Alzheimer's & Dementia 11 (2015) 161-174

Manual segmentation qualification platform for the EADC-ADNI harmonized protocol for hippocampal segmentation project

Simon Duchesne[a,*], Fernando Valdivia[a], Nicolas Robitaille[a], Abderazzak Mouiha[a], F. Abiel Valdivia[a], Martina Bocchetta[b,c], Liana G. Apostolova[d], Rossana Ganzola[a], Greg Preboske[e], Dominik Wolf[f], Marina Boccardi[b], Clifford R. Jack, Jr.,[e], Giovanni B. Frisoni[b,g], for the EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]*Department of Radiology, Université Laval and Centre de Recherche de l'Institut universitaire en santé mentale de Québec, Quebec City, Canada*
[b]*LENITEM (Laboratory of Epidemiology, Neuroimaging and Telemedicine) IRCCS – S. Giovanni di Dio – Fatebenefratelli, Brescia, Italy*
[c]*Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy*
[d]*Mary S. Easton Center for Alzheimer's Disease Research and Laboratory of NeuroImaging, David Geffen School of Medicine, University of California, Los Angeles, USA*
[e]*Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN, USA*
[f]*Klinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität, Mainz, Germany*
[g]*University Hospitals and University of Geneva, Geneva, Switzerland*

Abstract

Background: The use of hippocampal volumetry as a biomarker for Alzheimer's disease (AD) requires that tracers from different laboratories comply with the same segmentation method. Here we present a platform for training and qualifying new tracers to perform the manual segmentation of the hippocampus on magnetic resonance images (MRI) following the European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative (EADC-ADNI) Harmonized Protocol (HarP). Our objective was to demonstrate that the training process embedded in the platform leads to increased compliance and qualification with the HarP.

Method: Thirteen new tracers' segmentations were compared with benchmark images with respect to: (a) absolute segmentation volume; (b) spatial overlap of contour with the reference using the Jaccard similarity index; and (c) spatial distance of contour with the reference. Point by point visual feedback was provided through three training phases on 10 MRI. Tracers were then tested on 10 different MRIs in the qualification phase.

Results: Statistical testing of training over three phases showed a significant increase of Jaccard (i.e. mean Jaccard overlap $P < .001$) between phases on average for all raters, demonstrating that training positively increased compliance with the HarP. Based on these results we defined qualification thresholds which all tracers were able to meet.

Conclusions: This platform is an adequate infrastructure allowing standardized training and evaluation of tracers' compliance with the HarP. This is a necessary step allowing the use of hippocampal volumetry as a biomarker for AD in clinical and research centers.

Keywords: Alzheimer's disease; Qualification platform; Hippocampal segmentation; Harmonized protocol; Magnetic resonance imaging; Qualitative criteria; Quantitative criteria; Area; Spatial overlap; Spatial distance

[1]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.: +1-418-663-5741x4777; Fax: +1-418-663-5971.

E-mail address: simon.duchesne@crulrg.ulaval.ca

1552-5260/$ - see front matter
http://dx.doi.org/10.1016/j.jalz.2015.01.002

1. Introduction

Within the context of Alzheimer's disease (AD), hippocampal volumetry is an in vivo biomarker of major interest that has recently been accepted as part of newly revised diagnostic criteria [1–3]. The use of hippocampal volumetry has recently been approved for enrichment in mild cognitive impairment clinical trials by the European Medicine Agency [4]. Routine clinical diagnosis may also benefit of its informative value in cases where the clinical picture of patients does not clearly indicate whether AD is the origin of the cognitive symptoms. However, some methodological problems hurdle the use of hippocampal volumetry as a biomarker for AD. First, it is essential that different laboratories converge on a same neuroanatomical definition for the hippocampus and on a standard method for its measurement. Second, hippocampal anatomy experts, who must undergo specific training, define the current gold standard for hippocampal volumetry by manual delineating the structure. To date, these aspects do not guarantee, on the one hand, compliance from different laboratories to provide the same volume estimate for the same hippocampus, and on the other, limit wide clinical acceptance.

Measuring the hippocampus in such a reliable way requires, on the one hand, high-contrast images of the human brain, such as those obtained via T1-weighted, anatomical magnetic resonance images (MRI) in standardized protocols [5]; and on the other, a sound neuroanatomical protocol for the manual delineation of the structure on MRI. The different manual segmentation protocols that have been developed over the years [6–8] lead to hippocampal volume estimates in similar, age-matched control groups differing by up to 2.5-fold based on previous estimates [6], and even much higher as observed within our validation study of the Harmonized Protocol (HarP) [9]. This is due to the very wide variance of the different "local" protocols in including or excluding not only small parts (e.g. fimbria or subiculum) but even very large portions of the hippocampus, like the whole head, or the whole tail. As quantified and reported previously [10] the parts of hippocampus that are included or excluded by the different available segmentation protocols range from little portions consisting of 8% to 10% of the total hippocampal volume, to portions of intermediate volume (12–20% of total hippocampal volume) to the very large portion of the main hippocampal body (60% of total hippocampal volume), that includes the hippocampal head and most of the body in our operationalization study [4]. Although this latter unit was entirely included by all the segmentation protocols that we examined in Boccardi et al. [6], other protocols exist where only a small portion of this unit is included and used to provide hippocampal volume estimates (see for example Kaye et al. [11]). Protocols of this kind provide volume estimates in the range of 420 to 450 mm^3, whereas the most inclusive protocols generate volume estimates up to 2.6 cc (see tracers 8 and 16 in Table 2 and Figs. 1 to 4 in Frisoni et al. [6]). Such a heterogeneity in anatomic definitions and segmentation guidelines is thus the main reason that has hampered comparisons among different studies using hippocampal volumetry for diagnosis or as a surrogate marker for disease progression, and that limits its use as a diagnostic marker for clinical diagnosis. Moreover, the traditional manual segmentation paradigm requires initial training, and no further check is made on whether tracers keep being compliant with the original protocol along time, as may be necessary, to ensure that segmentation accuracy does not drift away from the officially adopted method. Ideally, and as long as the gold standard for hippocampal segmentation resides in manual segmentation, the use of hippocampal volumetry as a biomarker for AD should involve that such a periodical check be made to keep the performance of tracers from different laboratories compliant with the standard along time.

