ELSEVIER

Alzheimer's & Dementia ■ (2015) 1-9

Alzheimer's & Dementia

Research Article

Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol

Marina Boccardi[a,*], Martina Bocchetta[a,b], Félix C. Morency[c,d], D. Louis Collins[e], Masami Nishikawa[f], Rossana Ganzola[c], Michel J. Grothe[g], Dominik Wolf[h], Alberto Redolfi[a], Michela Pievani[a], Luigi Antelmi[a,i], Andreas Fellgiebel[h], Hiroshi Matsuda[f], Stefan Teipel[g,j], Simon Duchesne[c], Clifford R. Jack, Jr.,[k], Giovanni B. Frisoni[a,i], for the European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging initiative (ADNI) Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation*, and for the Alzheimer's Disease Neuroimaging Initiative**

[a]*LENITEM (Laboratory of Epidemiology, Neuroimaging and Telemedicine) IRCCS – Centro S. Giovanni di Dio – Fatebenefratelli, Brescia, Italy*
[b]*Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy*
[c]*Department of Radiology, Université Laval and Centre de Recherche de l'Institut universitaire de santé mentale de Québec, Quebec City, Canada*
[d]*Imeka, Sherbrooke, Québec, Canada*
[e]*McConnell Brain Imaging Center, Montreal Neurological Institute, McGill University, Montreal, Canada*
[f]*Kawamura Gakuen Woman's University, Abiko-city, Japan*
[g]*German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*
[h]*Klinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität, Mainz, Germany*
[i]*University Hospitals and University of Geneva, Geneva, Switzerland*
[j]*Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany*
[k]*Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN, USA*

Abstract

Background: The European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative (ADNI) Harmonized Protocol (HarP) is a Delphi definition of manual hippocampal segmentation from magnetic resonance imaging (MRI) that can be used as the standard of truth to train new tracers, and to validate automated segmentation algorithms. Training requires large and representative data sets of segmented hippocampi. This work aims to produce a set of HarP labels for the proper training and certification of tracers and algorithms.

Methods: Sixty-eight 1.5 T and 67 3 T volumetric structural ADNI scans from different subjects, balanced by age, medial temporal atrophy, and scanner manufacturer, were segmented by five qualified HarP tracers whose absolute interrater intraclass correlation coefficients were 0.953 and 0.975 (left and right). Labels were validated as HarP compliant through centralized quality check and correction.

Results: Hippocampal volumes (mm^3) were as follows: controls: left = 3060 (SD 502), right = 3120 (897); mild cognitive impairment (MCI): left = 2596 (447), right = 2686 (473); and Alzheimer's disease (AD): left = 2301 (492), right = 2445 (525). Volumes significantly correlated with atrophy severity at Scheltens' scale (Spearman's $\rho = <-0.468$, $P = <.0005$).

The manuscript has been approved by the ADNI Data and Publication Committee on November 14, 2013.

*Please see www.hippocampal-protocol.net for the complete list of EADC-ADNI Harmonized Protocol for Manual Hippocampal Segmentation Investigators.

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

*Corresponding author. Tel.: ■■■; Fax: ■■■.
E-mail address: mboccardifbf@gmail.com

http://dx.doi.org/10.1016/j.jalz.2014.12.002

Cerebrospinal fluid spaces (mm^3) were as follows: controls: left = 23 (32), right = 25 (25); MCI: left = 15 (13), right = 22 (16); and AD: left = 11 (13), right = 20 (25). Five subjects (3.7%) presented with unusual anatomy.

Conclusions: This work provides reference hippocampal labels for the training and certification of automated segmentation algorithms. The publicly released labels will allow the widespread implementation of the standard segmentation protocol.

Keywords: Harmonized protocol; Benchmark images; Automated segmentation algorithms; Algorithm training; MRI; Hippocampus; Hippocampal segmentation

1. Introduction

Between the years 2008 and 2013, a joint European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) effort was carried out to provide a consensual, harmonized protocol (HarP) for the manual segmentation of the whole hippocampus on magnetic resonance imaging (MRI). The protocol was defined through an evidence-based Delphi panel that converged on a consensus definition based on personal experience, evaluation of a common set of ad hoc data [1–3], and recursive re-evaluation of choices expressed by other panelists and justifications thereof [4]. The panel converged on a most inclusive definition of the outer hippocampal boundaries, where the whole hippocampal head, body, and tail are included in the segmentation, together with the alveus, fimbria, and both Andreas Retzius and the fasciolar gyri. The HarP has been validated in three different phases. First, its concurrent validity was compared against local protocols [5]. Results showed significant increase of absolute interrater intraclass correlation coefficients (ICCs) between tracers segmenting based on the HarP rather than on local protocols. Analysis of variance (ANOVA) denoted a very limited effect of tracer (0.9% of the total variance) in the use of HarP segmentations, corresponding to a very small coefficient of variation (2.4%). This method-related variance is notably smaller compared with coefficients of variation observed for other Alzheimer's disease (AD) biomarkers, ranging between 13% and 36% and more [5–7]. The HarP has finally been validated versus pathological evaluation, in a study with 7T post-mortem MRI where HarP hippocampal volumes correlated consistently with Braak and Braak stages and pertinent AD pathology [8].

