed. To qualify as a master tracer for the HarP project, an intra- and interrater intraclass correlation coefficient (ICC) ≥ 0.90 was required for the global hippocampal volume computed as the sum of all most inclusive SUs [10].

2.1.3. Benchmark phase

To provide certified labels, we set a two-stage procedure consisting of segmentation by expert HarP tracers and check by an independent HarP expert not involved in segmentation. This two-stage procedure by two independent experts guarantees that any incidental divergences or tendencies to systematic biases are corrected before the final release of the certified set.

Specifically, global hippocampal segmentations were made according to the HarP as defined by the Delphi panel [14]. When completed, all the benchmark segmentations underwent two rounds of slice-by-slice inspection by the independent expert of the HarP (M.Bocca.) whose role was to ascertain that all segmentations were fully compliant with

the HarP criteria defined by the Delphi panel [14]. For both rounds, all master tracers were required to correct any divergence from the HarP criteria. Master tracers received visual feedback displaying the five segmentations simultaneously mapped onto the correspondent MRI slices, together with written descriptions of the divergences from the HarP needing correction. Tracers were asked to discuss with M.Bocca. any required corrections in case they disagreed based on their prior experience of hippocampal morphology and their understanding of the HarP. When this happened, both the referee and the tracer navigated together in three dimensions (3D) through the problematic slices, and landmarks were examined and discussed in detail until the segmentation correctly resembling the HarP was identified. Regions of disagreement were scrutinized for being due to (1) anatomic ambiguity where all traces were still in compliance with the HarP (no corrections required), (2) segmentation errors in tissue attribution or noncompliance to the HarP (corrections required), or (3) ambiguous or insufficiently detailed landmark definition in the HarP. In the latter case, the HarP landmark definitions and descriptions were revised or better detailed to disambiguate the written instructions reported in the HarP text (www.hippocampal-protocol.net).

2.2. Feedback and improvement of HarP written instructions

The interactions between the tracers and the independent referee described in 2.1.3 allowed to feedback on the writing

of the HarP instructions improving them where ambiguous definitions could be found or where additional information is necessary to guide the tracer in paying greater attention to potential sources of mistake. For example, the levels of the head and tail were frequent sources of mistakes to be paid particular attention to. Special attention is needed, for example, at the boundary with the most caudal amygdala nuclei and with the indusium griseum or for the inclusion of the vertical digitation, vestigial tissue, and alveus/fimbria at the most caudal level. These interactions allowed us to enrich the protocol of details that better disambiguated hippocampal tissue in 3D.

During the process, the Delphi panelists were asked to check the amendments of the HarP and monitor that the HarP editing did resemble the Delphi panel consensual decisions. The final version of the HarP was then redistributed also to the master tracers.

2.3. Scan selection

Three-dimensional, T1-weighted, structural MRIs and sociodemographic, clinical, and genetic information were taken from the ADNI database (www.loni.ucla.edu/ADNI/Data). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and not-for-profit organizations, as a \$60 million, 5-year, public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as reduce time and costs of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from more than 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55 to 90 years, to participate in the research—approximately 200 cognitively normal elderly individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, please see www.adni-info.org.

As reported by Boccardi et al. [10], for the preliminary phase (Fig. 1), we selected a sample of 20 subjects, four for each degree of severity on the Scheltens visual medial temporal atrophy (MTA) scale [15].

For the training phase (Fig. 1), we selected 10 ADNI subjects with different diagnoses (four controls, three MCI subjects, and three AD subjects) and atrophy severity (two subjects from each atrophy severity score on the MTA scale

[15]). These subjects were different from those included in the preliminary sample.

For the benchmark phase (Fig. 1), 10 new ADNI subjects scanned at both 1.5 and 3 T (for a total sample of 20 images) were selected based on the following criteria:

1. Two subjects from each atrophy severity score on the MTA scale [15] to include the full atrophy range encountered in clinical trials and clinical practice;
2. Subjects for whom both 1.5 and 3 T were available; this allowed to compute a “proxy” of intrarater reliability measure of the segmented images for each master tracer, instead of segmenting and resegmenting the hippocampi on the same images;
3. At least one image from each ADNI scanner manufacturer (Philips Medical Systems, GE Medical Systems, and Siemens);
4. Subjects different from those included in the preliminary and training phases.

2.4. Image processing

A combination of freely available tools was used to preprocess the ADNI MRIs for manual segmentation. Raw MINC (Medical Imaging NetCDF) images were directly downloaded from the ADNI database (www.loni.ucla.edu/adni). Before segmentation, the 3D images were aligned along the line passing through the anterior and posterior commissures of the brain (AC-PC line) using a six-parameter linear registration from the Montreal Neurological Institute (MNI) package AutoReg (version 0.98v; www.bic.mni.mcgill.ca) and the MNI ICBM152 nonlinear symmetric template with $1 \times 1 \times 1$ mm voxel dimensions as the reference. Resampling was carried out with a linear transformation in AutoReg. As our goal was to test hippocampal segmentation in native space, the images were not normalized for intracranial volume differences. No additional preprocessing steps were performed.

2.5. Segmentation procedures

Segmentations were performed using the interactive MultiTracer software (<http://air.bmap.ucla.edu/MultiTracer/>) developed at the Laboratory of Neuro Imaging at the University of California, Los Angeles (Los Angeles, CA, USA). Tracers received detailed instructions describing how to load the images, adjust visualization and segmentation settings, segment the hippocampus, save the segmentation files, and compute volumes. They were required to trace in the coronal view magnified five times, while consulting sagittal view magnified three times and the axial view with no magnification. Magnification was kept constant throughout the segmentation. Segmentations were performed manually from rostral to caudal on approximately 30 to 35 contiguous 1-mm thick coronal brain sections on both sides, with contemporary 3D visualization of the axial and sagittal planes. Tracers were also required to use the same computer and monitor. For the

benchmark labels, tracers were asked to exclude any internal cerebrospinal fluid (CSF) pools through an additional label created in MultiTracer, to comply with the HarP criteria [14]. Tracers were blinded to age, sex, diagnosis, MTA score, field strength, and codes, including correspondence between 1.5 and 3 T MRIs for each subject.