An effort has been undertaken by European Alzheimer's Disease Consortium (EADC) and Alzheimer's Disease Neuroimaging Initiative (ADNI) centers to develop a HarP for the manual segmentation of the hippocampus on MRI [12] (www.hippocampal-protocol.net). It represents the largest effort of the scientific and pharmaceutical community to address this issue. This project defined a consensual definition of landmarks and segmentation procedures for a standard protocol for the manual segmentation of the hippocampus through the following steps. First, the most cited protocols in the AD literature were surveyed [8]. Second, the differences between these protocols were operationalized into so called "segmentation units", i.e., "pieces" of hippocampus that can be included or excluded in segmentation [10], and that correspond to the different landmark definitions. Third, these "pieces", and other aspects like image orientation, were investigated quantitatively, to estimate their associated reliability in segmentation, and the informative value as to AD-related atrophy [4,13]. Fourth, decisions were taken about which landmarks and which procedures to include in a standard protocol providing optimal estimates for AD. These decisions were taken through an evidence-based "Delphi" panel; the latter implies a specific procedure consisting of recursive voting sessions, aimed to facilitate consensual decision making [14], and thus is used to provide answers and solutions to complex issues. This procedure consisted in selecting experts with specific expertise in the hippocampal segmentation for AD as Delphi panelists. We drafted a questionnaire defining a number of questions aiming to provide an optimal standard for the hippocampal segmentation for AD. Panelists were asked to answer a questionnaire on a first round, and then in subsequent rounds were informed about the motivated answers provided by the other, anonymous panelists, and invited to answer again the same questions. The principle beneath this Delphi procedure is that panelists, accessing the answers and reasons thereof provided by the other participants, progressively converge toward the most relevant choice [14]. In five voting sessions of this kind, the panel converged on a definition for the HarP that was significantly agreed upon in a statistical fashion by panelists [10,15]. We have then translated the panelists' decisions into a detailed user manual (http://www.hippocampal-protocol.net/SOPs/

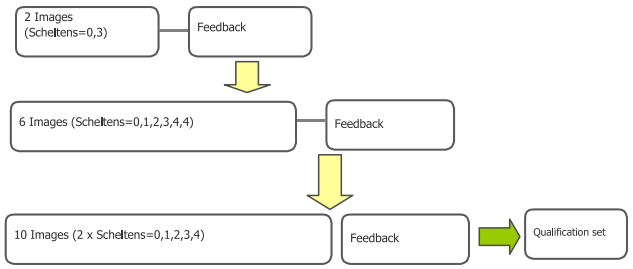

Fig. 1. Training and qualification workflow. The panel describes the three different rounds of the training phase. After the three rounds, a tracer will complete the qualification set (10 additional images).

LINK_PAGE/HarmonizedProtocol_ACPC_UserManual_biblio.pdf), extensively describing landmarks and segmentation procedures so that any segmentation ambiguity is disambiguated as long as possible. Following these instructions, a small group of "master tracers" segmented a set of *benchmark images* according to the consensual definition [16].

At that point, although a protocol had been obtained and exemplars provided, the project lacked a unified platform to train new tracers and certify their compliance with the HarP.

1.1. Objective

The aim of this work was to provide a web-platform accessible by remote users, allowing the standardized training and qualification of manual hippocampal segmentation based on the HarP. More exactly, we planned to create a platform where users could:

a) access all the necessary information and tools to learn and perform manual hippocampal segmentation based on the HarP (*learning phase*);
b) receive quantitative and visual feedback about the compliance of their segmentation based on the benchmark reference segmentations, allowing them to progressively approach correct segmentation (*training phase*); and
c) receive statistics of their tested compliance (or lack thereof) versus the standard, in a separate and final *test phase.*

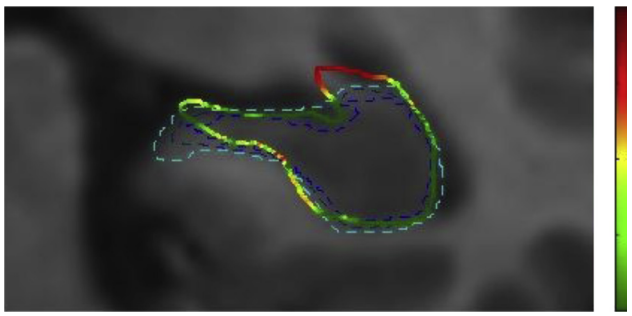

Fig. 2. Segmentation example with masters' maximum contour in cyan; mean contour in blue; and minimum contour in dark blue. User segmentation is shown superimposed with coloring proportional to the distance map ratio *D* shown in Fig. 1 (bottom left), set between 0 and 1 (cf. section 2).

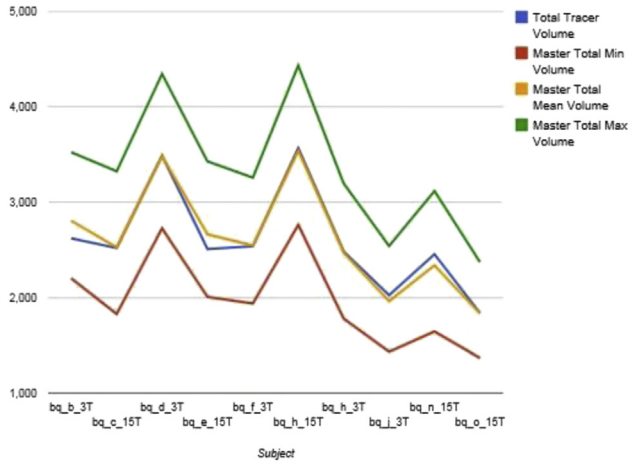

Fig. 3. Immediate quantitative feedback is provided to users on their performance for each of the specific metrics. Shown here (*y*-axis, in cm^3) is the rater's hippocampi volumes (in blue) against the maximum (green), minimum (red), and mean (orange) volumes from the expert tracings, for selected de-identified subjects (*x*-axis labels).

In the present article, we describe how we used this set of benchmark images to produce an interactive web system allowing protocol *learning*, segmentation *training*, and periodical *qualification* of the ability of new tracers to segment the hippocampus according to the HarP, through quantitative comparisons versus the reference segmentations.

2. Methods

2.1. Ethics

Each participant from the ADNI (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 magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment 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, 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 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 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have