To the purpose of the EADC-ADNI harmonized hippocampal segmentation project, benchmark segmentations (i.e., hippocampal segmentations proposed as a concrete standard reference and certified to resemble all the HarP criteria) were produced by professionals with previous experience in hippocampal segmentation who received further specific training on the HarP [9]. Segmentations were uploaded on a web-platform designed to help in training an independent group of tracers [10,11] who would take part in the validation of the HarP [5].

Although segmentations with the HarP proved to be very reliable between tracers, with reliability values of up to 0.90 for absolute interrater, and up to 0.99 for absolute intrarater ICCs [5,9], manual hippocampal segmentation remains a time-consuming task and an impractical one to be used in clinical routine or large scientific image data sets. However, not unlike new or naïve human raters, most algorithms require a sizable sample of segmented hippocampi, representative of physiological and pathological variability and technical factors (field strength, scanner manufacturer), to learn exemplars and properly generalize the knowledge of hippocampal boundaries to new subjects. The design of the EADC-ADNI harmonized hippocampal segmentation project required a very limited number of subjects for its full validation (n = 26 ADNI subjects in total), and only 10 subjects to generate the initial benchmark labels, far too low for algorithm training. This work was aimed to provide benchmark hippocampal segmentations based on the HarP for a large sample of hippocampi with an appropriate balance of key image analysis factors such as age, dementia severity, field strength, and scanner manufacturer.

2. Methods

2.1. Images

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). 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 nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. The 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, and lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of

Table 1
Frequencies for scanner, diagnosis, and age bins

	1.5 T (N = 68)			3 T (N = 67)		
Scanner	Siemens	GE	Philips	Siemens	GE	Philips
N	23	24	21	23	22	22
Diagnosis	CTRL	MCI	AD	CTRL	MCI	AD
N	22	24	22	22	22	23
Age	60–70	70–80	80+	60–70	70–80	80+
N	19	30	19	21	25	21

Abbreviations: CTRL, controls; MCI, mild cognitive impairment; AD, Alzheimer's disease.

California–San Francisco. ADNI is the result of efforts of many co-investigators 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 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited more than 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Raw MINC (http://www.bic.mni.mcgill.ca/ServicesSoftware/MINC) MP-RAGE T1-weighted structural MR images (slice thickness: 1.2 mm; acquisition plane: sagittal) of 135 different ADNI subjects were chosen and balanced by magnet field strength, manufacturer, diagnosis, qualitative medial temporal atrophy (MTA) severity [12], and age ranges (Tables 1 and 2). In detail, around 150 subjects were selected randomly, among groups with different diagnosis, age, and scan manufacturer. On these subjects, the MTA scores were used to rate atrophy severity. Next, we extracted 135 cases to obtain an optimal balance for all the aforementioned variables and atrophy severity. Besides the attention to these variables, subjects were taken randomly. Variables relating to type of machine and site were not balanced. The scans were obtained on the following 3 Tesla machines: Philips Achieva (phases: ADNI-2 and ADNI-GO), Philips Gemini (ADNI-2), Philips Intera (ADNI-2, ADNI-GO), Philips Ingenia (ADNI-2), GE Signa (ADNI-GO), GE Signa Excite (ADNI-GO), GE Signa HDx (ADNI-2, ADNI-GO), GE Signa HDxt (ADNI-2), GE Signa Excite (ADNI-GO), GE Signa HDx (ADNI-GO), Siemens Trio (ADNI-GO), Siemens TrioTim (ADNI-GO); and 1.5 Tesla machines: Siemens Sonata (ADNI-GO), Siemens Symphony (ADNI-GO), Siemens Avanto (ADNI-GO), Philips Gyroscan Intera (ADNI-GO), and Philips Intera (ADNI-GO). Detailed information about the specific acquisition protocol for each machine in each project phase can be found at http://www.adni-info.org/scientists/MRIProtocols.aspx.

2.2. Preprocessing

Image preprocessing was done centrally, and tracers received reoriented images ready to be segmented.

The ADNI images were downloaded in MINC format and reoriented along the AC-PC line using a six-parameter linear registration from either the Montreal Neurological Institute (MNI, Montreal, Canada) package AutoReg (version 0.98v) (www.bic.mni.mcgill.ca) or the functional MRI of the brain Software Library (FSL, Oxford, UK) package FLIRT (version 4.1, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL). The MNI ICBM152 template with $1 \times 1 \times 1$ mm voxel dimensions was used as the reference space for reorientation of the scans. Resampling was carried out with a trilinear interpolation.

