Alzheimer's & Dementia 11 (2015) 111-125

Alzheimer's & Dementia

Project Report

The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity

Giovanni B. Frisoni[a,b], Clifford R. Jack, Jr.,[c], Martina Bocchetta[a,d], Corinna Bauer[e], Kristian S. Frederiksen[f], Yawu Liu[g], Gregory Preboske[c], Tim Swihart[h], Melanie Blair[i], Enrica Cavedo[a], Michel J. Grothe[j], Mariangela Lanfredi[k], Oliver Martinez[l], Masami Nishikawa[m], Marileen Portegies[n], Travis Stoub[o], Chadwich Ward[c], Liana G. Apostolova[p], Rossana Ganzola[q], Dominik Wolf[r], Frederik Barkhof[s], George Bartzokis[t], Charles DeCarli[l], John G. Csernansky[u], Leyla deToledo-Morrell[o], Mirjam I. Geerlings[n], Jeffrey Kaye[h], Ronald J. Killiany[e], Stephane Lehéricy[v], Hiroshi Matsuda[m], John O'Brien[w], Lisa C. Silbert[h], Philip Scheltens[x], Hilkka Soininen[g], Stefan Teipel[j,y], Gunhild Waldemar[f], Andreas Fellgiebel[r], Josephine Barnes[i], Michael Firbank[w], Lotte Gerritsen[n,z], Wouter Henneman[s], Nikolai Malykhin[aa], Jens C. Pruessner[bb], Lei Wang[cc], Craig Watson[l], Henrike Wolf[dd,ee], Mony deLeon[ff], Johannes Pantel[gg], Clarissa Ferrari[k], Paolo Bosco[a], Patrizio Pasqualetti[hh,ii], Simon Duchesne[q], Henri Duvernoy[jj], Marina Boccardi[a,*], for the EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Volumetry and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]*LENITEM (Laboratory of Epidemiology, Neuroimaging and Telemedicine) IRCCS – Istituto Centro S. Giovanni di Dio – Fatebenefratelli, Brescia, Italy*
[b]*Memory Clinic and LANVIE - Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland*
[c]*Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN, USA*
[d]*Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy*
[e]*Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA, USA*

Project Collaborators include: Marilyn S. Albert (Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA), David Bennet (Department of Neurological Sciences, Rush University, Chicago, Illinois), Richard Camicioli (Department of Biomedical Engineering, Centre for Neuroscience, University of Alberta, Edmonton, Alberta, Canada), D. Louis Collins (MNI, McGill University, Montreal, Canada), Bruno Dubois (Service de Neuroradiologie, Hopital de la Pitie-Salpetriere, Paris, France), Harald Hampel (Department of Psychiatry, University of Frankfurt, Germany), Tom denHeijer (Department of Neurology, Sint Franciscus Gasthuis, Rotterdam, The Netherlands), Christofer Hock (Northwestern University Feinberg School of Medicine, Department of Psychiatry and Behavioral Sciences, United States), William Jagust (University of California, Berkeley, CA), Leonore Launer (Neuroepidemiology Section, Intramural Research Program, NIA/NIH, Bethesda MD 20892), Jerome J. Maller (Monash Alfred Psychiatry Research Centre, Melbourne, Australia), Susan Mueller (University of California (UCSF), San Francisco, CA; Perminder Sachdev (Centre for Healthy Brain Ageing (CHeBA), University of New South Wales, Sydney, Australia), Andy Simmons (Department of Neuroimaging, Institute of Psychiatry, King's College London, London, UK), Paul M. Thompson (Imaging Genetics Center, Institute for Neuroimaging and Informatics, Depts. of Neurology, Psychiatry, Engineering, Radiology, & Ophthalmology, Keck/USC School of Medicine, University of Southern California, Los Angeles, CA, USA), Peter-Jelle Visser (Department of Neurology and Alzheimer Center, VU University Medical Cente and Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; Department of Psychiatry and Neuropsychology, Maastricht University, School for Mental Health and Neuroscience, Alzheimer Center Limburg, Maastricht, the Netherlands), Lars-Olof Wahlund (Division of Clinical Geriatrics, NVS Department, Karolinska Institutet, Stockholm, Sweden), Michael W. Weiner (University of California (UCSF), San Francisco, CA; San Francisco Veterans Affairs Medical Center, San Francisco, CA), Bengt Winblad (KI Alzheimer Disease Research Center, NVS Department, Karolinska Institutet, Stockholm).

The manuscript received ADNI approval on October 16, 2013.

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

*Corresponding author. Tel.: +39-(0)303501553; Fax: +39-0303501592.