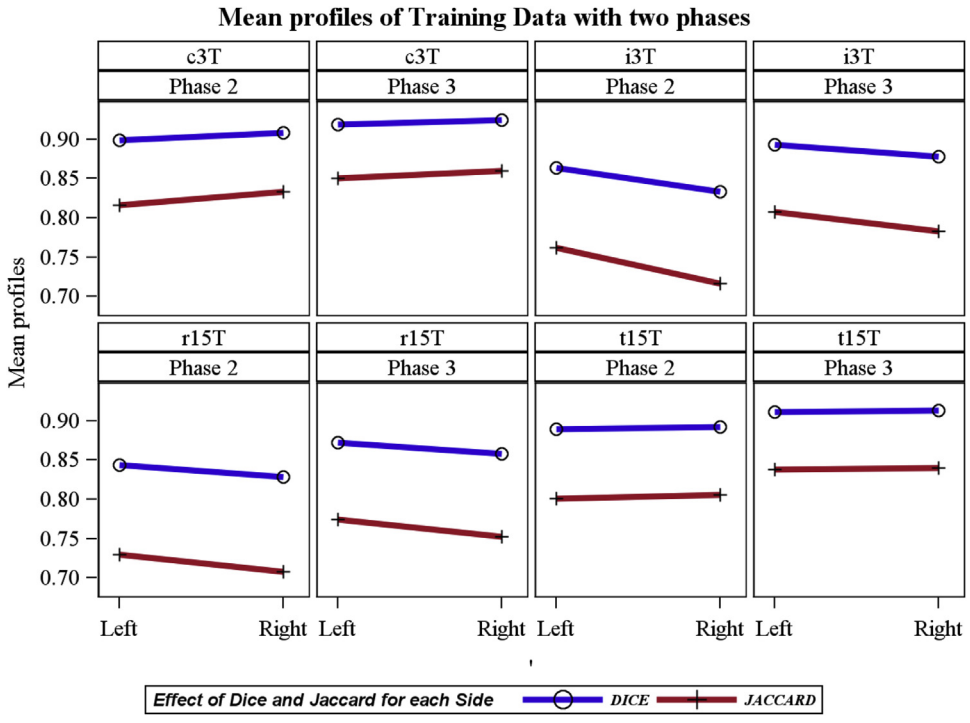

Fig. 4. Repeated measures analysis for all beta raters that completed Phases II and III. In each plot is shown mean Jaccard indices on the *y*-axis (in red; mean Dice overlap statistic also shown for comparison) for left and right hippocampi belonging to different subjects, per phase (e.g. subject "c" at 3.0 T in phases II and III, "r" at 1.5 T, and so on). Due to the experimental design, we therefore had access to four images segmented twice by the 13 new raters. Results show a significant increase between thephases, indicative of increased compliance with the Harmonized Protocol, but also an effect of SIDE (i.e., significant difference between left and right hippocampi in a same subject).

recruited more than 1500 adults, ages 55 to 90 years, 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) cohort was formally evaluated using eligibility criteria that are described in detail elsewhere (http://www.adni-info.org/index.php?option=com_content&task=view&id=9&Itemid=43). The institutional review boards of all participating institutions approved the procedures for this study. Written informed consent was obtained from all participants or surrogates. More information about ADNI investigators is given in the Acknowledgment section.

2.2. Validation formalism

For the purpose of developing the qualification platform as a verification, validation, and evaluation tool, we followed the formalism of Jannin et al. [17] with respect to ensuring that all model components were present to efficiently conduct and report validation results.

2.3. Subjects' images

For this study, we selected 10 subjects from the ADNI database housed at the Laboratory of NeuroImaging (Los Angeles, CA) (www.loni.ucla.edu/ADNI/Data) according to visual atrophy ratings of the medial temporal lobe, using Scheltens's medial temporal atrophy scale [18], to represent the full range of hippocampal atrophy. These subjects are the same as those selected for other sections of the HarP project and were described in detail in ref. [10]. For each subject we downloaded Medical Image NetCDF (MINC) formatted, distortion-corrected, three-dimensional (3D) T1-weighted structural MRIs at 1.5 T and 3.0 T from the Laboratory of NeuroImaging database, and aligned these images through a six-parameters registration (translations, rotations) using the Montreal Neurological Institute package AutoReg (version 0.98v) (www.bic.mni.mcgill.ca) along the slope determined by the line that passes through the anterior and posterior commissures of the brain. We used the Montreal Neurological Institute ICBM152 Nonlinear Symmetric template with $1 \times 1 \times 1$ mm^3 voxel dimensions as the reference. No additional preprocessing steps were performed.

2.4. Master hippocampal segmentations

The benchmark hippocampal segmentations based on the EADC-ADNI HarP to be used as the reference for the qualification of the new tracers were provided by five master tracers and described in detail in ref. [16]. The sample is composed of 200 labels, as each of the five different tracers provided labels for both hippocampi

of the same 10 ADNI subjects, and for both 1.5 T and 3T MRIs. Briefly, segmentations of left and right hippocampi were manually performed using the interactive MultiTracer 1.0 software (http://www.loni.ucla.edu/Software/MultiTracer) on approximately 30 contiguous 1-mm-thick coronal brain sections, with the simultaneous visualization of the axial and sagittal planes. The oversampling interpolation factor was kept constant (coronal view ×5, sagittal view ×3; default [FFT/Chirp-z] interpolation) and the direction of segmentation was from rostral to caudal. Tracers segmented the hippocampus according to the HarP procedure (centroalzheimer.it/public/SOPs/Suppl_Simon/HarmonizedProtocol_ACP-C_UserManual_biblio.pdf). Briefly, segmentations were performed from rostral to caudal, including the whole hippocampal head, the alveus and fimbria, and the whole tail together with the Andrea Retzius and the fasciolar gyri. The first slice was defined with the help of 3D views, allowing to detect where hippocampal tissue begins relative to the amygdala and to the alveus separating the two structures. The last slice was defined as the most caudal hippocampal issue, bordering the indusium griseum and the isthmus.

2.5. New tracer training and qualification

A call for tracers was sent to all centers participating in the HarP project. The ultimate goal of recruitment was validation of the HarP [9]: tracers were asked to segment both hippocampi on 20 ADNI MRIs using the protocol that was normally adopted by their laboratories. Then they were asked to resegment the same hippocampi with the HarP, after the completion of all training and qualification phases on the platform described in this article. Tracers with best performance on the qualification phase described here were involved in subsequent parts of the project. One consisted in the exact quantification of the sources of variance in HarP segmentations [9], the other consisted in the generation of additional HarP benchmark segmentations [5].

For the purposes of this study, these tracers were required to access the platform (see section 2.6) and to segment the same set of 20 ADNI images following the same settings/procedures as master tracers [16]. Specifically, 10 images were assigned to a "training" set (for a total of 20 hippocampi), and the remaining assigned to the "qualification" set (cf. Fig. 1). We paid particular attention to provide tracers the full range of atrophy in both sets. Specifically, if a 1.5 T MRI for a given subject was included in the training sample, then the 3T MRI of the same subject was included in the qualification sample. To measure the learning effect, we segregated the training set in three phases, whereby users had to segment a few images in Phase I, and successively more in following phases. Each phase included the images from the previous phase. Once completed the training phase, new tracers could move on to segment, once, the qualification set. All data were managed via our qualification platform.

2.6. Qualification platform

We developed a web-based environment for protocol learning, training and qualification of hippocampal segmentations made by new tracers against the masters' benchmark images. The environment can be accessed via the "Certification" section of the HarP website (www.hippocampal-protocol.net). We developed the system with three levels of access, namely:

(1) a common area, where visitors can access the protocol definition, presentation and examples;
(2) a registered user area, where users can download training and qualification datasets (contours), upload their training examples, receive qualitative and quantitative feedback on their performance, upload their qualification segmentations, and receive notification of their having succeeded in segmenting the hippocampus according to the HarP; and
(3) a section reserved for system administrators, allowing them to authorize user registrations, upload training and qualification examples, and access user statistics.