2.3. Hippocampal measurements

Before segmentation, and in the phase of image selection, a larger number of MRI were assigned a medial temporal

Table 2
Sociodemographic and clinical features of controls, MCI, and AD subjects

	CTRL, N = 22	MCI, N = 24	AD, N = 22	*P*	*P* (MCI)	*P* (AD)
1.5 T						
Age (yr)	76 (7)	74 (8)	74 (8)	n.s.	–	–
Gender (F/M)	10/12	12/12	11/11	n.s.	–	–
Education (yr)	16 (3)	16 (3)	15 (3)	n.s.	–	–
MMSE	29 (1)	27 (3)	23 (2)	**<.0005**	**<.0005**	**<.0005**
Scheltens	1.2 (1.2)	1.7 (1.3)	2.4 (1.3)	**.007**	n.s.	**.002**
3T						
Age (yr)	76 (7)	76 (8)	75 (8)	n.s.	–	–
Gender (F/M)	12/10	7/15	13/10	n.s.	–	–
Education (yr)	16 (3)	16 (3)	14 (3)	.061	n.s.	**.025**
MMSE	29 (1)	25 (3)	20 (5)	**<.0005**	**<.0005**	**<.0005**
Scheltens	1.0 (1.1)	1.9 (1.1)	2.9 (1.1)	**<.0005**	**.019**	**<.0005**

Abbreviations: CTRL, controls; MCI, mild cognitive impairment; AD, Alzheimer's disease; MMSE, Mini-Mental State Examination; F/M, female or male; n.s., not significant.

NOTE. Values denote mean (standard deviation) or frequencies. *P* computed with analysis of variance, t-tests versus controls, and Fisher's exact test. MMSE was lacking for nine controls, two MCI, and eight AD at 3T.

score based on visual evaluation as defined by Scheltens et al. [12]. MTA scores were rated by a single rater (M Bocch) expert in MTA visual evaluation. Her ICCs (95% confidence interval) were: intra-rater: 0.969 (0.924–0.981), interrater: 0.940 (0.851–0.976). These ICCs were computed on an independent sample of 20 subjects. The score attributed to each subject consisted of the mean score between the right and left hippocampi, as described by DeCarli et al. [13].

Hippocampi were segmented once each by five different tracers. Segmentations were performed using MultiTracer 1.0, http://www.loni.usc.edu/Software/MultiTracer, developed at the Laboratory of Neuro Imaging, LONI, at UCS, Los Angeles, CA), using the same software version and settings used in the previous project phases [5,9].

Tracers were selected for being among the most qualified HarP tracers within the HarP project. They were from five different centers: LENITEM, Brescia, Italy (MBocch); Laval University, Québec City, Canada (RG); Germany; DZNE, Rostock, Germany (MG); Kawamura Gakuen Woman's University, Abiko-city, Japan (MN); and University of Medicine, Mainz, Germany (DW). The ICC values across all five raters were as follows: consistency method: left = 0.970 (95% confidence interval or CI 0.928–0.991); right = 0.988 (0.970–0.997); absolute method: left = 0.953 (95% CI 0.873–0.987); right = 0.975 (0.920–0.994). All the tracers involved in this study were researchers in the dementia field, and specifically in the field of neuroimaging. Four of them (MBocch, RG, MG, DW) also had previous extensive experience in manual hippocampal segmentation, whereas MN learned hippocampal segmentations for the HarP project, achieved the highest results in the qualification phase [10,11] and completed all segmentations of Validation Phase I described in [5]. MBocch and RG had been in the group that coordinated the HarP project, and had extensive knowledge of the HarP for their experience in having worked at many key steps of its development. The specific training on HarP segmentation received by the five tracers was as follows: MBocch, RG, and DW carried out the whole training as "Master Tracers" [1,9], and provided the benchmark images for the qualification platform (the central web-system allowing standard training and qualification for new remote tracers. The platform was used to train and qualify tracers for the HarP project, and is now publicly accessible from the home page of www.hippocampal-protocol.net). Such training consisted of learning the tracing of the so-called segmentation units (SUs), the "pieces" of hippocampus that are included or excluded by the currently available segmentation protocols, and that therefore represents the landmark variability among protocols. MG and MN carried out the training and qualification on the standard web platform, and performed all segmentations of Validation Phase I [5]. Their individual performance on the platform was Jaccard = 0.85 and Dice = 0.92 (MG) and Jaccard = 0.83 and Dice = 0.91 (MN).

Segmentations were carried out based on the HarP (Appendix II in this special issue). Briefly, only the outer con- Q44 tour of the hippocampus was delineated, including the whole hippocampal head, the alveus and fimbria from the head to the tail, the subiculum, and the whole tail including the Andreas Retzius and the fasciolar gyri. Any internal spaces, i.e. the sets of voxels that appear as hypointense compared with the hippocampal gray matter, were excluded. These spaces, or CSF pools, are considered to be remnants of the hippocampal sulcus and cists, and have been segmented using separate labels to subtract their volume from the volume of the whole hippocampus. Quality check was carried out by a HarP expert not involved in segmentation (MBocca): segmentations of all hippocampi were examined slice by slice and compliance to all HarP criteria was evaluated for all boundaries and segmentation procedures. Corrections were required through written feedback to tracers. Corrected segmentations were again checked for full compliance, until complete compliance was achieved for each slice segmented for each hippocampus. Hippocampal volumes presented in this work consist of the volumes computed from the outer hippocampal contour minus the volumes of the CSF pools (if any) segmented for that hippocampus.