2.6. Volume computation and statistical analyses

Tracers computed hippocampal volumes using the “Frustr Volume” computation in MultiTracer, which multiplies each segmented area by the slice thickness and sums up the obtained volumes. The “Frustr Volume” option was chosen as it resulted in more accurate volume computation by performing volume interpolation when the segmented areas were not perfectly aligned between slices. This function assumes that the structure extends from the center of the first plane on which it was drawn to the center of the last plane with the square root of that segmented areas varying linearly when moving from the center of one plane to the center of the next. The internal CSF total volume was computed as the sum of each CSF segmented area multiplied by the slice thickness. If segmented, it was subtracted from the total hippocampal volume.

Statistical analyses were performed in SPSS software (SPSS Inc., Chicago, IL, USA), version 12.0, and in R language, v.2.13.0. ICCs for both intra- and interrater reliability were computed with a two-way random-effects model. During the benchmark phase, a proxy of intrarater ICCs was computed comparing the segmentations on images at 1.5 T with segmentations on 3 T images of the same subjects. For interrater ICCs, we considered “scan” as a random effect and “tracer” as a random n-level effect, where n = number of tracers (n = 2 for the preliminary phase, n = 3 for the training phase, and n = 5 for the benchmark phase). During the benchmark phase, the more conservative “absolute” model was used. During the preliminary and the training phase, the less stringent and more commonly used “consistency” model was permitted because of the increased difficulty of segmenting multiple subunits of the hippocampus. The “consistency” model is used when systematic differences between tracers are not considered to be relevant, whereas the “absolute” is chosen when systematic differences are considered relevant.

For each hippocampus, the spatial overlapping agreement among benchmark segmentations made by the five tracers was computed with the formula:

$$\text{Similarity coefficient} = \frac{5 |A \cap B \cap C \cap D \cap E|}{(|A| + |B| + |C| + |D| + |E|)},$$

where |A|, |B|, |C|, |D|, and |E| is the number of voxels of the segmented hippocampal regions A, B, C, D, and E, respectively. This formula was adapted from the Dice similarity coefficient [16]. Higher values indicate higher spatial concordance.

Table 1
Reliability of the benchmark labels

	Left hippocampus	Right hippocampus
Proxy intrarater (n = 10)		
Tracer 1	0.981 (0.928–0.995)	0.986 (0.776–0.997)
Tracer 2	0.968 (0.879–0.992)	0.974 (0.902–0.994)
Tracer 3	0.943 (0.335–0.989)	0.968 (0.541–0.994)
Tracer 4	0.966 (0.819–0.992)	0.971 (0.818–0.993)
Tracer 5	0.981 (0.930–0.995)	0.986 (0.944–0.997)
Interrater at 1.5 T (n = 10)		
All 5 tracers	0.957 (0.881–0.988)	0.971 (0.916–0.992)
Interrater at 3 T (n = 10)		
All 5 tracers	0.943 (0.791–0.986)	0.962 (0.863–0.990)

NOTE. Figures denote ICC values (95% confidence interval). Intra- and interrater ICCs were computed with a two-way random-effect “absolute” model. Proxy intrarater denotes agreement of segmentations carried out by the same tracer on scans of the same subjects scanned at 1.5 and 3 T.

The Wilcoxon nonparametric test for repeated measures was used to compare the overlapping coefficient values between 1.5 and 3 T (separately for the left and right hippocampi). The spatial overlapping for CSF internal pools was not computed. To test the variance components of the benchmark sample and to identify the source of variability, we used a three-way analysis of variance (ANOVA) for repeated measures (computed on the 10 subjects), considering the following as “within” factors: “field strength” (two levels), “tracer” (five levels), and “side” (two levels). The coefficients of variation (CVs) of these factors were computed as follows: [(standard deviation of factor)/(mean of hippocampal volumes)] × 100.

3. Results

3.1. Preliminary and training phases

Sociodemographic, clinical, and genetic features of subjects in the preliminary and training phase are listed in [Supplementary Table 1](#).

Both intra- and interrater ICCs for tracers 1 and 2 [10] were above 0.88, for all SUs, except for A/F (0.75). Reliability for the global volume of the “Maximum Hippocampus” (MaxH), computed as the sum of all SUs and using the “Morphology” criterion as the medial border of the hippocampal body [10], was 0.97 (95% CI, 0.92–0.99) and 0.99 (95% CI, 0.98–1) for intrarater ICC and 0.99 (95% CI, 0.96–0.99) for the interrater ICCs ([Supplementary Table 2](#)).

The three new tracers had very high reliability for MaxH: both intra- and interrater ICCs were ≥0.94 ([Supplementary Table 2](#)). Of the individual SUs, the “Maximum Hippocampus” (MinH) and the “Subiculum-Morphology” showed the highest reliability values, whereas the A/F showed the lowest interrater ICC value ([Supplementary Table 2](#)). As previously reported for tracers 1 and 2 [10], also for the other three tracers, the ICCs of the MinH and A/F combined were higher than ICCs of both these SUs alone: intrarater ICC