The environment has been developed using the Model-View-Controller paradigm, a software design approach used to organize code in such a way that the business logic and data presentation are separate. The back-end and front-end were coded in the server-side scripting language PHP, which is interpreted by a web server and generates the HTML pages seen by users on the Qualification Platform.

2.7. Segmentation assessment framework

The Qualification Platform was designed to assess a new tracer's performance in either training or qualification. It was built using a framework for validating segmentation performance, inspired by Jannin's model, in which we proposed that any measurement fulfill the following criteria: (a) representativity of the task; (b) specificity and sensitivity to the task; and (c) orthogonality of measurement (i.e. limited correlation between measures). A number of measurements are available (cf. Appendix A) however, for purposes of validating segmentation accuracy, and within the context of similar image segmentation between a new rater and the experts (i.e. existence of reference labels), we proposed the following elements to be measured:

a) Absolute segmentation volume, as the expression of the finality of the task;
b) Spatial overlap of contour with the proposed reference; and

c) Spatial distance of contour with the proposed reference.

A detailed description of each feature is provided in Appendix B.

2.8. Statistical analyses

Our objective was to assess the increased compliance of tracers that had gone through the training phases with the benchmark segmentations performed by the expert HarP tracers. The segmentations by the five expert tracers [7] that have been uploaded as the HarP reference for naïve tracers denoted a collective absolute intraclass coefficient (ICC) >0.95. Here we sought to have trainees of our platform increase their performance through the phases of learning the HarP, creating initial segmentations, performing multiple training rounds, and qualifying. To test this increase in compliance, even though volumes and spatial distance metrics were shown and recorded, we performed a repeated measures analysis of the Jaccard overlap statistic, averaged over all tracers, on a per-image basis, taking into consideration the effect of side as a further indicator of increased compliance. Following training, the degree of divergence associated to new tracers' segmentations that could be considered compliant to the HarP was evaluated. This evaluation allowed to set thresholds for acceptable divergence for the qualification of segmentations.

3. Results

The Qualification Platform came online on 5 October 2012. We segregated the training set in three phases, whereby users segmented 2 images (4 hippocampi) in Phase I, 6 images (12 hippocampi) in Phase II, and 10 images (20 hippocampi) in Phase III. Each phase included the images from the previous phase, corrected based on feedback. Users could visualize their segmentations, see their statistical results and visualize the distance map ratios online (cf. Figs. 2 and 3). Further individual verbal feedback was provided to users by an independent expert of HarP for the first two phases, to improve their comprehension and compliance with the HarP criteria, following the same procedure as reported in ref. [16].

Following training, new tracers had to segment a further 20 hippocampi on 10 images from the qualification set. In this phase they did not receive visual feedback for their performance, nor individual slice statistics, but only a compendium of volumes and overlap statistics versus the reference (Fig. 3). In this qualification phase no correction of segmentation was required or admitted.

3.1. Users

We launched a call to laboratories that had shown an interest in participation to the HarP Project, and had expertise in hippocampal segmentation in the field of AD. Tracers from these laboratories can all be considered experienced tracers based on the segmentation protocol locally used in their laboratories. Intra- and inter-rater of at least 0.80 based on local protocols were the admission criteria. In this project, 18 users responded to our invitation and registered on the platform, and 13 completed all three steps of the training with five dropouts for various professional reasons. (The 13 tracers who completed training and qualification were from the following 12 centers—PI [Tracer/s] centre—Charles DeCarli [Oliver Martinez] Department of Neurology, University of California, Davis, CA; Leyla de Toledo-Morrell [Travis Stoub] Department of Neurological Sciences, Rush University, Chicago, IL; Giovanni Frisoni [Enrica Cavedo, Mariangela Lanfredi] LENITEM [Laboratory of Epidemiology, Neuroimaging and Telemedicine] IRCCS—Istituto Centro S. Giovanni di Dio—Fatebenefratelli Brescia, Italy; Nick Fox [Melanie Blair] Dementia Research Centre, UCL Institute of Neurology, Box 16, National Hospital for Neurology and Neurosurgery, Queen Square, London WC1N 3BG, UK; Mirjam Geerlings [Marileen Portegies] University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, the Netherlands; Clifford Jack [Chadwick Ward] Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN; Hilkka Soininen [Yawu Liu] Dept of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland; Ron Killiany [Corinna Bauer] Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA; Stefan Teipel [Michel Grothe] German Center for Neurodegenerative Diseases, Rostock; Jeffrey Kaye [Tim Swihart] Oregon Health and Science University, Portland, OR; Hiroshi Matsuda [Masami Nishikawa] Kawamura Gakuen Woman's University, Abiko-city, Japan; Gunhild Waldemar [Kristian Frederiksen] Memory Disorders Research Group, Dept. of Neurology, Rigshospitalet, Copenhagen, Denmark.)

3.2. Statistical results

To test the increase in compliance between phases, we performed a repeated measures analysis of the Jaccard overlap statistic, on a per-image basis, averaged over all tracers that completed Phases II and III, and tracers that completed Phases I, II, and III. From the experimental design we therefore had access to four images (eight hippocampi) that were segmented twice, and two images (four hippocampi) that were segmented three times.

Statistical testing of training with two phases showed a significant effect of Jaccard ($P < .0001$) (i.e. Jaccard overlap increased significantly between phases for all images, on average for all raters), and a significant effect for SIDE ($P < .001$) for all variables except one (i.e., there was a difference between performance between the left and right hippocampi) (cf. Fig. 4). Testing for those raters that