2.4. Statistics

Fisher's exact test was used to evaluate the homogeneous representation of MTA severity. The homogeneity of variance and normality of data distributions were evaluated with the Levene and Kolmogorov-Smirnov tests. ANOVA and t-tests were used to estimate the significance of volumes group differences, Spearman's rho for correlations with MTA scores.

2.5. Labels

The contours segmented in the AC-PC oriented images have been voxelized using a custom Matlab routine. First, the contour is loaded and represented on a grid whose dimensions are identical to the AC-PC image. The interior of the contour of each coronal slice is then filled using Matlab's inpolygon() function (Fig. 1A).

One can notice that the contour is roughly approximated due to the fact that the contours have been traced in a subvoxel space. The approximation can be refined by representing the contour on a grid whose dimensions are greater than the original AC-PC image resolution (Fig. 1B).

A good approximation of the label was obtained using a grid 10 times the original AC-PC resolution. This oversampled contour can be represented as a binary image (Fig. 1C) and was downsized to the original AC-PC image dimensions using the Matlab imresize() function using a bicubic interpolation (Fig. 1D). Voxels for which the segmentation contour covered less than 50% of the total voxel volume were discarded using the im2bw() Matlab method with a 0.5 threshold (Fig. 1E).

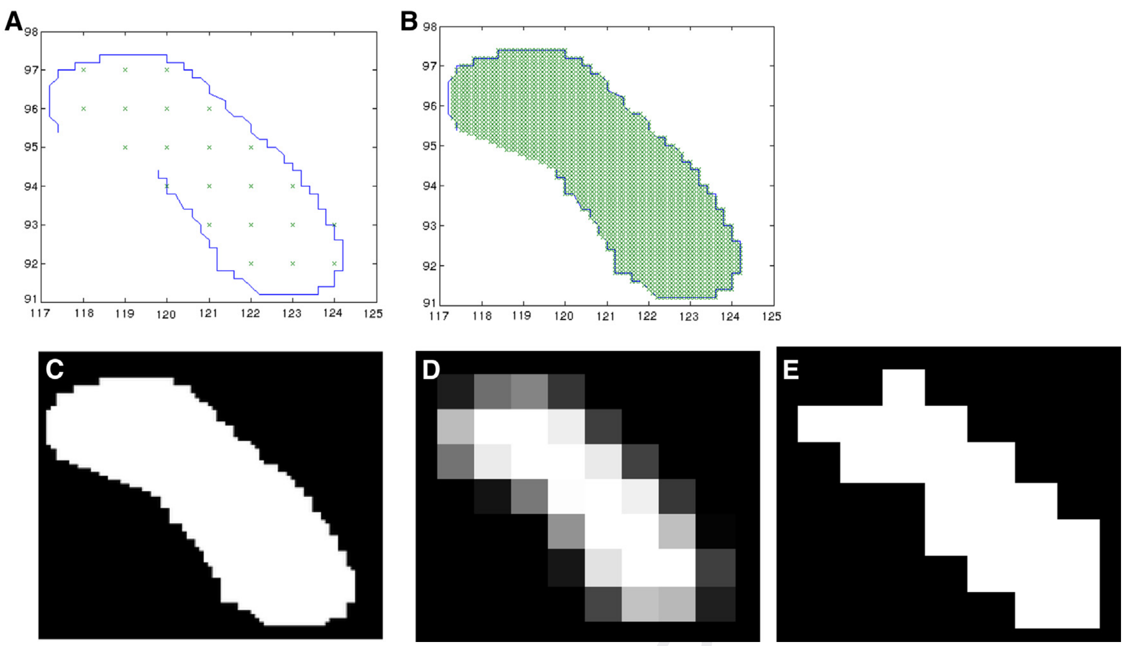


Fig. 1. Steps of contour voxelization (label processing). The *x* and *y* axes represent the voxel indexes on the coronal axis. (A) Coronal contour filled using the inpolygon() Matlab function. The blue line represents the contour traced by the expert and the green cross represents the center of a voxel; (B) oversampled coronal contour (10×) filled using inpolygon(); (C) coronal oversampled (10×) contour binary segmentation; (D) coronal label downsized to original AC-PC image dimensions using the imresize() Matlab function and bicubic interpolation; (E) thresholded label where each voxel covering less than 50% of the total voxel volume was discarded using the im2bw() Matlab function. Q43

AC-PC voxelized labels were then back-transformed to native space using the inverse linear transform and trilinear interpolation using MINC. A HarP expert then checked the labels mapped onto the MRIs in native space to ascertain that reorientation did not influence the appropriate mapping with the hippocampus as defined in the HarP.

3. Results

Consistent with the initial selection, scanner manufacturer, diagnosis, age bins, and gender were homogenously represented in the sample (Table 1). A slight difference emerged for education: AD patients in the 1.5 T sample had a mean of 15 years, and AD patients in the 3T sample had 14 years, the latter differing significantly from the 16 years of controls (Table 2).