E-mail address: mboccardifbf@gmail.com

http://dx.doi.org/10.1016/j.jalz.2014.05.1756

[f]*Memory Disorders Research Group, Department of Neurology, Rigshospitalet, Copenhagen, Denmark*
[g]*Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland*
[h]*Department of Neurology, Oregon Health & Science University, Portland, OR, USA*
[i]*Dementia Research Centre, Department of Neurodegenerative Disease, UCL Institute of Neurology, National Hospital for Neurology and Neurosurgery, London, UK*
[j]*German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*
[k]*Unit of Psychiatry, IRCCS – Centro S. Giovanni di Dio – Fatebenefratelli, Brescia, Italy*
[l]*Department of Neurology, University of California, Davis, CA, USA*
[m]*Kawamura Gakuen Woman's University, Abiko-city, Japan*
[n]*University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, The Netherlands*
[o]*Department of Neurological Sciences, Rush University, Chicago, IL, USA*
[p]*Mary S. Easton Center for Alzheimer's Disease Research and Laboratory of NeuroImaging, David Geffen School of Medicine, University of California, Los Angeles, CA, USA*
[q]*Department of Radiology, Université Laval and Centre de Recherche de l'Institut universitaire de santé mentale de Québec, Quebec City, Canada*
[r]*Klinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität Mainz, Mainz, Germany*
[s]*Department of Radiology and Nuclear Medicine, Image Analysis Center, VU University Medical Center, Amsterdam, The Netherlands*
[t]*Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA*
[u]*Washington University, Northwestern University, Chicago, IL, USA*
[v]*Service de Neuroradiologie, Hopital de la Pitie-Salpetriere, Paris, France*
[w]*Institute for Ageing and Health, Newcastle University, Newcastle upon Tyne, UK*
[x]*Department of Neurology and Alzheimer Center, VU University Medical Cente and Neuroscience Campus Amsterdam, Amsterdam, The Netherlands*
[y]*Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany*
[z]*Department of Medical epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden*
[aa]*Department of Biomedical Engineering, Centre for Neuroscience, University of Alberta, Edmonton, Alberta, Canada*
[bb]*Department of Psychiatry, McGill Centre for Studies in Aging, McGill University, Montreal, Quebec, Canada*
[cc]*Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
[dd]*Department of Psychiatry Research and Geriatric Psychiatry, Psychiatric University Hospitals, University of Zurich, Zurich, Switzerland*
[ee]*German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany*
[ff]*New York University School of Medicine, Center for Brain Health, New York, NY, USA*
[gg]*Institute of General Practice, Goethe-University Frankfurt, Frankfurt, Germany*
[hh]*SeSMIT (Service for Medical Statistics and Information Technology), AFaR (Fatebenefratelli Association for Research), Fatebenefratelli Hospital, Rome, Italy*
[ii]*Unit of Clinical and Molecular Epidemiology, IRCCS "San Raffaele Pisana", Rome, Italy*
[jj]*Chemin des Relançons, Besançon, France*

Abstract

Background: An international Delphi panel has defined a harmonized protocol (HarP) for the manual segmentation of the hippocampus on MR. The aim of this study is to study the concurrent validity of the HarP toward local protocols, and its major sources of variance.

Methods: Fourteen tracers segmented 10 Alzheimer's Disease Neuroimaging Initiative (ADNI) cases scanned at 1.5 T and 3T following local protocols, qualified for segmentation based on the HarP through a standard web-platform and resegmented following the HarP. The five most accurate tracers followed the HarP to segment 15 ADNI cases acquired at three time points on both 1.5 T and 3T.

Results: The agreement among tracers was relatively low with the local protocols (absolute left/right ICC 0.44/0.43) and much higher with the HarP (absolute left/right ICC 0.88/0.89). On the larger set of 15 cases, the HarP agreement within (left/right ICC range: 0.94/0.95 to 0.99/0.99) and among tracers (left/right ICC range: 0.89/0.90) was very high. The volume variance due to different tracers was 0.9% of the total, comparing favorably to variance due to scanner manufacturer (1.2), atrophy rates (3.5), hemispheric asymmetry (3.7), field strength (4.4), and significantly smaller than the variance due to atrophy (33.5%, $P < .001$), and physiological variability (49.2%, $P < .001$).

Conclusions: The HarP has high measurement stability compared with local segmentation protocols, and good reproducibility within and among human tracers. Hippocampi segmented with the HarP can be used as a reference for the qualification of human tracers and automated segmentation algorithms.

Keywords: Hippocampal volumetry; Magnetic resonance; Alzheimer's disease; Biomarkers; Diagnostic criteria; Enrichment; Clinical trials; Validation; Harmonized protocol; Standard operating procedures; Manual segmentation

1. Introduction

Hippocampal volume measured on single time point high resolution T1-weighted magnetic resonance (MR) images is a recognized biomarker of Alzheimer's disease (AD) [1]. Hippocampal atrophy is one of the core biomarkers in the revised National Institute on Aging-Alzheimer's Association (NIA-AA) diagnostic criteria for AD [2], and has been qualified by the European Medicines Agency for enrichment in regulatory clinical trials in the predementia stage of AD [3]. Qualification at the US Food and Drug Administration (FDA) is under way. Hippocampal atrophy rate is among the most sensitive markers of disease progression in AD [1,4–6] and is currently being used as a secondary outcome in a number of clinical trials with candidate disease modifiers [7].

Manual outlining by an expert rater is the most validated procedure used to estimate hippocampal atrophy [8]. Manual volumetry is also used as the standard against which automated segmentation algorithms are assessed [7]. Historically, different laboratories have used different anatomical landmarks and measurement procedures. Estimates of "normal" hippocampal volumes have differed as much as 2.5-fold [9]. The lack of an agreed reference procedure for manual volumetry is a major barrier to the widespread acceptance and the use of hippocampal volumetry for clinical diagnosis, disease tracking, and qualification of automated segmentation algorithms.

An international effort to harmonize existing protocols was funded by the Alzheimer's Association in 2010 following a smaller initial grant from pharmaceutical companies. The working group comprises 91 scientists from 38 research groups in four continents. The group began activities by surveying the protocols for manual hippocampal segmentation used in the AD literature, and the 12 most frequently cited were selected as the starting point for harmonization. The landmarks of the selected protocols were catalogued, semantics were harmonized, and the authors of the protocols were personally contacted to check appropriate interpretation [10]. Then we have reduced the highly variable and in some cases ill-defined landmarks defining the different protocols into a limited number of units, amenable to quantitative investigation. These have been named "Segmentation Units" [11]; as Lego blocks, different combinations of Segmentation Units allow to reconstruct the shapes of hippocampi segmented by different protocols. The four Segmentation Units (minimum hippocampus, alveus/fimbria, tail, and subiculum) so defined summarize and account for the whole landmarks variability of currently used protocols. Measurement properties of Segmentation Units were empirically estimated [11] and fed to a panel of 16 international experts (including protocols' authors) through a Delphi procedure. The experts were invited to answer questionnaires based on their experience and on the measurements provided. They were informed about the answers of other participants, and could iteratively vote and converge on a single combination of Segmentation Units, the EADC (European Alzheimer's Disease Consortium)-ADNI (Alzheimer's Disease Neuroimaging Initiative) Harmonized Protocol (HarP). Specifically, the Delphi panel converged on a protocol where all the most inclusive Segmentation Units were included. This means that a complete HarP hippocampal segmentation includes the whole hippocampal head, body, and tail; the alveus/fimbria, up to the most caudal slices, the whole subiculum, based on the visible morphology of its boundary with the entorhinal cortex, or on a horizontal line drawn from the top of the parahippocampal white matter, and the caudal tissue of the Andreas Retzius and fasciolar gyri, excluded as vestigial tissue from current protocols [12]. Five expert ("master") tracers then segmented 40 hippocampi following the HarP. These master segmentations were checked, corrected and certified as the "benchmark" labels to be used for the qualification of any future human tracer or automated segmentation procedure [13]. An online platform, freely available at www.hippocampal-protocol.net, was developed to qualify new ("naïve") tracers to the use of the HarP based on the benchmark labels [14].