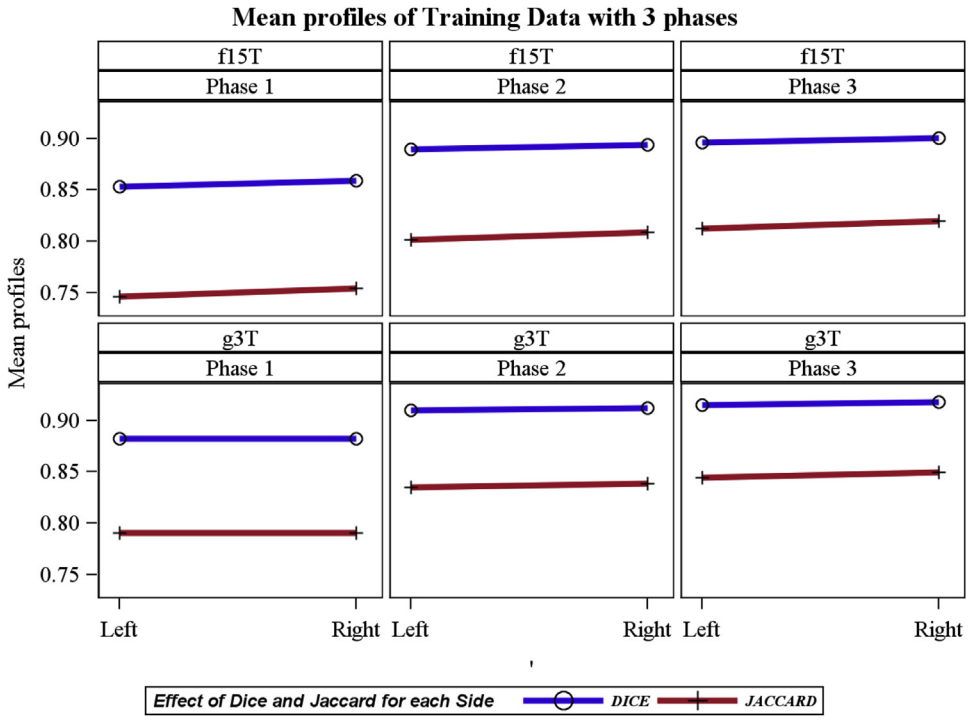

Fig. 5. Similar repeated measures analysis of Jaccard indices for N = 13 new raters on the two images segmented as part of Phases I, II, and III. Results show again a significant increase between phases, but a disappearance of the hemispheric effect, indicative of convergence to the Harmonized Protocol.

performed all three phases for those selected images which were present in each phase again showed a significant effect for Jaccard overlap ($P < .0001$), but SIDE fell below significance ($P > .05$) (cf. Fig. 5).

Results from the repeated effects model (mean effect with confidence intervals on Jaccard index between phases) are shown in Table 1.

3.3. Qualification thresholds

Based on tracers' initial training, and on the visual evaluation of compliance of segmentations with the HarP, we estimated thresholds for qualification as follows:

a) Tracer total volume for any hippocampi must fall within minimum and maximum masters' volumes;
b) Minimum Jaccard similarity index for any hippocampus >0.75;
c) Maximum distance ratio summation for any hippocampus <15; and
d) Average distance ratio summation for all hippocampi <20.

All tracers met these thresholds in the qualification set (cf. Table 2).

Table 1
Results from the repeated effects model

Variable	*P* value (95% CI)
Two phases	
Subject A: Phase 2, phase 3	<.0001 (−0.02535, −0.01491)
Subject B: Phase 2, phase 3	<.0001 (−0.05400, −0.03098)
Subject C: Phase 2, phase 3	<.0001 (−0.04624, −0.03185)
Subject A: Phase 2, phase 3	<.0001 (−0.03477, −0.02184)
Three phases	
Subject E: Phase 1 vs phase 2	<.0001 (−0.05406, −0.03691)
Subject E: Phase 1 vs phase 3	<.0001 (−0.06054, −0.04298)
Subject E: Phase 2 vs phase 3	.1541 (−.01506, 0.02497)
Subject F: Phase 1 vs phase 2	<.0001 (−0.04347, −0.03092)
Subject F: Phase 1 vs phase 3	<.0001 (−0.05171, −0.03887)
Subject F: Phase 2 vs phase 3	.0153 (−0.01452, −0.00168)

4. Discussion

4.1. Scope

We have developed a web-based Qualification Platform for the training of new tracers on the HarP for the segmentation of the hippocampus on MRI, including automated feedback and qualification features. This online system can provide standard qualitative and quantitative results in comparison with benchmark images, allowing to go through the same training and qualification procedures from different remote laboratories. This kind of service is required to guarantee a homogeneous performance in the hippocampal segmentation performed in different laboratories, and thus to implement the practical use of hippocampal volumetry as a biomarker for AD. Although the system will shortly be improved thanks to additional benchmark segmentations that have recently been completed [5], the platform demonstrated to work in the validation phase of the

Table 2
Naive tracer qualification results

Tracer	Jaccard	Dice
14	0.85	0.92
20	0.83	0.91
21	0.82	0.90
11	0.81	0.90
16	0.81	0.90
2	0.81	0.89
19	0.80	0.89
18	0.80	0.89
4	0.80	0.89
17	0.80	0.89
6	0.79	0.88
10	0.78	0.88
13	0.78	0.88

whole EADC-ADNI project for the Harmonization of Hippocampal Protocols [9], and is already freely available and accessible from the official website of the project at www.hippocampal-protocol.net.

4.2. Outcome

The main outcomes of the work described in this article are threefold: (1) the creation and release of the infrastructure allowing standard centralized training for HarP tracers; (2) the demonstration that the performance of our tracers improved through the three training phases; and (3) the definition of thresholds for compliance with the HarP, based on tracers performance and on a qualitative evaluation of compliance.

4.3. Tracers improvement through the three training phases

Statistical analysis has shown that the effect of training positively increased the compliance with the HarP and therefore served to reduce between-rater variance. The training paradigm, namely that users had to segment two images in the first phase, then four new images in both of the subsequent rounds, was also validated by the fact that three phases were required before the effect of side (in essence a pseudorandomization of results) was removed. It is, therefore, recommended to maintain all three phases of training to increase the rater's chance of complying with the HarP.

The fact that the side effect was no longer significant in images that have been segmented three times could be explained by the fact that there were limited discrepancies between hippocampi from the outset. In the four images segmented twice, the difference between left and right is far more pronounced, and thus this difference may still exist if they were segmented for a third time.

The group of tracers that completed all training and qualification phases had then been involved in the validation phase of the HarP project, as described in ref. [9].

4.4. Thresholds denoting compliance of hippocampal segmentations with the HarP

We selected qualification thresholds on the current naive tracers, which is a relatively recursive situation for this group. However, these thresholds were defined not only based on their performance, but also after a visual evaluation of qualitative compliance of segmentations on the key features defined in the HarP and considering the range of statistics associated with good compliance. The defined thresholds are also determined by a ceiling effect beyond which a degree of error cannot be ruled out in tracers who had step by step training, but still perform manual segmentation. New tracers on the other hand will need to fulfill the same criteria, following the same procedure described in this article. As our knowledge evolves, these criteria may be modified to reflect growing expertise in using the platform. For example, although we have elected to base our qualification criteria on volumes, Jaccard similarity index, and distance ratio computations, we computed other metrics that can be used for similar purposes (cf. Appendix A). All these metrics are measured–where applicable–on the minimum, mean and maximum masters' contours, slice by slice. Thus, a different metric combination can be proposed. This could also include different thresholds applied to various key regions, e.g. tighter controls on the hippocampal tails, by identifying a specific set of slices.