3.1. Hippocampal and CSF pools volumes

AD patients had 20–27% smaller hippocampal volumes (mm^3) than controls: left = 2301 (SD = 492), right = 2445 (525); MCI had about 14% smaller volumes: left = 2596 (447), right = 2686 (473). Controls volumes (mm^3) were left = 3060 (SD = 502), right = 3120 (897). The difference among groups was significant at $P < .0005$ at ANOVA. The pattern of results remained unchanged when stratifying groups by magnet field strength (Table 3; Fig. 2).

CSF pool spaces (mm^3) were: controls: left = 23 (32), right = 25 (25), MCI: left = 15 (13), right = 22 (16), and AD: left = 11 (13), right = 20 (25). When stratified by magnet field strength, slightly larger volumes and more frequent outliers occurred in the 1.5 T sample (Table 3, Fig. 3). The overall group differences appear to be due to a relatively small number of subjects with larger CSF pools (Fig. 3).

Hippocampal volumes significantly correlated with atrophy severity at MTA (Table 4).

3.2. Unusual anatomy

Five subjects (3.7%) had unusual anatomy: ADNI subjects 023_S_0061 (image: 132164) and 067_S_1185 (image: 65946) had part of the hippocampal head located medial to the amygdala, rather than ventral/ventro-medial, in the coronal view; subjects 002_S_1280 (image 233435) and 098_S_0172 (image 11398) had very large CSF pools; subject 002_S_0954 (image 108600) had gray voxels of the same intensity as hippocampal gray matter above hippocampal body, beyond the alveus/fimbria.

3.3. Digital labels

A maximum of two rounds of corrections were required from tracers to achieve full compliance with the HarP for all the segmented hippocampi.

The voxelized and reoriented labels resembled the HarP segmentation criteria at visual quality check.

Part of the data (in.ucf, MINC and NIFTI format, linear Q6 transformations and.mnc reoriented voxelized and interpolated labels in native space) are available at www.

Table 3
Hippocampal and CSF pools volumes

	Control	MCI	% Change	*P* (MCI vs ctrl)	AD	% Change	*P* (AD vs ctrl)	*P* (ANOVA)
1.5 T	N = 22	N = 24			N = 22			
Hippocampus								
L	3119 (533)	2620 (447)	16.0	**.002**	2405 (507)	22.9	**<.0005**	**<.0005**
R	3156 (506)	2647 (506)	16.1	**.001**	2487 (543)	21.2	**<.0005**	**<.0005**
CSF pools								
L	31 (38)	17 (10)	–	n.s.	16 (17)	–	n.s.	n.s.
R	29 (28)	27 (16)	–	n.s.	31 (30)	–	n.s.	n.s.
3 T	N = 22	N = 22			N = 23			
Hippocampus								
L	3001 (452)	2569 (456)	14.3	**.003**	2203 (467)	26.6	**<.0005**	**<.0005**
R	3084 (479)	2729 (441)	11.5	**.004**	2405 (516)	22.0	**<.0005**	**<.0005**
CSF pools								
L	15 (22)	13 (15)	–	n.s.	5 (7)	–	.058	n.s.
R	21 (22)	16 (13)	–	n.s.	9 (12)	–	**.039**	n.s.

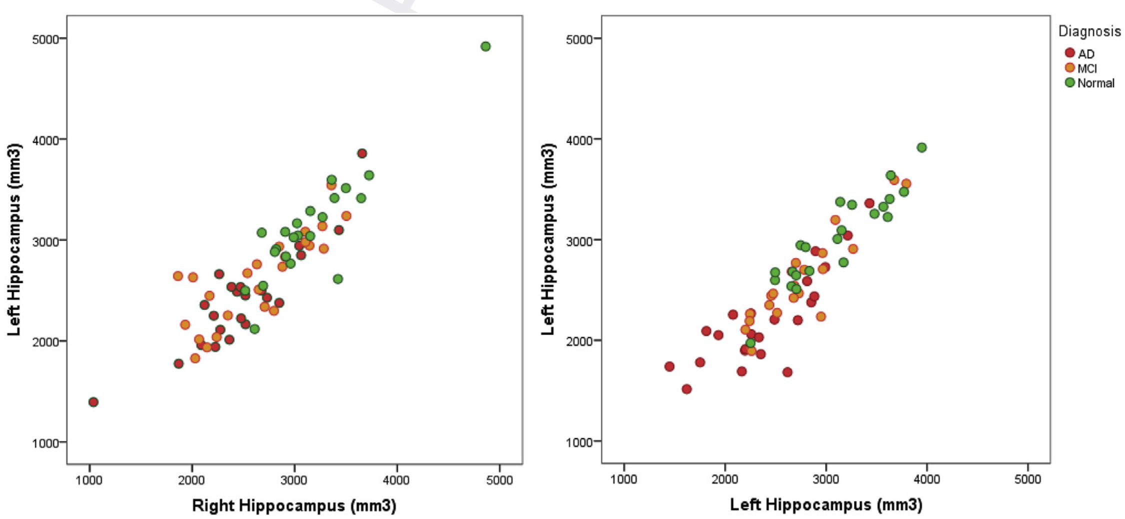
Abbreviations: CSF, cerebrospinal fluid; MCI, mild cognitive impairment; ctrl, control; AD, Alzheimer's disease; ANOVA, analysis of variance; L, left; R, right; n.s., not significant.