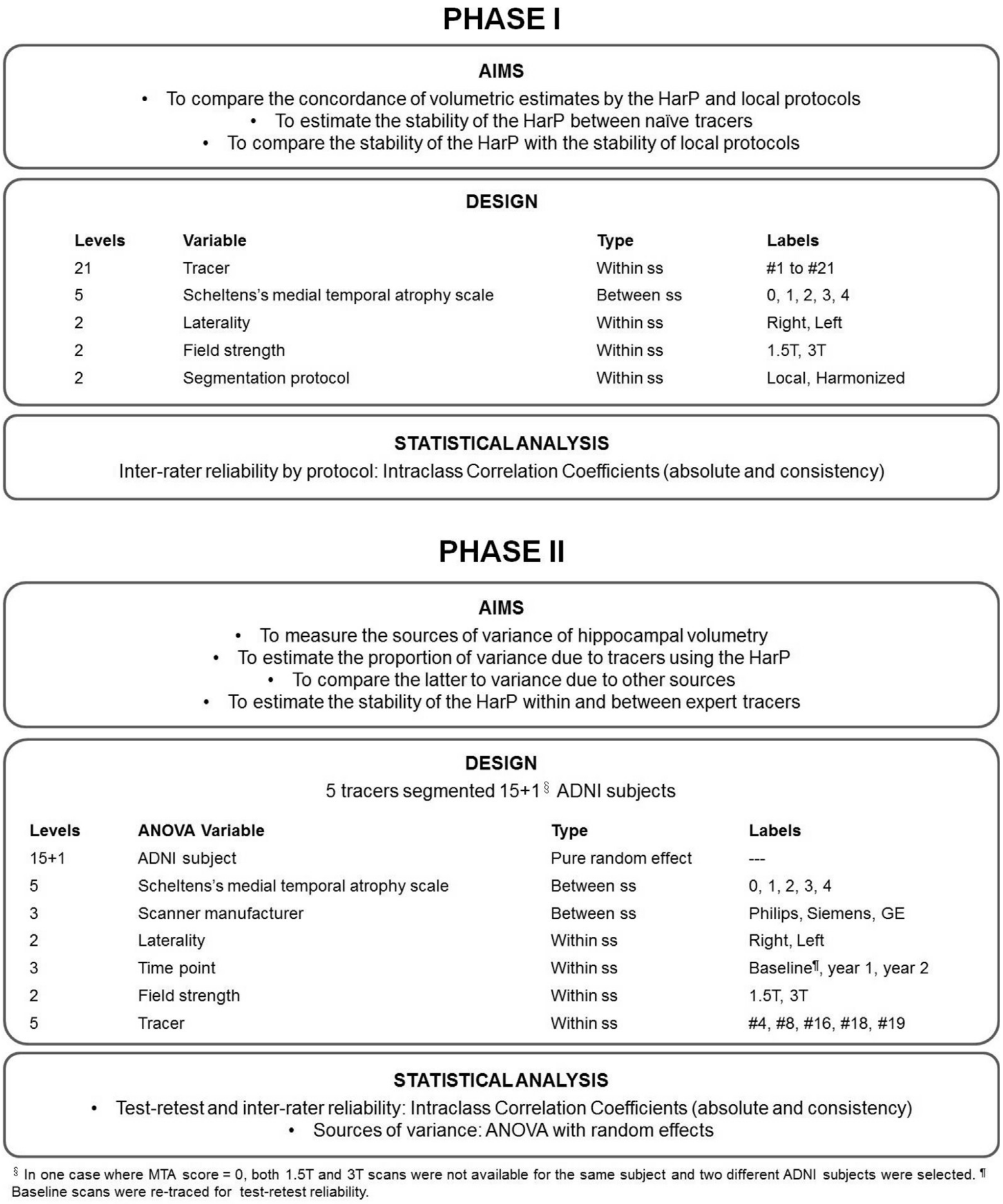
This report describes the final step of the initiative where the concurrent validity of the HarP was compared with local protocols (Phase I), and the major sources of variance of hippocampal volumes segmented with the HarP were estimated (Phase II). We hypothesized that agreement between raters would be greater with the HarP than with local protocols, and that the variability of the HarP-based segmentations due to different tracers would compare favorably to variability due to other sources.

2. Methods

This study was conceived in two logically sequential phases. In Phase I, 21 tracers naïve to the HarP and coming from different research centers were recruited to: segment 20 ADNI MR brain scans following the local segmentation protocol in use in their imaging laboratory: qualify for the HarP; and resegment the same hippocampi following the HarP. In Phase II, the five most accurate naïve tracers blindly resegmented the same images, allowing evaluation of test-retest intraclass correlation coefficient (ICC), and segmented an additional set of ADNI scans balanced for a number of variables, allowing to estimate the amount of variance due to the human tracers using the HarP and that due to other relevant factors (Figure 1).