4.5. Ongoing relevance

With this platform, we have provided the infrastructure to train and define how precise is the performance of a new tracer, not just in terms of volume consistency coefficients as traditionally done, but in terms of much more accurate statistics that are never routinely used for assessing manual hippocampal segmentation and allow point by point and slice by slice evaluation. We provided accuracy estimates (Jaccard, spatial distance) that allow a much stricter evaluation of tracers performance and demonstrated that new tracers are able to approach expert tracers with a precision and reproducibility that approaches that demonstrated by automated algorithms. The collective interrater coefficient across the 14 tracers (one expert rater, and 13 new tracers qualified in this platform) in the validation phase (which included different ADNI subjects than those used for the platform) was close or beyond 0.90 even when using an "absolute" method to assess it. If we consider the tracers selected among those with best performance for the subsequent parts of the harmonization project, their reliability was even greater, and, to our knowledge, never observed so far for manual tracers. In particular, the tracers who took part to the II phase of the validation study described in ref. [9] had the following values A: individual Jaccard in the qualification phase described in this study, B: absolute volume intrarater ICC for 1.5 T as described in [9], C:

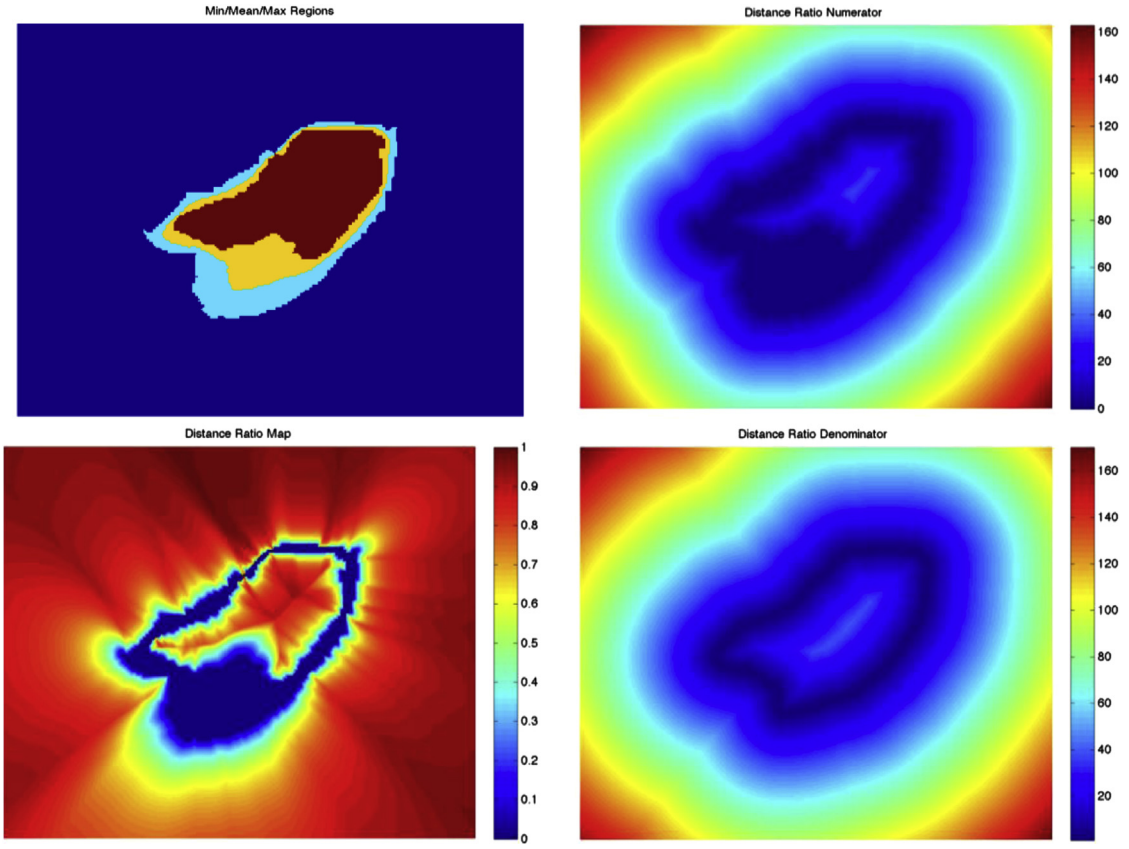

Fig. 6. (Top left) Master tracers' minimum (red), mean (yellow), and maximum (cyan) regions for a given hippocampal slice; (top right) the corresponding distance map for the maximum contour, used as the numerator. Colors in this Chamfer distance map represent arbitary distance units from contours; (bottom right) the corresponding distance map for the minimum contour, used as the denominator, with a similar color coding scheme; and (bottom left) the distance map ratio *D*, expressed as a continuous variable between 0 and 1.

absolute volume intrarater ICC for 3T images as described in [9]: Tracer "4": A:0.796, B: 0.942, C: 0.960; Tracer "18": A: 0.799, B: 0.968, C: 0.980; Tracer "19": A: 0.801, B: 0.986, C: 0.993; Tracer "16": A: 0.811, B: 0.993, C: 0.985. The tracers who were involved in the generation of additional HarP benchmark labels [5] had individual Jaccard in the qualification phase described in this study: 0.850, 0.831, and collective absolute volume inter-rater ICC with other three HarP expert tracers: left hippocampus = 0.953; right = 0.975.

The practical implications of such results are of paramount importance for the use of hippocampal volumetry in AD. First, the manual segmentation has always been considered liable to a certain degree of subjectivity. This study shows that, through the platform that we have produced and released, human tracers can achieve the same segmentation intra-rater reliability that is comparable with those characterizing automated algorithms. The investigation of the sources of variance in HarP segmentations performed in [9] showed that the factor "tracer" explained the smallest percentage of variability compared with the other factors (subject, atrophy, scanner, magnet field strength, side), and namely it explained 0.9% of the whole variability. This corresponds to a coefficient of variation of 2.4%, which is notably lower than the coefficient of variation known to be associated to the batches of reagents used in different laboratories for quantifying AD biomarkers from cerebrospinal fluid. These range from 13% to 36% [19], although plasma biomarkers are associated to even lower reliability [20]. This finding on one side attests the validity of the training system that we described in this article; on the other side it shows that the final results of the whole harmonization process, of which the infrastructure described here is a key element, concretely allows to use hippocampal volumetry as a biomarker for AD diagnosis. This is important to enable the comparability between different clinical trials currently searching for disease-modifying drugs, and will be essential for clinical diagnosis and progression monitoring as soon as such drugs will be available for patients.

Segmentations for these studies will be performed in the future mainly by automated algorithms. The importance of this study and of the whole harmonization project consisted in defining a standard protocol for manual segmentation, to get different tracers (either human or automated) aligned to the same standard [21]. Notwithstanding the very high intrarater reliability of automated segmentation, the gold standard has always consisted in manual tracing, given that the structure is defined neuroanatomically. In fact,

algorithms must be validated versus a segmentation template built in such a way that structure boundaries are defined and guaranteed by an expert of cerebral anatomy [22–25]. Automated systems so far are not yet able to define, in a digitalized MR, whether the single gray matter voxel belongs to the amygdala, to the entorhinal cortex, choroid plexus, isthmus, indusium griseum, or to the hippocampus. They can do reliably, however, when validated versus segmentations performed by expert tracers, and alignment to a single standard segmentation method is what is guaranteed by the platform generated with this work.