NOTE. *P* values refer to significance at ANOVA among controls, MCI and AD, and t-tests to comparisons of patient groups versus controls. Hippocampal volumes are expressed in mm^3. The volume of internal CSF pools was excluded from total hippocampal volume.

hippocampal-protocol.net. The ADNI subject IDs, image codes, and conversion files reporting the orientation function used for the reorientation of each MRI along the AC-PC line, as required by the HarP, are also reported.

4. Discussion

With this work, we carried out a natural extension of the project on the Harmonization of Protocols for Manual Hippocampal Segmentation. This was aimed to define an optimal procedure allowing the proper transference of the standard segmentation of the whole hippocampus outer contour into concrete everyday usage. We have provided a relatively large set of benchmark hippocampal segmentations based on the HarP, that cover a wide range of physiological and pathological variability. This set is meant to provide the appropriate reference to automated algorithms so that they can generalize the learning and appropriately segment hippocampi of new subjects. Moreover, this work can be used to improve the current qualification platform, and allow the periodical check of qualified tracers by testing them on different images that can be taken from this larger set of certified labels. This work follows the completion of the HarP project, defining the new standard for the measurement of hippocampal volumetry and its use as a biomarker for AD. So far, another large set of benchmark images was produced during the project aimed to develop the HarP itself. However, these came from a very limited number of ADNI subjects (n = 10, considering only certified benchmark labels used for the training platform [9–11]), an insufficient sample to train automated segmentation algorithms. On the contrary, the set of benchmark images described in this article is



web 4C/FPO

Fig. 2. The distribution of hippocampal volumes versus diagnosis at 1.5 T (left panel) and 3 T (right panel). Hippocampal volumes were computed subtracting the volume of cerebrospinal fluid (CSF) pools.

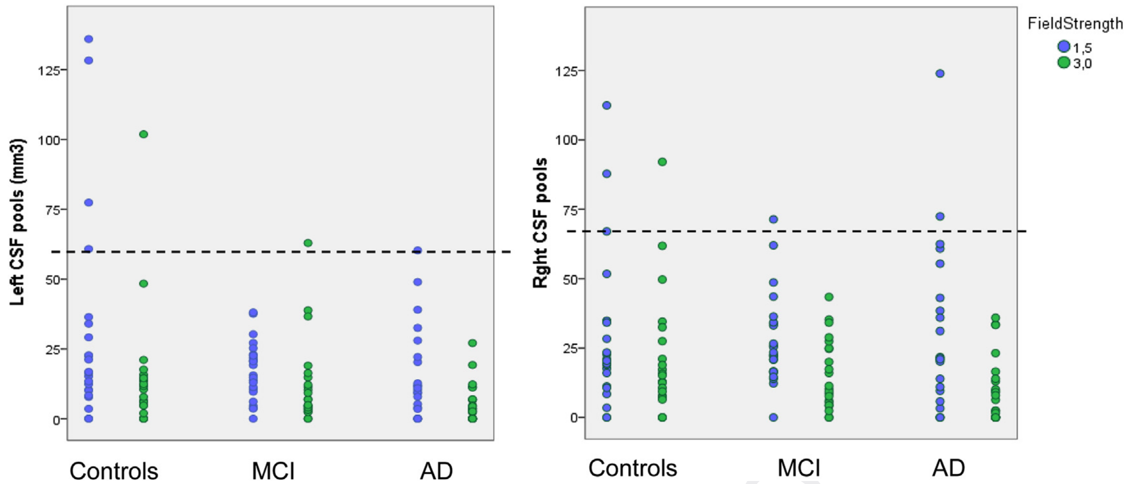


Fig. 3. Distributions of cerebrospinal fluid (CSF) pools volumes. The dotted line denotes the threshold for outliers, computed as two standard deviations beyond the mean of all volumes (59 mm^3 for the left, 66 mm^3 for the right CSF pools).

obtained from 135 different subjects and captures a wider range of physiological morphologies. Thanks to this larger variability, this work additionally provided evidence of known group validity to the HarP, whose proper validation [5] was carried out on MRIs taken from a maximum of 16 different subjects.

Segmented labels were checked by an expert of the HarP, and segmentation corrections were performed where needed until final certification was provided for full compliance with the HarP. Labels are available on the official web site of the project (www.hippocampal-protocol.net) and can be freely downloaded in the most commonly used formats.