2.1. MR Scans

Raw Medical Image NetCDF (MINC) 3D T1-weighted structural magnetic resonance (MR) images of 16 ADNI cases were downloaded from the ADNI database (www.adni.loni.usc.edu). Cases for Phase I were (ADNI IDs): 005_S_0324, 005_S_0814, 016_S_1121, 018_S_0335, 023_S_1046, 023_S_1190, 023_S_1262, 100_S_0190, 126_S_0605, 131_S_0441. The additional cases used for Phase II were 002_S_1018, 005_S_0572, 010_S_0422,



PHASE I

AIMS

- To compare the concordance of volumetric estimates by the HarP and local protocols
- To estimate the stability of the HarP between naïve tracers
- To compare the stability of the HarP with the stability of local protocols

DESIGN

Levels	Variable	Type	Labels
21	Tracer	Within ss	#1 to #21
5	Scheltens's medial temporal atrophy scale	Between ss	0, 1, 2, 3, 4
2	Laterality	Within ss	Right, Left
2	Field strength	Within ss	1.5T, 3T
2	Segmentation protocol	Within ss	Local, Harmonized

STATISTICAL ANALYSIS

Inter-rater reliability by protocol: Intraclass Correlation Coefficients (absolute and consistency)

PHASE II

AIMS

- To measure the sources of variance of hippocampal volumetry
- To estimate the proportion of variance due to tracers using the HarP
- To compare the latter to variance due to other sources
- To estimate the stability of the HarP within and between expert tracers

DESIGN

5 tracers segmented 15+1§ ADNI subjects

Levels	ANOVA Variable	Type	Labels
15+1	ADNI subject	Pure random effect	---
5	Scheltens's medial temporal atrophy scale	Between ss	0, 1, 2, 3, 4
3	Scanner manufacturer	Between ss	Philips, Siemens, GE
2	Laterality	Within ss	Right, Left
3	Time point	Within ss	Baseline¶, year 1, year 2
2	Field strength	Within ss	1.5T, 3T
5	Tracer	Within ss	#4, #8, #16, #18, #19

STATISTICAL ANALYSIS

- Test-retest and inter-rater reliability: Intraclass Correlation Coefficients (absolute and consistency)
- Sources of variance: ANOVA with random effects

§ In one case where MTA score = 0, both 1.5T and 3T scans were not available for the same subject and two different ADNI subjects were selected. ¶ Baseline scans were re-traced for test-retest reliability.

Fig. 1. Aim and design of the validation study of the EADC (European Alzheimer's Disease Consortium)-ADNI (Alzheimer's Disease Neuroimaging Initiative) Harmonized Protocol (HarP). *Phase I*: 21 naïve tracers segmented the hippocampi of 10 ADNI subjects (right and left, scanned at 1.5T and 3T, for a total of 40 hippocampi) balanced by atrophy following local protocols; qualified for the HarP; and then retraced the same hippocampi following the HarP. The total number of hippocampi was 80 per tracer. *Phase II*: the most accurate of the naïve tracers completing Phase I followed the HarP to segment 15 ADNI cases balanced by atrophy and scanner manufacturer and acquired at three time points at 1.5T and 3T; 20 baseline scans (already segmented in Phase I) were retraced in Phase II, while the other 10 baseline scans were blindly segmented twice within Phase II. The total number of hippocampi was 240 per tracer.

012_S_1009, 018_S_0450, 023_S_0625. Further information regarding the images used for the HarP validation are available at: www.centroalzheimer.it/public/SOPs/online/Appendix.doc. Cases were balanced by atrophy severity score on the Medial Temporal Lobe (MTA) scale [15] and scanner manufacturer (Figure 1 and Table 1). Diagnoses for Phase I were: three controls, three MCI and four AD. Phase II was meant to have five additional subjects and three time points at both 1.5 T and 3T with the same scanner manufacturer. No subjects with MTA equal to 0 were available having all three time points scans at both magnetic field strengths at Philips scanner manufacturer. Therefore, for MTA equal to 0 and "Philips" as scanner manufacturer, we selected two different subjects for the two magnet field strengths. For this reason, the total number of ADNI subjects is 16 (four controls, seven patients with MCI, and five with AD), rather than the 15 required by the experimental design.

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and

Table 1
Descriptive features of the ADNI (Alzheimer's Disease Neuroimaging Initiative) subjects selected for Phases I and II of the validation study

	MTA scale									
Phase I	0		1		2		3		4	
Age, yrs	73	77	56	76	73	75	72	79	71	84
Sex	F	F	F	F	F	F	M	M	F	F
ApoE genotype	33	33	33	34	33	33	34	34	34	23
Diagnosis	Ctr	Ctr	MCI	Ctr	MCI	AD	MCI	AD	AD	AD
Manufacturer	GE	Si	Si	GE	Si	GE	Si	Ph	GE	Ph

	MTA scale														
Phase II	0			1			2			3			4		
Age, yrs	62/71	73	77	56	76	76	69	73	75	72	79	79	71	76	84
Sex	M/F	F	F	F	F	M	M	F	F	M	M	M	F	M	F
ApoE genotype	33/33	33	33	33	34	44	34	33	33	34	44	34	34	33	23
Diagnosis	MCI/AD	Ctr	Ctr	MCI	Ctr	Ctr	MCI	MCI	AD	MCI	MCI	AD	AD	MCI	AD
Manufacturer	Ph/Ph	GE	Si	Si	GE	Ph	Ph	Si	GE	Si	GE	Ph	GE	Si	Ph

Abbreviations: MTA, medial temporal atrophy [16]; Ctr, healthy control; MCI, mild cognitive impairment; AD, Alzheimer's disease.

NOTE. In Phase II, no case was available satisfying the criteria of having MTA equal to 0 and being scanned on a Philips scanner at both 1.5T and 3T at three time points; thus, two different subjects, one scanned at 1.5T and one scanned at 3T, for three time points, were selected. Ph, Philips; GE, General Electric; Si, Siemens.

Bioengineering, FDA, private pharmaceutical companies, and nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI was to develop markers to track disease progression to be used in clinical trials of disease modifiers of early AD. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California–San Francisco. ADNI has recruited 200 cognitively normal individuals, 400 persons with MCI, and 200 with mild cognitive impairment aged 55 to 90, and followed them for at least 3 years. For up-to-date information, see www.adni-info.org.