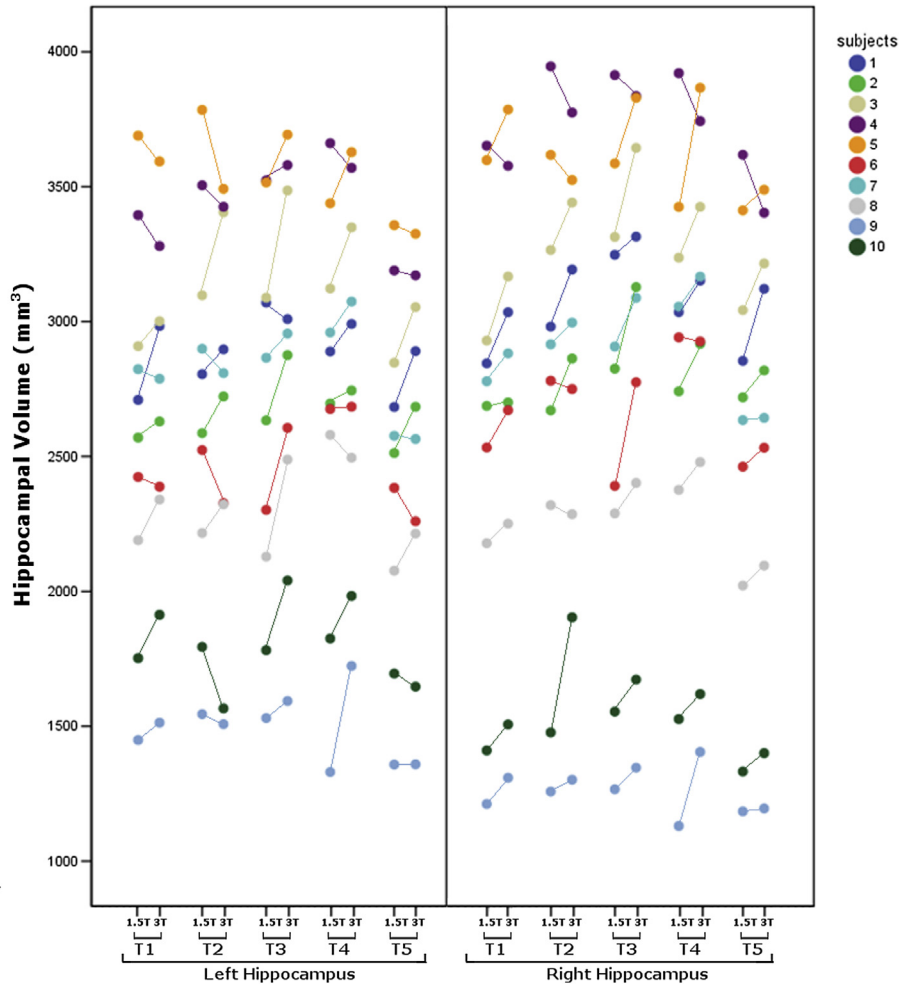


Fig. 2. Sources of variability of the benchmark labels. Plots of hippocampal volumes (y-axis) obtained by the five master tracers in the benchmark phase, by magnetic field strength by side (x-axis). Colors denote individual subjects. The statistics are reported in Table 2. (For interpretation of colors, the reader is referred to the Web version of this article.)

≥ 0.95 (95% CI, 0.83–0.99) and interrater ICC ≥ 0.96 (95% CI, 0.88–0.99).

3.2. Benchmark phase

Sociodemographic, clinical, and genetic features of subjects in the benchmark phase are listed in Supplementary Table 1. 1.5 and 3T images were obtained using the same type of manufacturer scanner for each but two subjects.

The proxy intrarater ICCs computed between 1.5 and 3 T images of the same subjects ranged from 0.943 to 0.986 (Table 1, Fig. 2, and Supplementary Fig. 1). Interrater ICCs were 0.957 for the left and 0.971 for the right hippocampus at 1.5 T and 0.943 for the left and 0.962 for the right hippocampus at 3 T (Supplementary Fig. 2).

Similarity coefficients among the five tracers were 0.73 (standard deviation [SD], 0.04) for both the left and right hippocampi at 1.5 T and 0.74 (SD, 0.03) and 0.76 (SD, 0.03) for the left and right hippocampi at 3 T (Fig. 3). Similarity coefficients were higher at 3 T than 1.5 T (Wilcoxon test exact $P = .049$ for the left and $P = .004$ for the right). See

Fig. 4 for two exempla of the simultaneous mapping of the corrected benchmark hippocampal segmentations of the five master tracers.

Using a three-way ANOVA for repeated measures model, we found a statistical significance for the factor “tracer” ($P = .004$) but not for “side” ($P = .202$), nor “field strength” ($P = .335$). Moreover, through the variance decomposition analysis, we found that the variance due to “subject” explained the 96% of the whole variance of the model, whereas “tracer” contributed only for 0.25% (Table 2). The variance due to “subject” did not differ from the whole variance of the model [$F(199, 9)$, $P = .525$], whereas the variance due to the other factors did [“field strength”, $F(199, 1)$, $P = .024$; “tracer”, $F(199, 4)$, $P < .001$; “side”, $F(199, 1)$, $P = .025$]. The variance due to “subject” significantly differed from “field strength” [$F(9, 1)$, $P = .024$], “tracer” [$F(9, 4)$, $P < .001$], and “side” [$F(9, 1)$, $P = .025$]. Variance due to “tracer” did not differ from the “field strength” one [$F(4, 1)$, $P = .423$] nor from the “side” one [$F(4, 1)$, $P = .435$]; interestingly, it differed from the residual error variance [$F(184, 4)$, $P = .009$] (Table 2). The highest CV was for

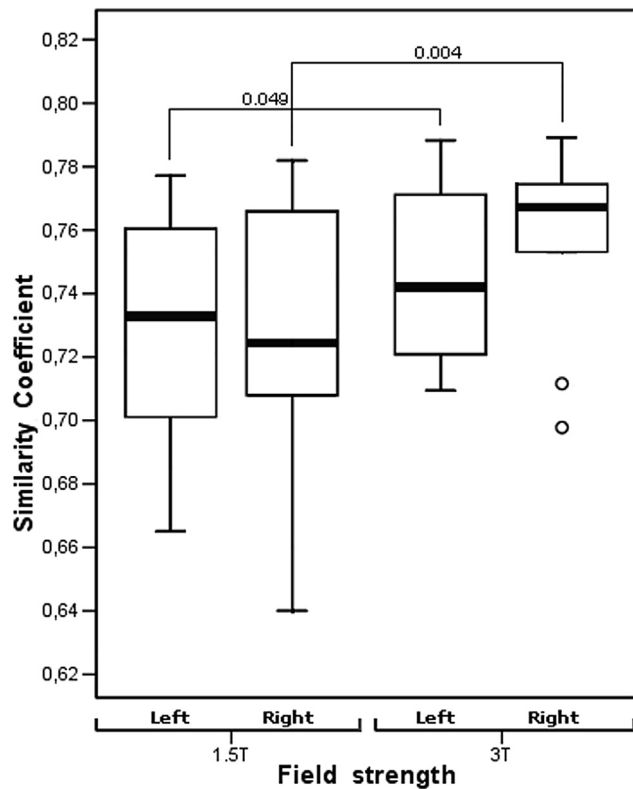


Fig. 3. Similarity coefficients of the benchmark labels. Similarity coefficients denote spatial overlap among the five master tracers. Upper and lower box boundaries represent 25th and 75th percentiles of the distribution, lines represent the mean, whiskers represent extreme data (i.e., the highest and lowest values that are not outliers), and circles represent outliers. Coefficients were significantly higher at 3 T than 1.5 T (Wilcoxon exact test).

the factor “subject,” whereas for “tracer,” it was 1.4% (Table 2).