4.1. Characteristics of the segmented benchmark hippocampal labels

As expected from the HarP features [4], hippocampal volumes were in the range of those obtained by the most inclusive protocols in the literature [14]. The volume of CSF pools in the context of hippocampal tissue tended to be higher in controls, but the visual assessment of data distribution shows that this may be due to a rather limited number of outliers that were observed more often in the 1.5 T sample (Fig. 3). Indeed, internal CSF pools relate to the normal physiological variability in the morphology of the hippocampal sulcus residual cavity, that, unlike other perihippocampal CSF spaces, appears to be unrelated to both ageing and AD neurodegeneration [15]. Our finding is in line with other published evidence, indicating that particularly large CSF pools were most frequently observed in controls than in MCI or AD groups, an otherwise unexplained finding so far [15].

4.2. Digital labels

Segmentations are available in.ucf, MINC and NIFTI formats.

This will allow to modify the qualification platform [10,11], previously used for the training and qualification of human tracers, enabling the use by developers of automated algorithms. The final aim is to allow algorithm training based on the benchmark segmentations (or part of them, working as training set) produced in this work, upload of labels segmented by algorithms, and perform comparisons of automated

Table 4
Mean hippocampal volume (mm^3) and Spearman's rho correlation values of hippocampal volumes by MTA

Scheltens	0	1	2	3	4	Spearman's	
1.5 T							
N	13	19	15	11	10	ρ	*P*
Left	3025 (785)	2932 (406)	2517 (459)	2501 (292)	2408 (665)	−.405	**.001**
Right	3168 (669)	2999 (364)	2565 (408)	2474 (439)	2379 (720)	−.497	**<.0005**
3 T							
N	12	14	17	13	11		
Left	3029 (463)	2889 (470)	2432 (525)	2302 (487)	2286 (455)	−.531	**<.0005**
Right	3142 (516)	3031 (466)	2535 (576)	2476 (395)	2526 (422)	−.467	**<.0005**

Abbreviations: MTA, medial temporal atrophy severity evaluated visually (Scheltens et al. [12]); CSF, cerebrospinal fluid.
NOTE. Values denote mean (standard deviation). The volume of internal CSF pools was excluded from total hippocampal volume.

segmentations versus the benchmark reference (or part of them, working as a test sample) produced in this work. Comparisons are planned to be performed with respect to volume, spatial overlap, and spatial distance of the external boundary.

4.3. Limitations

One of the main limitations of this study consists in the lack of longitudinal images of the same subjects; this will not serve algorithms that exploit differences between scans, nor allow for the validation of atrophy rate estimations and other longitudinal behavior (e.g., transitivity, linearity).

A second limitation lies in the segmentation of each hippocampus by a single tracer rather than by more experts as for the previously generated benchmark labels [9]. It was felt that the very accurate definitions provided by the HarP reduced the range of alternative segmentations that may be considered to be correct for each hippocampus. This is consistent with the very high absolute interrater ICCs among the tracers involved in this work. Nonetheless, some divergence may be considered acceptable due to a certain degree of ambiguity in tissue definition from MRIs, which do not provide the perfect visualization of subtle features of brain morphology. Certification criteria that can flexibly account for these ambiguities depending on the different anatomical regions will need to be defined, based on quantitative data and quality check of the performance of a large set of segmentations by new tracers, to make certification both possible and highly accurate for human tracers and algorithms.

RESEARCH IN CONTEXT

Hippocampal volumetry is a useful biomarker for Alzheimer's disease (AD), and recently a standard protocol has been defined to enable different tracers from different laboratories obtain consistent volume estimates. Hippocampal segmentation from magnetic resonance imaging is anyway a time consuming task, and large clinical trials, and routine clinical needs, may benefit of segmentation by automated algorithms. The variability of hippocampal anatomy is large, therefore the training of automated algorithms requires a very large set of segmentation examples in order for them to learn and be able generalize to new subjects. In this work, such a data set of segmentations has been produced. The segmentations have been certified for full compliance with the Harmonized Protocol and released for public use of the community.

This work is the step that allows the concrete and widespread use of the Harmonized Protocol for research and clinical purposes.

Acknowledgments

This project was carried out thanks to a public-private partnership including the Alzheimer's Association as the main partner, and Bioclinica, Brain Image Analysis LLC, IXICO Ltd., Roche, Synarc, and True Positive Medical Devices Inc. Chahin Pachai, Ronald Pierson, Derek Hill, Emilio Merlo-Pich, Joyce Suhy and D. Louis Collins gave and helpful suggestions. We wish to thank Adam Schwartz for help with the coordination of funding efforts.

The Alzheimer's Association, Wyeth, part of the Pfizer group, and Lilly have supported the previous steps regarding the development of the HarP.