2.2. *Phase I*

The 21 naïve tracers followed local protocols to segment the right and left hippocampi of 10 Phase I ADNI cases. The total number of hippocampi was 40 per tracer (2 subjects × 5 score of MTA scale × 2 field strength × 2 sides). The naïve tracers' and local protocols were (initials of the tracers, either listed among authors or in the acknowledgments, and reference in brackets): AC [16], CBa [17], CBo [18], EB [19,20], EC [21], FvD [22], KF [23], MB [24], MG [21], SH [unpublished], ML [21], YL [25], OM [unpublished], MN [21], MPo [26], GP [27], MPr [22], TSt [28], TSw [29], MT [30], CW [27].

To improve preprocessing homogeneity, the images were oriented along the anterior-posterior commissure (AC-PC) line using a 6 degree of freedom (DoF) function using the Montreal Neurological Institute (MNI) package AutoReg (version 0.98v) (www.bic.mni.mcgill.ca) and the MNI ICBM152 Nonlinear Symmetric template with 1 × 1 × 1 mm voxel dimensions as the reference. Resampling was carried out with a linear transformation with a linear interpolation scheme in AutoReg. However, the tracers were allowed to reorient images according to local protocols if these required a specific orientation (as was the case for CBa [17], FvD [22], KF [23], MB [24], SH [unpublished], YL [25], OM [unpublished], MPo [26], GP [27], MPr [22], TSt [28], MT [30], CW [29]). In these cases, tracers were recommended to use of a six DoF function without normalization or other preprocessing. All tracers were asked to use the same segmentation software (MultiTracer 1.0, http://www.loni.usc.edu/Software/MultiTracer, developed at the Laboratory of Neuro Imaging, LONI, at UCLA, Los Angeles, USA). Detailed instructions were provided covering all aspects of the segmentation process from image loading to volume computation. Qualification was initiated by sharing the HarP (Figure 2) in the form of a manual (see Appendix II of this Special Issue) providing detailed description of landmarks and segmentation procedures [12]. Naïve tracers also received instructions on how to create an account on the qualification platform (http://medics.crulrg.ulaval.ca/hippocampus), download the reoriented images (along the AC-PC line, as required by the HarP), segment the hippocampus using MultiTracer 1.0, save the segmentation files, and upload the segmented labels to the platform. Tracers were required to segment n = 10 images in three training rounds (n = 2, n = 4, n = 4). The platform provided color-coded visual feedback on compliance (or departure) from benchmark labels' segmentation, by color-coding the extent of departure on a red-to-green scale, where red denoted departure and green compliance. The platform also provided Dice and Jaccard measures of accuracy for each segmented slice and for the overall hippocampus [14]. At the end of the first two rounds, tracers received detailed written feedback from the project manager (M. Boccardi) illustrating HarP violations according to the color-coded visual feedback. Segmentation inaccuracies were

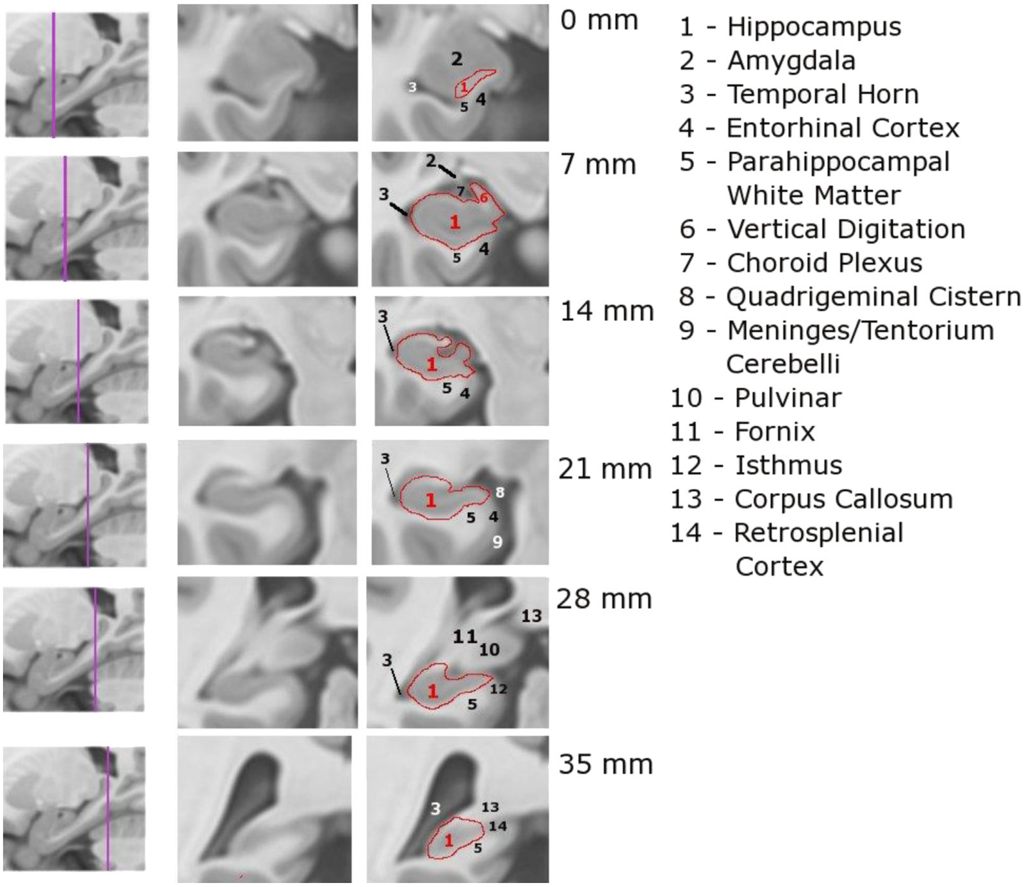


Fig. 2. The EADC (European Alzheimer's Disease Consortium)-ADNI (Alzheimer's Disease Neuroimaging Initiative) Harmonized Protocol (HarP): selected illustrative slices.

checked by the project manager, tracers were asked to edit segmentation inaccuracies if identified, re-upload edited images, and segment new scans. After 3 such training rounds, tracers were asked to segment and upload 10 new scans (qualification phase [14]). In the Qualification phase, tracers received only feedback regarding general performance, and corrections were not allowed. Tracers' performance on the Qualification phase featured high overlapping values: mean Dice (SD, range): 0.89 (0.01, 0.88–0.92); Jaccard: 0.81 (0.02, 0.78–0.85).