4. Discussion

In the present work, we have described a necessary methodological step for the definition of standard operating procedures for hippocampal volumetry. We have generated 200 benchmark labels of the hippocampus based on the EADC-ADNI HarP, as each of the five different tracers provided labels for both hippocampi of the same 10 ADNI subjects, scanned at both 1.5 and 3 T. In the next step of the project, these benchmark labels will be uploaded onto a web platform and used as the reference for the qualification of the naïve tracers who will take part to the final validation of the HarP [17].

To generate a certified set of benchmark HarP segmentations as a standard reference, the segmented labels underwent quality check and corrections when even minimum divergences from the HarP were observed. Notwithstanding the consolidated experience of tracers and their strict training, the complexity of instructions and hippocampal morphology did require corrections in some points, and

the presence of these divergences was used to further clear up the instructions provided in the protocol.

A very high reliability was observed within and among the five tracers, even before the quality check and corrections were performed (data not shown). However, the present work was not aimed to demonstrate the reliability of the HarP, which will be tested with the proper design in the subsequent phase of the project [18]. We underline that the reported reliability measures are proposed here only to describe the delivered benchmark labels and not to provide demonstration of HarP validity, which requires of course a different experimental design. However, these high reliability data (absolute ICCs ≥ 0.94 and five-level similarity coefficient ≥ 0.73) may be considered as a very preliminary and proxy measure of protocol reliability. These data give us a hint of confidence that the segmentation criteria and landmark descriptions in the HarP are sufficiently well defined to be used as a reliable standard worldwide. Under this respect, it is worth to underline that master tracers were recruited from different centers—two in Europe (Italy and Germany) and three in North America (Los Angeles, CA, Rochester, MN, and Quebec City, Canada)—and were originally familiar with different segmentation protocols.

In exploratory ANOVA sources, the volumetric variability of the benchmark labels due to tracer was substantially lower than anatomic variability, much lower than residual error variance, and of similar magnitude as the variability due to side and field strength. It should be underlined that the CV due to factors such as interlaboratory procedures (i.e., analytical kits, assay lots) for measurement of CSF biomarkers (A β 42, tau, and p-tau) ranges between 9% and 30% [19,20]. This is more than 6.5-fold greater than the variability due to HarP tracers in this study. Although only preliminary and needing quantification with the proper experimental design [18], this is a notable finding considering that measurements were produced through manual segmentation.

Besides reorienting along the AC-PC line, no additional image preprocessing was performed to avoid any use of arbitrary procedures. Indeed, segmentations on 3 T images displayed significantly higher similarity coefficient than the ones on 1.5 T, suggesting a role of the image contrast on segmentation stability. However, the definition of specific preprocessing possibly improving the hippocampal segmentation should be covered in further projects specifically aimed at identifying an optimal procedure.

As a part of a larger project, this work has some limitations. One limitation consists in the fact that we computed a proxy of intrarater ICC for the benchmark images, resegmenting the same subjects on their different 1.5 and 3 T scans, rather than on the same scans. However, the variance due to field strength was quantified in 0.1% of the whole variance, which suggests that this method can represent a reliable proxy of the correct quantification of intrarater reliability.

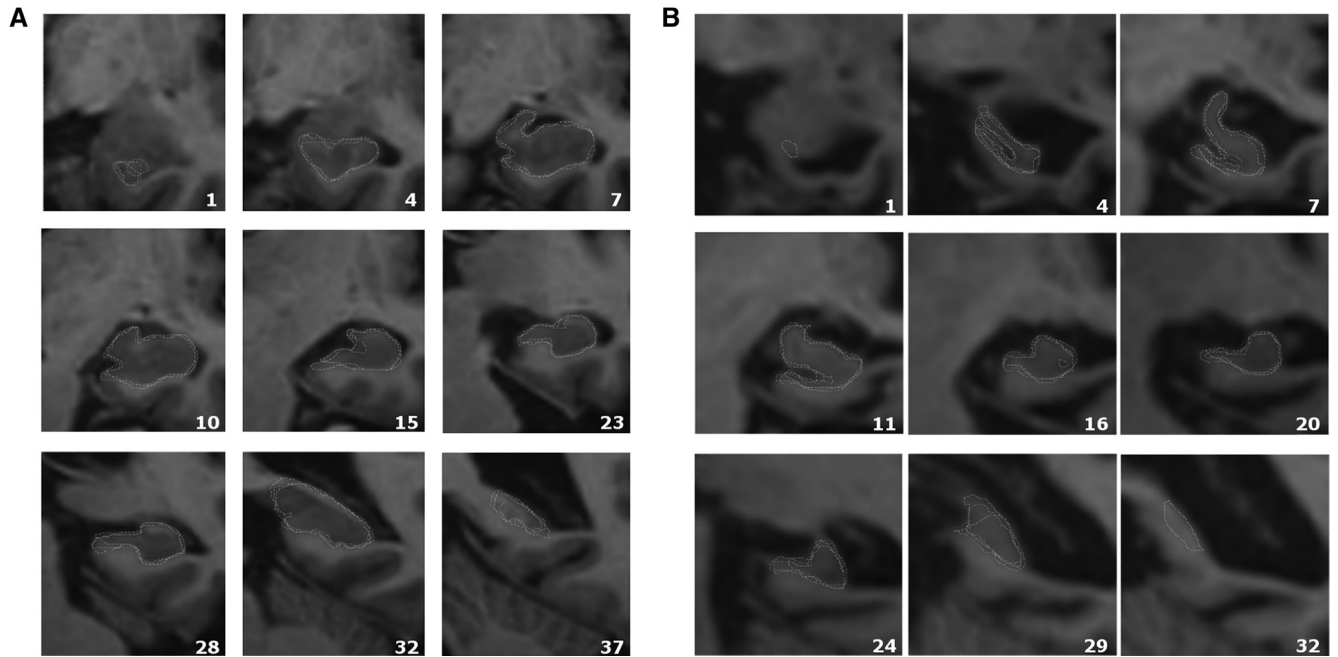


Fig. 4. Selection of color-coded benchmark label segmentations. For illustrative purposes, scans were selected where similarity coefficient was highest (panel A: 0.789; right hippocampus of a control subject, 3 T MRI) and lowest (panel B: 0.640; right hippocampus of an AD subject, 1.5 T MRI). Slices were selected on the following basis: (1) the first rostral slice where the hippocampus was segmented by at least one master tracer (number 1), (2) the last caudal slice where the hippocampus was segmented by at least one master tracer (number = 37 and 32), and (3) the central ones. As to the central ones, we chose equally distant slices whenever possible. High resolution figures of all slices are available at: http://centroalzheimier.it/public/MB/SOPs/PaperBenchmark/Figure4a_HIGHEST_hr.tif and http://centroalzheimier.it/public/MB/SOPs/PaperBenchmark/Figure4b_LOWEST_hr.tif.