4.6. Limitations

The platform as presented is geared toward measuring the compliance of manual raters with the HarP, whereby tracing generates two-dimensional (2D) contours on coronal, subvoxellized images on a slice by slice basis in pseudo-Talairach space. It does not provide for the training and testing of results from automated algorithms, which typically generate 3D, voxellized objects. First, a learning set will be provided; this is the purpose of the so called "label expansion project", in which a set of 270 manually segmented harmonized hippocampal labels from 135 individuals representing wider physiological variability than the current group is being released to the community to train and test automated segmentation algorithms (cf. "Training Labels" article by Boccardi et al. in this Special Issue [26]). Second, the testing of voxellized labels will require the adaptation of the current platform metrics, from 2D to 3D and contours to objects, albeit in keeping with the same paradigm (i.e. volumes, spatial overlap, and spatial surface distance).

As to the qualification of manual raters, the first limitation consists in the separation of the MRI to be segmented into training and qualification sets. Although in this instance subjects were present in both (albeit at different field strengths), in future implementations they will be separate. A second limitation, as mentioned previously, is with respect to the qualification thresholds, which will be iteratively refined with time and experience. Notwithstanding both these limitations, the subsequent performance of the tracers trained through this platform was very good as reported in the article describing the validation of the HarP in this special issue [9], and as described previously.

4.7. Conclusions

The platform provided with this part of the EADC-ADNI project on the HarPs for hippocampal segmentation enabled remote participants to get trained and qualified as HarP tracers through a standard procedure, that demonstrated very good results in the qualification phase described in this article and, mainly, in the subsequent validation of the HarP. The standard training provided with our infrastructure is thus a key step supporting the concrete implementation of hippocampal volumetry for the diagnosis of AD in clinical and research settings worldwide.

Acknowledgments

The Alzheimer's Association has provided logistic support for update meetings of the HarP project. Wyeth, part of the Pfizer group, and Lilly have provided unrestricted grants in support of the HarP project. A follow-up project has been funded by the Alzheimer's Association: "A Harmonized Protocol for Hippocampal Volumetry: an EADC-ADNI Effort", grant n. IIRG -10-174022.

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 P.I.) 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 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); NEUROMED, Department of Neuroimaging, King's College London, London, United Kingdom (A Simmons). ADNI Centres are: Mayo Clinic, Rochester, MN (CR 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 NeuroImaging (LoNI), University of California, Los Angeles (UCLA), CA (LG Apostolova); University of Southern California, Los Angeles, 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); Institute for Ageing and Health, Wolfson Research Centre, Newcastle General Hospital, Newcastle, United Kingdom (J O'Brien). Population-based Studies

participants: PATH through life, Australia (P Sachdev and JJ Maller); SMART-Medea Study, University Medical Center Utrecht, The Netherlands (MI Geerlings); Rotterdam Scan Study, The Netherlands (T denHeijer); L Launer, National Institute on Aging (NIA), 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); MNI, McGill University, Montreal, Canada (L Collins). Advisors: Clinical issues: PJ 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 & Education: G Waldemar, Copenhagen University Hospital, Copenhagen, Denmark.

The MR images used in this paper belong to the ADNI dataset. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; 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; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SD, FV, AV, and NR have received funding support from the Ministère du Développement Économique, de l'Innovation et de l'Exportation du Québec (PSR-SIIRI 631), and the Alzheimer's Society of Canada (#13 32). S.D. is a Junior 1 Research Scholar from the Fonds de Recherche Québec–Santé (#22424).

Disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Role of the funding source: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors′ contributions:

- Guarantors of integrity of entire study: all authors;
- Study concepts and design: all authors;
- Literature research: S.D., N.R., M.B.;
- Clinical studies: ADNI;
- Data acquisition and processing: ADNI, M.Bocchetta, M. Boccardi, L.A.G., R.G., G.P., D.W.;
- Methods, analysis and interpretation: S.D.; F.V; F.A.V., N. R.;
- Statistical analysis: S.D.; N.R., A.M.;
- Manuscript preparation: S.D., N.R., A.M.; revision/review, all authors; and.
- Manuscript definition of intellectual content, editing, and final version approval: all authors.

RESEARCH IN CONTEXT

1. Systematic review: Within the context of AD, hippocampal volumetry is an in vivo biomarker of major interest. The Harmonized Protocol for Hippocampal Segmentation Project is an effort undertaken by European Alzheimer's Disease Consortium and Alzheimer's Disease Neuroimaging Initiative centers for the manual segmentation of the hippocampus on magnetic resonance scans. "master tracers" segmented a set of benchmark images according to the consensual definition obtained following evidence-based Delphi panels.
2. Interpretation: This article describes our work toward implementing an interactive web system allowing *protocol learning*, segmentation *training*, and periodical *qualification* of the ability of new tracers to segment the hippocampus according to the HarP. We demonstrate that the training process embedded in the platform led to increased compliance with the HarP.
3. Future directions: The platform is geared toward measuring compliance of manual raters with the protocol, however, a thorough statistical validation must be performed to determine metric qualification thresholds for new users. Furthermore, the training set will need to substantially expand and metrics adapted for automated algorithms qualification.

References

[1] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging and the Alzheimer's Association workgroup. Alzheimers Dement 2011;7:263–9.

[2] 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.

[3] 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.

[4] Committee for Medicinal Products for Human Use. 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; 17 November 2011.

[5] Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27:685–91.

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

[7] 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.

[8] 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.

[9] Frisoni GB, Jack CR Jr, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, et al. The EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance: evidence of validity. Alzheimers Dement 2015;11:111–25.

[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;11:192–202.

[11] Kaye JA, Swihart T, Howieson D, Dame A, Moore MM, Karnos T, et al. Volume loss of the hippocampus and temporal lobe in healthy elderly persons destined to develop dementia. Neurology 1997; 48:1297–304.

[12] 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.

[13] 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.

[14] Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess 1998;2:1–88.

[15] Boccardi M, Bocchetta M, Apostolovac 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.

[16] Bocchetta M, Boccardi M, Ganzola R, Apostolova LG, Preboske G, Wolf D, et al. Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project. Alzheimers Dement 2015;11:151–65.

[17] Jannin P, Grova C, Maurer CR. Model for defining and reporting reference-based validation protocols in medical image processing. Int J Comp Assist Radiol Surg 2006;1:63–73.

[18] 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.

[19] 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.

[20] 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. J Alzheimers Dis 2012;32:939–47.

[21] 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.