All of the original 21 tracers entered Phase I and segmented the images according to local protocols, but seven tracers did not enter or complete the qualification procedure due to withdrawal for logistical reasons (e.g. tracers changing job, or failure to complete segmentations within deadlines). Thirteen tracers successfully completed the qualification procedure and Phase I; one tracer was trained and had served as master tracer and did not need to undergo the qualification procedure. At the end of the qualification procedure tracers had very good reliability indices, with Dice values ranging from 0.88 to 0.92 for 3T images, and 0.87 to 0.91 for 1.5 T images [14]. These 14 tracers resegmented the same ADNI cases following the HarP (mean delay following local protocol segmentation of one year). In this article, only results from these tracers are shown for both Phases I and II. Tracers carried out segmentation with the same version of MultiTracer and the same settings. Each tracer used the same computer and monitor across Phases I and II. Tracers were required to segment in the coronal view magnified five times, while consulting the sagittal view magnified three times and the axial view with no magnification. This setting allowed tracers to visualize all hippocampi of the same size, at the same time fitting any computer screen. Magnification was kept constant throughout the segmentation. Segmentations were performed manually from rostral to caudal on approximately 30 to 35 contiguous coronal slices. Brain sections were 1 mm thick, and hippocampi were segmented on both the left and right sides.

2.3. *Phase II*

Five of the 14 tracers were selected to take part in Phase II of the study based on their accuracy during the qualification procedure (Jaccard of at least 0.80). One of these (GP) was trained and served as a "master" tracer after the segmentation based on local protocols. The others were chosen for having a Jaccard of at least 0.80, among those completing all segmentations of Validation Phase I. They were asked to use the HarP to segment 15 ADNI cases balanced by atrophy and scanner manufacturer and acquired at three time points at both 1.5 T and 3T (Figure 1 and Table 1). More exactly, for Phase II, tracers had to segment 3 subjects × 5 score of MTA scale × 2 field strength × 2 sides × 3 time points, and retrace the baseline sample: 3 subjects × 5 scores of MTA scale × 2 field strength × 2 sides. This led to a total of 180 + 60 (40 of them were already segmented during Phase I) = 240 hippocampi. Among these, 40 were blindly resegmented from Phase I, while 20 were blindly segmented twice within Phase II. Tracers were blinded to image code and clinical and socio-demographic features of the subjects all the times. Tracers were instructed to adhere to the procedures and software settings used in Phase I of the study.

2.4. Statistical analysis

Agreement within (test-retest) and between tracers (inter-rater reliability) was estimated with ICC and their 95% confidence intervals (95% CI) derived from the following two-way random analysis of variance (ANOVA):

$$H_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij},$$

where H_{ij} is the hippocampal volume of subject j segmented by tracer i, μ is the overall mean of tracers; α_i is the difference from μ of the mean of ith tracer (normal distributed with zero mean and variance $\sigma^2{}_\alpha$); β_j is the difference from μ of the jth subject (normal distributed with zero mean and variance $\sigma^2{}_\beta$); $(\alpha\beta)_{ij}$ is the degree to which the ith tracer departs from his/her rating tendencies when confronted by jth subject (normal distributed with zero mean and variance $\sigma^2{}_I$); and ε_{ij} is the random error (normal distributed with zero mean and variance $\sigma^2{}_\varepsilon$). In particular, the absolute ICC is given by the following ratio: $\sigma^2{}_\beta/(\sigma^2{}_\beta + \sigma^2{}_\alpha + \sigma^2{}_I + \sigma^2{}_\varepsilon)$ and it is estimated as the ratio of between subject mean square error and the total mean square error [31]. Consistency and absolute methods were used to compute ICCs. The difference between absolute and consistency agreement is defined in terms of how the systematic variability ($\sigma^2{}_I$) is treated. If that variability is considered irrelevant, it is not included in the denominator of the estimated ICCs, and measures of consistency are produced. If systematic differences among levels of tracers are considered relevant, rater variability contributes to the denominators of the ICC estimates, and measures of absolute agreement is produced.

Test-retest reliability was computed with a two-way random two-level ANOVA; inter-rater reliability was computed with a two-way 14-level (for Phase I) and five-level (for Phase II) random ANOVA. Analyses were carried out separately for 1.5 T and 3T images.

The amount of variance due to the HarP and other sources of variability was estimated with a multi-way ANOVA model with between and within factors (Figure 1). "Within" factors were modeled as nested random effects; ADNI case identifier was modeled as a pure random effect. The coefficient of variation was computed as the ratio of the standard deviation to the mean. The difference in ICC among protocols was tested with a Student's *t* test. Statistical analyses were performed using the SPSS software version 12.0 (http://www-01.ibm.com/software/analytics/spss/products/statistics/) and the R language v.2.13.0 (www.r-project.org).