The overlapping coefficients may be considered relatively low, and the inferior 95% confidence limit was low for some ICC values for the benchmark labels. This is due to the following reasons. First, we aimed to obtain benchmark segmentations that were compliant with the HarP but that also represented the variability that can be observed among HarP-compliant segmentations and that should be accepted in the learning of future tracers. Indeed, this variability in HarP-compliant segmentations is due to the complexity of the identification of hippocampal boundaries from MRI without correspondent histologic determination. Second, we sought volume reliability values through the most restrictive comparison of *absolute volumes* rather than on correla-

tions, an approach that is very sensitive to minimal differences. Finally, we computed the overlapping values across five rather than two tracers, adapting the Dice similarity coefficient [16]. For this latter case, the lack of a similar approach in the literature as well as of ranges of acceptability prevents the comparison of these results with previous ones. A generalized metric could be used [21] but conceptually would provide the same information. A final limitation was that when measuring spatial agreement, we did not exclude the internal CSF pools as we did for crude volumes.

In this phase, we produced a major product of the EADC-ADNI harmonization project, that is, the benchmark hippocampal labels representing the HarP criteria as defined by the Delphi panel [14]. These labels are the concrete standard reference for the future definition of qualification criteria of human tracers and automated algorithms. These will be based on not only the benchmark labels themselves but also statistics of new tracers' performance [17].

The next phase of the project will consist of the training of human tracers who will be involved in the validation of the HarP [18] and the expansion of this set of certified benchmark labels, to provide a wider representation of physiological variability and offer an adequate training set for automated algorithms ([22] and www.hippocampal-protocol.net).

The benchmark labels may serve hippocampal segmentation training not only for the AD field but possibly for other fields where hippocampal volumetry is an informative feature since the HarP definition is sufficiently inclusive to

Table 2
Statistics of the sources of variability of the benchmark labels

Factor	CV (%)	ANOVA		
		Variance; df	% of variance	P
Tracer	1.4	1435; 4	0.25	–
Subject	27.5	546,026; 9	96	.00002
Field strength	0.9	527; 1	0.1	.423
Side	0.9	563; 1	0.1	.435
Residual error	5.3	20,469; 184	3.55	.009

Abbreviations: CV, coefficient of variation; ANOVA, analysis of variance; df, degree of freedom.

NOTE. Figures denote CV, amount of variance, df, and percentage of variance represented by each factor. P denotes the difference of variance of the factor “tracer” versus all other factors (three-way ANOVA for repeated measures model). The data this statistics refer to are shown in Fig. 2.

be adapted to different aims. Of course, the use of the HarP would need to be specifically validated for these different fields.

Acknowledgments

The Alzheimer's Association has provided logistic support for the update meetings of the project. Wyeth, part of the Pfizer group, and Lilly have provided unrestricted grants in support of the work reported in this article. A follow-up project has been funded by the Alzheimer's Association: "A Harmonized Protocol for Hippocampal Volumetry: an EADC-ADNI Effort," grant IIRG -10-174022. Part of [Supplementary Table 2](#) was reprinted from "Alzheimer's & Dementia, 10.1016/j.jalz.2013.03.001, Boccardi et al., Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation, Copyright 2013, with permission from Elsevier."

The project principal investigator (PI) is Giovanni B. Frisoni, IRCCS Fatebenefratelli, Brescia, Italy; the co-PI is Clifford R. Jack, Mayo Clinic, Rochester, MN; the Statistical Working Group is led by Simon Duchesne, Laval University, Quebec City, Canada; and the project coordinator is Marina Boccardi, IRCCS Fatebenefratelli, Brescia, Italy. EADC centers (local PI) are: IRCCS Fatebenefratelli, Brescia, Italy (G.B. Frisoni); University of Kuopio and Kuopio University Hospital, Kuopio, Finland (H. Soininen); Hôpital Salpêtrière, Paris, France (B. Dubois and S. Lehericy); University of Frankfurt, Frankfurt, Germany (H. Hampel); University Rostock and DZNE, Rostock, Germany (S. Teipel); Karolinska Institutet, Stockholm, Sweden (L.-O. Wahlund); Department of Psychiatry Research, Zurich, Switzerland (C. Hock); Alzheimer Centre, Vrije Univ Medical Centre, Amsterdam, The Netherlands (F. Barkhof and P. Scheltens); Dementia Research Group Institute of Neurology, London, United Kingdom (N. Fox); Dep. of Psychiatry and Psychotherapy, University Medical Center, Mainz (A. Fellgiebel); and NEUROMED, Department of Neuroimaging, King's College London, London, United Kingdom (A. Simmons). Alzheimer's Disease Neuroimaging Initiative (ADNI) centers are: Mayo Clinic, Rochester, MN (C.R. Jack); University of California, Davis, CA (C. DeCarli and C. Watson); University of California, Los Angeles (UCLA), CA (G. Bartzokis); University of California, San Francisco (UCSF), CA (M. Weiner and S. Mueller); Laboratory of Neuro Imaging, UCLA, CA (L.G. Apostolova); University of Southern California, Los Angeles, CA (P.M. Thompson); Rush University Medical Center, Chicago, IL (L. deToledo-Morrell); Rush Alzheimer's Disease Center, Chicago, IL (D. Bennet); Northwestern University, IL (J. Csernansky); Boston University School of Medicine, MA (R. Killiany); John Hopkins University, Baltimore, MD (M. Albert); Center for Brain Health, New York, NY (M. De Leon); and Oregon Health & Science University, Portland, OR (J. Kaye). Other centers are: McGill University, Montreal, Quebec, Canada (J. Pruessner); University of Alberta, Edmonton, AB, Canada (R. Camicioli and N. Malykhin);