The project 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; project Coordinator is Marina Boccardi, IRCCS Fatebenefratelli, Brescia, Italy. EADC Centres (local PI) are: IRCCS Fatebenefratelli, Brescia, Italy (GB Frisoni); University of Kuopio and Kuopio University Hospital, Kuopio, Finland (H Soininen); Höpital Salpètriere, Paris, France (B Dubois and S Lehericy); University of Frankfurt, Frankfurt, Germany (H Hampel); University Rostock, 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, UK (N Fox); NEUROMED, Centre for Neuroimaging Sciences, London, UK (A Simmons). ADNI Centres are: Mayo Clinic, Rochester, MN (CR Jack); University of California Davis, CA (C DeCarli); University of California, Los Angeles (UCLA), CA (G Bartzokis); University of California San Francisco (UCSF), CA (M Weiner and S Mueller); Laboratory of NeuroImaging (LoNI), University of California, Los Angeles (UCLA), CA (PM Thompson); Rush University Medical Center, Chicago, IL (L deToledo-Morrell); Rush Alzheimer's Disease Center, Chicago, IL (D Bennet); Nortwestern 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); Oregon Health&Science University, Portland, OR (J Kaye). Other Centres 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); Wayne State University (WSU), Detroit, MI (C Watson); Institute for Ageing and Health, Wolfson Research Centre, Newcastle General Hospital, Newcastle, UK (J O'Brien). Population-based studies: PATH through life, Australia (P Sachdev and JJ Maller); SMART-Medea Study, The Netherlands (MI Geerlings); Rotterdam Scan Study, The Netherlands (T denHeijer).

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); MNI, McGill University, Montreal, Canada (L Collins). Advisors: Clinical issues: PJ Visser, Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands; EADC PIs: B Winbald, Karolinska Institute, Sweden and L Froelich, Central Institute of Mental Health, Mannheim, Germany; Dissemination & Education: G Waldemar, Copenhagen University Hospital, Copenhagen, Denmark; ADNI PI: M Weiner, University of California San Francisco (UCSF), CA; Population studies: L Launer, National Institute on Aging (NIA), Bethesda and W Jagust, University of California, Berkeley, CA.

Data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation;BioClinica.; Biogen Idec.; Bristol-Myers Squibb Company; Eisai; 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; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to Rev October 14, 2013 support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the UCLA.

References

[1] 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 2013; http://dx.doi.org/10.1016/j.jalz.2013.03.001. pii:S1552–5260(13)00078-2.

[2] Boccardi M, Bocchetta M, Apostolova LG, Preboske G, Robitaille N, Pasqualetti P, et al. Establishing magnetic resonance images orientation for the EADC-ADNI manual hippocampal segmentation protocol. J Neuroimaging 2014;24:509–14.

[3] 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.

[4] 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 2014; http://dx.doi.org/10.1016/j.jalz.2014.02.009. pii: S1552–5260(14)02418-2.

[5] Frisoni GB, Jack CRJ, Bocchetta M, Bauer C, Frederiksen K, Liu Y, et al. The EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement 2014; http://dx.doi.org/10.1016/j.jalz.2014.05.1756. pii: S1552–5260(14)02468-6.

[6] Mattsson N, Andreasson U, Persson S, Arai H, Batish SD, Bernardini S, et al. The Alzheimer's Association external quality control program for cerebrospinal fluid biomarkers. Alzheimers Dement 2011;7:386–3956.

[7] Höglund K, Bogstedt A, Fabre S, Aziz A, Annas P, Basun H, et al. Longitudinal stability evaluation of biomarkers and their correlation in cerebrospinal fluid and plasma from patients with Alzheimer's disease. Alzheimers Dement 2012;32:939–47.

[8] Apostolova LG, Zarow C, Biado K, Hurtz S, Boccardi M, Somme J, et al. Relationship between hippocampal atrophy and neuropathology markers: a 7T MRI study. Alzheimers Dement 2014. I revision.

[9] Bocchetta M, Boccardi M, Ganzola R, Apostolova LG, Preboske G, Wolf D, et al. Harmonized benchmark labels of the hippocampus on MR: the EADC-ADNI project. Alzheimers Dement 2013; http://dx.doi.org/10.1016/j.jalz.2013.12.019. pii: S1552–5260(14)00010-7.

[10] Duchesne S, Valdivia F, Robitaille N, Abiel Valdivia F, Bocchetta MM, Boccardi M, et al. Manual segmentation certification platform. IEEE 2013; Medical Measurements and Applications proceedings (MeMeA),:35–39.

[11] Duchesne S, Valdivia F, Robitaille N, Mouiha A, F, Valdivia A, et al. Manual segmentation qualification platform for the EADC-ADNI Harmonized Protocol for hippocampal segmentation project. Alzheimers Dement 2014 (First Revision).

[12] 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.

[13] DeCarli C, Frisoni GB, Clark CM, Harvey D, Grundman M, Petersen RC, et al. Qualitative estimates of medial temporal atrophy as a predictor of progression from mild cognitive impairment to dementia. Arch Neurol 2007;64:108–15.

[14] Geuze E, Vermetten E, Bremner JD. MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. Mol Psychiatry 2005;10:147–59.

[15] Li Y, Li J, Segal S, Wegiel J, De Santi S, Zhan J, et al. Hippocampal cerebrospinal fluid spaces on MR imaging: relationship to aging and Alzheimer disease. AJNR Am J Neuroradiol 2006;27:912–8.