3. Results

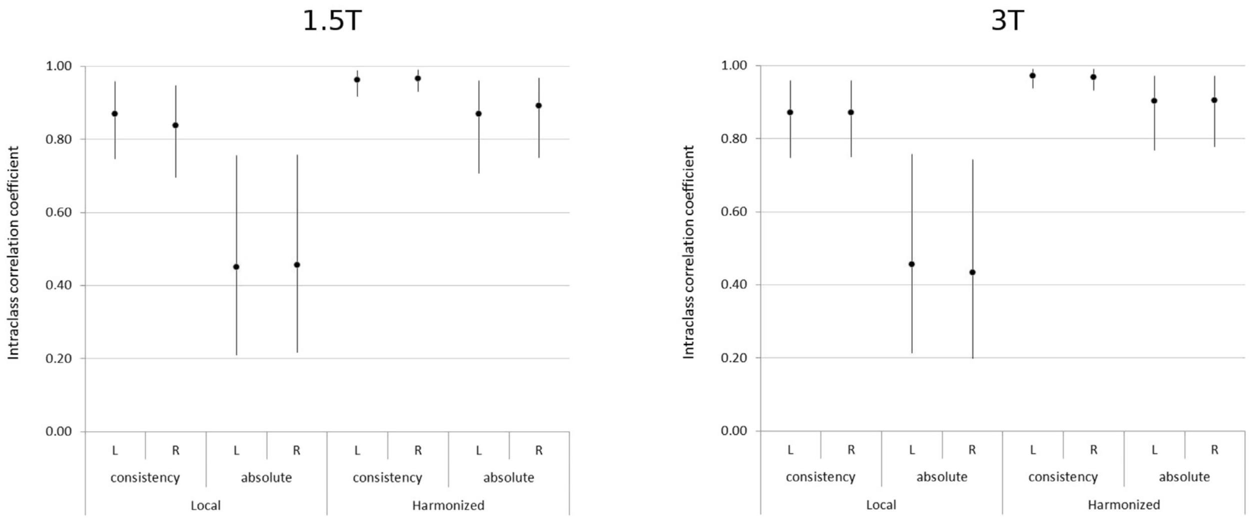
3.1. Phase I: concurrent validity of the HarP with local protocols.

Raw hippocampal volumes were higher for HarP segmentations (mean volumes across all subjects and tracers: left: 2781 mm^3, right: 2738 mm^3) than for local protocol segmentations (mean volumes across all subjects and tracers: left: 2143 mm^3, right: 2144 mm^3, Table 2). ICCs between tracers measured with the consistency method were in the mid-0.80s for the local protocols and high-0.90s for the HarP (Figure 3). The higher agreement of HarP segmentations was even more striking when agreement was estimated with absolute ICCs—in the mid-0.40s for the local protocols and high-0.80s for the HarP (Figure 3). Comparisons of homologous absolute ICCs between local and harmonized protocols were significant on t-test ($P < .01$ at 1.5 T and $P < .006$ at 3T), whereas consistency ICCs were not significantly different. The variability among segmentations based

Table 2
Mean hippocampal volumes computed for the right and left hippocampus for 10 ADNI (Alzheimer's Disease Neuroimaging Initiative) subjects, scanned at both 1.5T and 3T, and segmented based on local protocols and on the Harmonized Protocol (HarP)

	Local protocols		HarP	
Tracer	Left hippocampus	Right hippocampus	Left hippocampus	Right hippocampus
Tracer 2	2531.10	2545.77	2986.25	2951.12
Tracer 4	2454.81	2456.68	2922.38	2806.89
Tracer 6	2374.41	2351.98	2592.42	2536.09
Tracer 8	2590.64	2604.54	2776.08	2756.93
Tracer 10	2144.96	2113.38	2272.79	2283.27
Tracer 11	2092.75	2132.52	2812.33	2892.26
Tracer 13	2228.51	2195.47	3050.96	2873.79
Tracer 14	2528.05	2549.95	2685.96	2666.50
Tracer 16	419.66	450.40	2575.62	2604.04
Tracer 17	2283.06	2266.15	2991.21	2863.71
Tracer 18	2134.86	2164.29	2800.06	2741.31
Tracer 19	2294.80	2263.68	3071.77	3043.09
Tracer 20	2417.34	2427.94	2696.64	2590.98
Tracer 21	1504.23	1498.42	2701.79	2716.70
Mean	2142.80	2144.37	2781.16	2737.62
SD	782.12	773.03	713.277	684.27

NOTE. Values denote: mean hippocampal volume in mm^3 for each tracer who took part in the validation, global mean volume (and SD) for the right and left hippocampi computed including all ADNI subjects scanned at both 1.5T and 3T images for all tracers.



1.5T

		Local consistency L	Local consistency R	Local absolute L	Local absolute R	Harmonized consistency L	Harmonized consistency R	Harmonized absolute L	Harmonized absolute R
ICC		0.869	0.838	0.451	0.455	0.962	0.967	0.869	0.892
95% C.I.	inferior	0.746	0.696	0.211	0.218	0.918	0.930	0.707	0.750
	superior	0.958	0.947	0.755	0.757	0.988	0.990	0.960	0.968

3T

		Local consistency L	Local consistency R	Local absolute L	Local absolute R	Harmonized consistency L	Harmonized consistency R	Harmonized absolute L	Harmonized absolute R
ICC		0.871	0.871	0.455	0.433	0.971	0.968	0.902	0.904
95% C.I.	inferior	0.749	0.750	0.214	0.199	0.938	0.932	0.769	0.778
	superior	0.958	0.958	0.758	0.742	0.991	0.990	0.971	0.972

Fig. 3. Phase I: summary measures of the stability of the local and harmonized protocols among 14 naïve tracers (inter-rater ICC based on both absolute and consistency methods). ICC, intraclass correlation coefficient; CI, confidence interval. Comparisons of homologous absolute ICCs between local and harmonized protocols are significant on t-test at $P < .01$.

on local protocols with AC-PC orientation and HarP can be visually appreciated in Figure 4.

We computed inter-rater ICC among tracers using the same local protocols, to separate the components between tracer reliability and method variability. Tracers 2 and 8 used the same local protocol [27] and were from the same laboratory. Other four tracers (Tracers 6, 10, 14, and 20) used the same local protocol [21] but were from different laboratories. ICCs for these tracers were extremely high, with absolute ICC up to 0.899 for the four tracers coming from different laboratories, and absolute ICC up to 0.972 for the two coming from the same laboratory.