Department of Psychiatry, Psychosomatic, Medicine & Psychotherapy, Johann, Wolfgang Goethe-University, Frankfurt, Germany (J. Pantel); and Institute for Ageing and Health, Wolfson Research Centre, Newcastle General Hospital, Newcastle, United Kingdom (J. O'Brien). Population-based study participants: PATH through life, Australia (P. Sachdev and J.J. Maller); SMART-Medea Study, University Medical Center Utrecht, The Netherlands (M.I. Geerlings); Rotterdam Scan Study, the Netherlands (T. denHeijer); L. Launer, National Institute on Aging, Bethesda, and W. Jagust, University of California, Berkeley, CA. Statistical Working Group: AFAR (Fatebenefratelli Association for Biomedical Research) San Giovanni Calibita-Fatebenefratelli Hospital, Rome, Italy (P. Pasqualetti); Laval University, Quebec City, Canada (S. Duchesne); Montreal Neurological Institute, McGill University, Montreal, Canada (L. Collins). Advisors for clinical issues: P.J. Visser, Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, the Netherlands; EADC and ADNI PIs: B. Winbald, Karolinska Institute, Sweden and L. Froelich, Central Institute of Mental Health, Mannheim, Germany; M. Weiner, University of California, San Francisco (UCSF), CA. Dissemination and education: G. Waldemar, Copenhagen University Hospital, Copenhagen, Denmark.

The MRIs used in this article belong to the ADNI data set. Data collection and sharing for this project was funded by the ADNI (National Institutes of Health grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd; AstraZeneca; Bayer HealthCare; BioClinica, Inc; Biogen Idec Inc; Bristol-Myers Squibb Company; Eisai Inc; Elan Pharmaceuticals Inc; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc; GE Healthcare; Innogenetics, N.V.; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development, LLC; Medpace, Inc; Merck & Co., Inc; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer Inc; Servier; Synarc Inc; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is Rev March 26, 2012, 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 California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

N. Robitaille and S. Duchesne have received funding support from the Ministère du Développement Économique and de

l'Innovation et de l'Exportation du Québec. S. Duchesne is supported by a Fonds de Recherche Québec-Santé Junior 1 Research Scholar award.

Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.jalz.2013.12.019>.

RESEARCH IN CONTEXT

1. Systematic review: To date, manual segmentation on magnetic resonance images is the widely accepted *in vivo* gold standard for hippocampal volumetry. However, the availability and widespread use of a large number of different protocols for manual segmentation across laboratories has resulted in highly heterogeneous hippocampal volumetric estimates, preventing comparisons among different studies and outcomes of clinical trials for Alzheimer's disease. A globally harmonized protocol for manual hippocampal segmentation based on magnetic resonance has been recently defined by a panel of international experts.
2. Interpretation: Here, we developed 200 benchmark hippocampal segmentations manually traced by five experts. These labels were certified as reflecting the landmarks and segmentation modalities defined in the harmonized protocol criteria.
3. Future directions: This benchmark set of hippocampal labels will be used as the gold standard for training and qualification of future tracers.

References

- [1] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol* 2007; 6:734–46.
- [2] Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, et al. Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol* 2010;11:1118–27.
- [3] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:263–9.
- [4] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:280–92.
- [5] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:270–9.
- [6] Committee for Medicinal Products for Human Use (CHMP). Qualification opinion of low hippocampal volume (atrophy) by MRI for use in clinical trials for regulatory purpose—in pre-dementia stage of Alzheimer's disease. EMA/CHMP/SAWP/809208/2011. 17 November 2011.
- [7] Frisoni GB, Fox NC, Jack CR Jr, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* 2010; 6:67–77.
- [8] Geuze E, Vermetten E, Bremner JD. MR-based *in vivo* hippocampal volumetrics: 1. review of methodologies currently employed. *Mol Psychiatry* 2005;10:147–59.
- [9] Konrad C, Ukas T, Nebel C, Arolt V, Toga AW, Narr KL. Defining the human hippocampus in cerebral magnetic resonance images—an overview of current segmentation protocols. *Neuroimage* 2009;47:1185–95.
- [10] Boccardi M, Bocchetta M, Ganzola R, Robitaille N, Redolfi A, Duchesne S, et al. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. *Alzheimers Dement* 2015; 11:184–94.
- [11] Frisoni GB, Jack CR. Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. *Alzheimers Dement* 2011;7:171–4.
- [12] Jack CR Jr, Barkhof F, Bernstein MA, Cantillon M, Cole PE, DeCarli C, et al. Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. *Alzheimers Dement* 2011;7:474–4854.
- [13] Boccardi M, Ganzola R, Bocchetta M, Pievani M, Redolfi A, Bartzokis G, et al. Survey of protocols for the manual segmentation of the hippocampus: preparatory steps towards a joint EADC-ADNI harmonized protocol. *J Alzheimers Dis* 2011;26(Suppl 3):61–75.
- [14] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi Definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. *Alzheimers Dement* 2015;11:126–38.
- [15] Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in “probable” Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *J Neurol Neurosurg Psychiatry* 1992;55:967–72.
- [16] Robitaille N, Duchesne S. Label fusion strategy selection. *Int J Biomed Imaging* 2012;2012:431095.
- [17] Duchesne S, Valdivia F, Robitaille N, Abiel Valdivia F, Bocchetta M, Boccardi M, et al. Manual segmentation qualification platform for the EADC-ADNI harmonized protocol for hippocampal segmentation project. *Alzheimers Dement* 2015;11:161–74.
- [18] Frisoni GB, Jack CR Jr, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, et al. The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimers Dement* 2015;11:111–25.
- [19] Mattsson N, Andreasson U, Persson S, Carrillo MC, Collins S, Chalbot S, et al. CSF biomarker variability in the Alzheimer's Association Quality Control Program. *Alzheimers Dement* 2013;9:251–61.
- [20] Verwey NA, van der Flier WM, Blennow K, Clark C, Sokolow S, De Deyn PP, et al. A worldwide multicentre comparison of assays for cerebrospinal fluid biomarkers in Alzheimer's disease. *Ann Clin Biochem* 2009;46:235–40.
- [21] Crum WR, Camara O, Hill DL. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging* 2006;25:1451–61.
- [22] Boccardi M, Bocchetta M, Nishikawa M, Ganzola R, Grothe M, Wolf D, et al. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimers Dement* 2015;11:175–83.