[22] Duchesne S. Appearance-based segmentation of medial temporal lobe structures. NeuroImage 2002;17:515–31.

[23] Hasan KM, Pedraza O. Improving the reliability of manual and automated methods for hippocampal and amygdala volume measurements. Neuroimage 2009;48:497–8.

[24] Barnes J, Foster J, Boyes RG, Pepple T, Moore EK. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. NeuroImage 2008;40:1655–71.

[25] Brewer JB, Magda S, Airriess C, Smith ME. Fully-automated quantification of regional brain volumes for improved detection of focal atrophy in Alzheimer disease. AJNR Am J Neuroradiol 2009;30:578–80.

[26] Boccardi M, Bocchetta M, Morency FC, Collins DL, Nishikawa M, Ganzola R, et al. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimers Dement 2015;11:175–83.

Appendix A
Statistical measures available from the qualification platform

Volumes/areas

- Tracer volume (mm^3): total hippocampal volume obtained by summing the slice area multiplied by thickness.
- Tracer area (mm^2): area delimited by the tracer contour for the current slice.
- Master minimum area (mm^2): area commonly delimited by all master tracers for the current slice.
- Master mean area (mm^2): mean area delimited by all master tracers for the current slice.
- Master maximum area (mm^2): area maximally delimited by any master tracer for the current slice.
- True positive area w.r.t. min (mm^2), true positive area w.r.t. mean (mm^2), true positive area w.r.t. max (mm^2): tracer true positive area with respect to the area delimited by the minimum/mean/maximum master contour for the current slice.
- False negative area w.r.t. min (mm^2), false negative area w.r.t. mean (mm^2), false negative area w.r.t. max (mm^2): tracer false negative area with respect to the area delimited by the minimum/mean/maximum master contour for the current slice.
- False positive area w.r.t. min (mm^2), false positive area w.r.t. mean (mm^2), false positive area w.r.t. max (mm^2): tracer false positive area with respect to the area delimited by the minimum/mean/maximum master contour for the current slice.

Overlap measures

- Jaccard similarity index (see definition previously) between the areas delimited by the tracer and master minimum/mean/maximum contours for the current slice.
- Jaccard Index w.r.t. min, Jaccard Index w.r.t. mean, Jaccard Index w.r.t. max.
- Dice similarity index between the areas delimited by the tracer and master minimum/mean/maximum contours for the current slice.
- Dice Index w.r.t. min, Dice Index w.r.t. mean, Dice Index w.r.t. max.

Contour distances

- Distance integral: Line integral of the distance ratio (see definition previously for "distance ratio") values along the tracer contour, normalized by the contour's number of point, for the current slice.
- Hausdorff distance (mm): Hausdorff distance between the tracer contour and the master mean contour for the current slice.

Appendix B: Qualitative and quantitative performance estimates description

The Harmonized Protocol (HarP) is a consensual but theoretical neuroanatomical protocol. Once implemented however, interpretations vary slightly. Thus, by definition, each master tracer was correct in her or his interpretation of the HarP and of the boundaries it represents. Various rounds of consensus were performed between master tracers to ensure maximum convergence, however some variability remained. Thus, we could not discard the information provided by any one of the master's tracing. We therefore elected to compute the maximum contour as the outer hull of all master's tracings; and the minimum contour as the area of commonality between all master's. Thus, by definition, any contour submitted within the maximum and minimum set by the master's tracers is considered correct.

To assess tracing performance, however, we selected the following three metrics, each capturing one aspect of segmentation accuracy (additional—but correlated—metrics have been measured, and are listed in Appendix A), and defined with respect to either the maximum, mean, or minimum master's contours:

- Hippocampal volumes: We calculated total HC volumes stereologically by multiplying the segmented area on any given slice by its slice thickness, and summing up these partial volumes. New tracers volumes can be compared with the average masters' volume for that hippocampus on a pairwise basis.
- Spatial overlap: Although segmentations may have similar volumes to be accurate they must significantly overlap. To capture this variability, we calculated the Jaccard similarity index as a metric of spatial overlap. The Jaccard similarity index between regions A and B is given by

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}$$

For purposes of comparison, we calculated the Jaccard similarity index between the new tracers 'contour and the mean masters' contour. The final measure used in statistical tests is the average Jaccard similarity index over all slices for a given hippocampus.

- Spatial distance: To ensure further compliance with the definitions set forth in the HarP, we required a distance metric to assess whether the new tracers were within the minimum and maximum boundaries defined by the masters.

To this end we first computed a distance ratio map from the Euclidean distance maps of the regions delimited by the masters' minimum, mean, and maximum contours.

The distance ratio map D is given by

$$D = \frac{\mathrm{dist}\left(\overline{R_{\mathrm{min}}}\right) + \mathrm{dist}(R_{\mathrm{max}})}{\mathrm{dist}(R_{\mathrm{mean}}) + \mathrm{dist}\left(\overline{R_{\mathrm{mean}}}\right)}$$

where $\mathrm{dist}(R)$ and $\mathrm{dist}(\overline{R})$ are maps of the Euclidean distances calculated from the binary region R and the binary inverse of R (i.e. $\overline{R}$) (also known as Chamfer distance maps). In other words, $\mathrm{dist}(R)$ is the map of the distances from the region border toward the outside R, and $\mathrm{dist}(\overline{R})$ is the map of the distances from the region border toward the inside of R. Note that the distance maps are images and the summations and division in D are performed element by element.

In Fig. 6, we show the master minimum, mean, and maximum regions, an example of the numerator and denominator of D, and the map D. We note that when the distance between the masters' minimum and maximum contours is large, the values of D increase slowly as we move from the masters' contours. However, when this distance is short the values of D increase rapidly.

The distance ratio values are bound between [0, 1]. From the distance ratio map D, the distance ratio is interpolated for each point of the tracer contour, and a color (from green to yellow to red) is assigned according to the distance ratio value (from 0 to 1) in the contour plots. This provided useful feedback to tracers who could rapidly gauge their compliance with the HarP, when provided with this information (cf. example in Fig. 2).

A value of D equal to 0 means that we are inside the boundaries delimited by the masters' minimum and maximum contours, and hence by definition in agreement with the HarP. A value of D equal to 1 means that the distance from the mean contour is at least equal to or greater than the distance from the minimum or maximum contour. The value of D thus tends to 1 when a given point is far from the masters' contours (hence, in strong disagreement with the HarP) or when the masters' minimum, mean, and maximum contours are very close (and hence, users should be more careful in strictly adhering to landmark definitions).

The distance ratio map D thus gives a distance value that is weighted according to the distance between the masters' minimum and maximum contours. The values of D are thus adapted to capturing agreement with master tracers, and hence adherence to the HarP. The final measure used in statistical testing consists in the summation of distance ratios for each contour point for a new tracer's contour, averaged over all slices for a particular hippocampus.