Supplementary Table 1
Sociodemographic and clinical features

Variable	MTA scale				
	0	1	2	3	4
Preliminary phase (n = 20)*					
Age (y)	72, 74, 77, 78	70, 74, 76, 83	64, 72, 79, 80	74, 79, 80, 85	72, 80, 82, 83
Sex	M, M, F, F	M, M, M, M	M, M, F, F	M, M, M, M	M, M, F, F
APOE genotype	ε23, ε33, ε33, ε34	ε33, ε33, ε34, ε34	ε34, ε34, ε34, ε44	ε33, ε34, ε34, ε44	ε23, ε33, ε33, ε24
Diagnosis, CTRL/MCI/AD	4/0/0	4/0/0	0/4/0	0/3/1	0/2/2
Training phase (n = 10)					
Age (y)	69, 77	73, 79	77, 79	72, 75	74, 79
Sex	M, F	M, M	M, F	M, F	M, M
APOE genotype	ε33, ε44	ε44, ε44	ε33, ε34	ε34, ε34	ε33, ε34
Diagnosis, CTRL/MCI/AD	1/0/1	0/2/0	2/0/0	0/0/2	1/1/0
Benchmark phase (n = 10)					
Age (y)	69, 73	74, 85	73, 77	80, 83	77, 83
Sex	F, F	M, F	M, M	M, M	M, F
APOE genotype	ε23, ε33	ε23, ε23	ε33, ε44	ε33, ε34	ε33, ε34
Diagnosis, CTRL/MCI/AD	1/1/0	2/0/0	1/1/0	0/2/0	0/1/1








Abbreviations: MTA, medial temporal atrophy; CTRL, healthy control subjects; AD, Alzheimer's disease; MCI, mild cognitive impairment; APOE, Apolipoprotein E.

NOTE. Magnetic resonance scans segmented in the preliminary (n = 20 at 1.5 T, also reported in [13]), training (n = 10 at 1.5 T), and benchmark phases (n = 10 at both 1.5 and 3 T). Scans were balanced by the severity of MTA [15]. Tracers 1 and 2 were trained on the preliminary phase sample; tracers 3, 4, and 5 were trained on the training phase sample; and all tracers segmented the benchmark sample (see Fig. 1).

*Reported in [13].

Supplementary Table 2

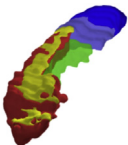



Reliability results of the master tracers in the preliminary (tracers 1 and 2) and training phase (tracers 3, 4, and 5)

		Intrarater (n = 10)					Interrater (n = 10)	
		Left hippocampus						
		Tracer 1*	Tracer 2*	Tracer 3	Tracer 4	Tracer 5	Tracers 1 and 2*	Tracers 3, 4, and 5
		Preliminary phase*		Training phase			Preliminary phase*	Training phase
	MinH	0.934 (0.841–0.973)	0.992 (0.980–0.997)	0.989 (0.955–0.997)	0.964 (0.863–0.991)	0.990 (0.962–0.998)	0.974 (0.936–0.990)	0.918 (0.783–0.977)
	Alveus/fimbria	0.748 (0.466–0.892)	0.863 (0.687–0.944)	0.961 (0.851–0.990)	0.861 (0.538–0.964)	0.906 (0.669–0.976)	0.885 (0.734–0.953)	0.854 (0.640–0.958)
	Subiculum - Oblique line	0.878 (0.719–0.950)	0.964 (0.911–0.985)	0.963 (0.859–0.991)	0.937 (0.770–0.984)	0.932 (0.753–0.983)	0.907 (0.781–0.962)	0.797 (0.528–0.939)
	Subiculum - Morphology	0.921 (0.811–0.968)	0.981 (0.952–0.992)	0.980 (0.923–0.995)	0.923 (0.724–0.980)	0.933 (0.755–0.983)	0.937 (0.848–0.975)	0.839 (0.610–0.953)
	Subiculum - Horizontal line	0.925 (0.822–0.970)	0.980 (0.951–0.992)	0.975 (0.905–0.994)	0.895 (0.636–0.973)	0.906 (0.668–0.976)	0.932 (0.836–0.972)	0.843 (0.617–0.954)
	Tail Crura	0.993 (0.982–0.997)	0.998 (0.994–0.999)	0.980 (0.921–0.995)	0.872 (0.568–0.967)	0.995 (0.980–0.999)	0.937 (0.847–0.974)	0.740 (0.429–0.919)
	Tail End	0.972 (0.931–0.989)	0.988 (0.971–0.995)	0.984 (0.935–0.996)	0.857 (0.526–0.962)	0.929 (0.740–0.982)	0.905 (0.775–0.961)	0.794 (0.523–0.938)

(Continued)

Supplementary Table 2





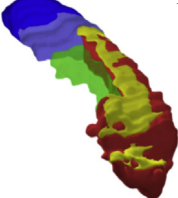
Reliability results of the master tracers in the preliminary (tracers 1 and 2) and training phase (tracers 3, 4, and 5) (*Continued*)

		Intrarater (n = 10)					Interrater (n = 10)	
		Left hippocampus						
		Tracer 1*	Tracer 2*	Tracer 3	Tracer 4	Tracer 5	Tracers 1 and 2*	Tracers 3, 4, and 5
		Preliminary phase*		Training phase			Preliminary phase*	Training phase
	MaxH	0.967 (0.919–0.987)	0.993 (0.982–0.997)	0.998 (0.991–0.999)	0.955 (0.830–0.989)	0.983 (0.932–0.996)	0.985 (0.963–0.994)	0.963 (0.895–0.990)
		Right hippocampus						
		Tracer 1*	Tracer 2*	Tracer 3	Tracer 4	Tracer 5	–	Tracers 3, 4 and 5
		Preliminary phase*		Training phase			–	Training phase
	MinH	–	–	0.991 (0.966–0.998)	0.969 (0.881–0.992)	0.983 (0.934–0.996)	–	0.927 (0.804–0.979)
	Alveus/ fimbria	–	–	0.969 (0.881–0.992)	0.843 (0.489–0.958)	0.950 (0.812–0.987)	–	0.588 (0.212–0.861)
	Subiculum - Oblique line	–	–	0.937 (0.770–0.984)	0.915 (0.696–0.978)	0.760 (0.291–0.935)	–	0.806 (0.545–0.942)

(Continued)

Supplementary Table 2

Reliability results of the master tracers in the preliminary (tracers 1 and 2) and training phase (tracers 3, 4, and 5) (*Continued*)

		Right hippocampus						
		Tracer 1*	Tracer 2*	Tracer 3	Tracer 4	Tracer 5	–	Tracers 3, 4 and 5
		Preliminary phase*		Training phase			–	Training phase
	Subiculum - Morphology	–	–	0.976 (0.905–0.994)	0.957 (0.838–0.989)	0.879 (0.588–0.969)	–	0.877 (0.690–0.965)
	Subiculum - Horizontal line	–	–	0.982 (0.929–0.995)	0.922 (0.718–0.980)	0.888 (0.613–0.971)	–	0.777 (0.492–0.932)
	Tail Crura	–	–	0.995 (0.981–0.999)	0.893 (0.629–0.972)	0.980 (0.921–0.995)	–	0.825 (0.582–0.948)
	Tail End	–	–	0.980 (0.923–0.995)	0.943 (0.788–0.986)	0.990 (0.960–0.998)	–	0.770 (0.480–0.930)
	MaxH	–	–	0.998 (0.990–0.999)	0.947 (0.801–0.986)	0.982 (0.930–0.996)	–	0.936 (0.828–0.982)

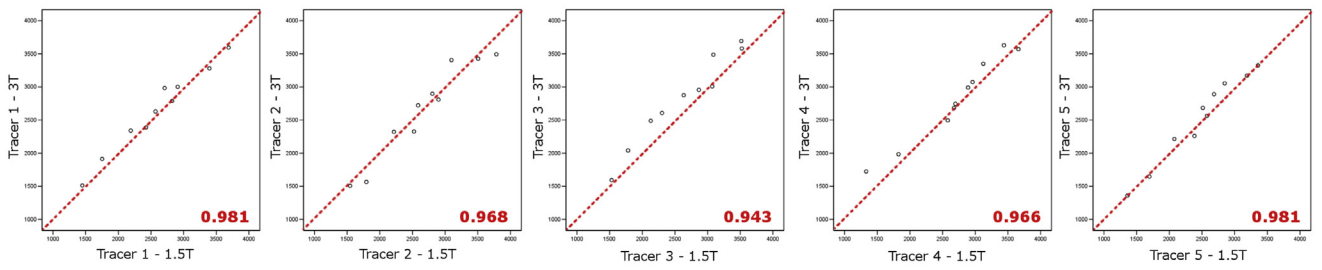
Abbreviations: ICC, intraclass correlation coefficient; MinH, minimum hippocampus; MaxH, maximum hippocampus.

NOTE. Figures denote ICC values (95% confidence interval). Intra- and interrater ICCs were computed with a two-way random-effect “consistency” model. MaxH is a derived summed measure from MinH, Alveus/Fimbria, Subiculum – Morphology, Crura and Tail End [13].

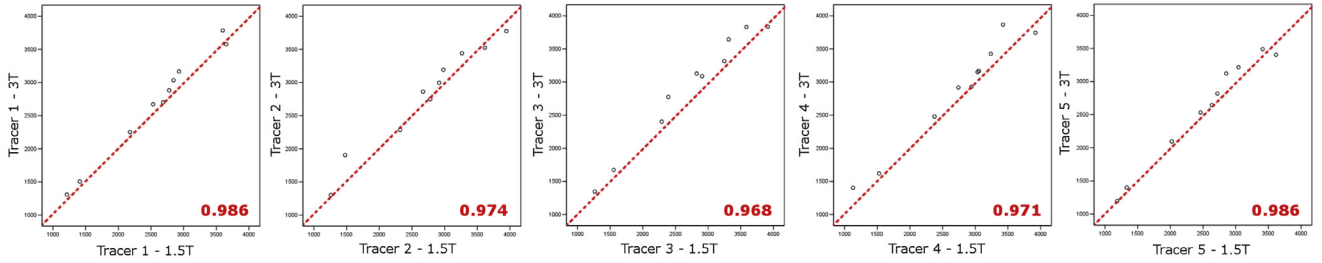
*Data published in [13].

Reprinted from *Alzheimer's & Dementia*, 10.1016/j.jalz.2013.03.001, Boccardi et al., Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation, Copyright 2013, with permission from Elsevier.

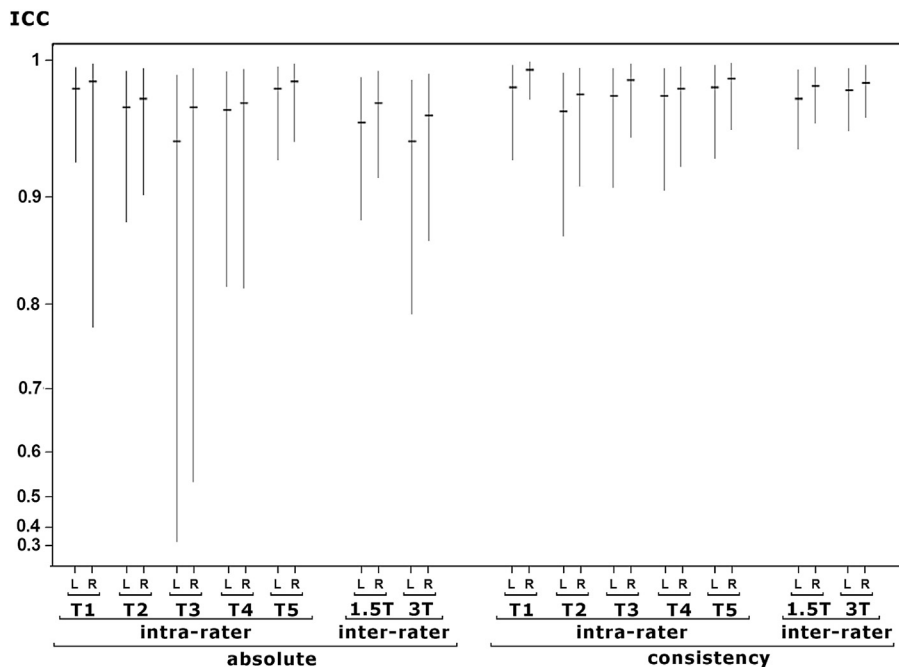
A Left Hippocampal Volume



B Right Hippocampal Volume



Supplementary Fig. 1. Intrarater agreement of the benchmark labels. Plots of left (upper line, A) and right (lower line, B) hippocampal volumes of the five master tracers in the benchmark phase at 1.5 and 3 T. Red dotted lines denotes perfect agreement; numbers denote intrarater “absolute” ICCs.



Supplementary Fig. 2. Intra- and interrater agreement of the five master tracers for the benchmark labels. “Absolute” and “consistency” ICCs are shown with 95% confidence intervals.