The extreme cases of agreement between pairs of tracers illustrate that in the case of best agreement between local protocols, the gain of using the HarP is marginal, while in the case of poorest agreement between local protocols, the gain is patent (Figure 5). A case of best agreement between local protocols occurred in the left hippocampus segmented at 3T between tracers #14 and #20 achieving an ICC of 0.971; when the same two tracers used the HarP, ICC marginally increased to 0.981. In contrast, the poorest case agreement between local protocols occurred in the left hippocampus segmented at 3T between tracers #4 and #16 achieving an ICC of 0.007; when the same two tracers used the HarP, their ICC increased to 0.922, indicating a critical and beneficial impact of the HarP on agreement between tracers.

3.2. Phase II: major sources of variance of the HarP.

The high stability of the HarP was confirmed by the five tracers taking part in Phase II of the study. Absolute figures of within tracer (test-retest) left/right ICC point estimates were in the mid- and high-0.90s. Measures of the absolute left and right ICC among all five tracers were in the high-0.80s to low-0.90s. Despite an obvious ceiling effect, consistency measures were generally even higher. Agreement tended to be slightly higher at 3T (ICC higher by about 0.02 units) though not statistically significant in any test (Table 3).

An ANOVA model including the sources of HarP variance listed in Table 4 accounted for 96.3% of the total variance, supporting its goodness-of-fit (Table 4). The largest proportion (82.7%) of this variance was due to inter-individual variability (the "case" factor) or atrophy. The residual variance (13.6%) was shared among scanner manufacturer, atrophy rate, hemispheric asymmetry (the "laterality" factor), "field strength", and "tracer". Of these, the factor accounting for the smallest proportion of variance was tracer (0.9%), indicating that when the HarP is used to measure hippocampal volume by reliable users, the "human factor" was less relevant than any other source of variability that we assessed here. It should be recognized, however, that the proportion of variance accounted for by "tracer" was not significantly different from the variance accounted for by "scanner manufacturer", "atrophy rates", "laterality", and "field strength", whereas it was significantly smaller than that accounted for by inter-individual variability and atrophy. Importantly, however, the coefficient of variation due to the factor "tracer" was very low (2.4%).

4. Discussion

We found that the use of the EADC-ADNI HarP for manual segmentation on MR scans was extremely stable

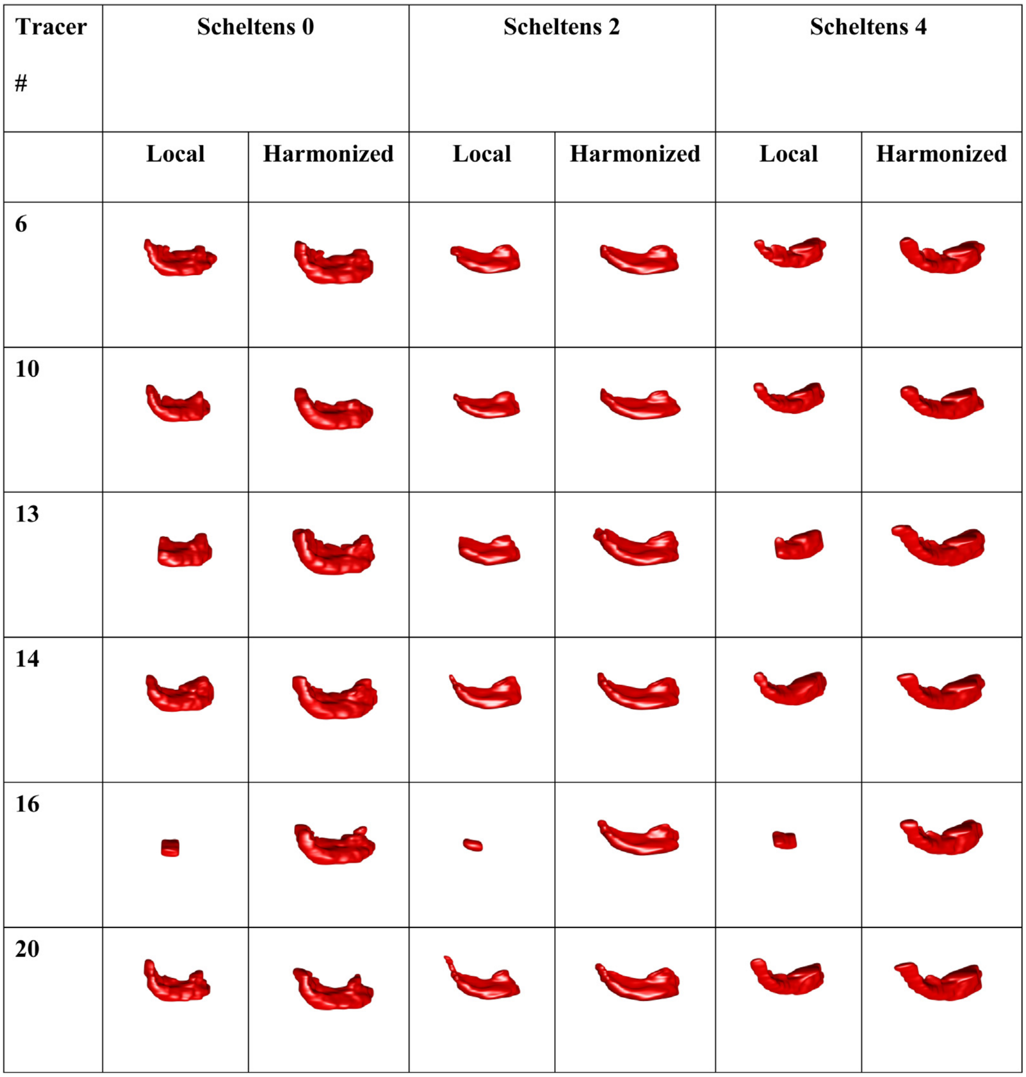


Fig. 4. Phase I: 3D rendering of three sample hippocampi with no, mild to moderate, and severe atrophy (Scheltens' atrophy scores of 0, 2, and 4) [16] segmented with local protocols requiring anterior-posterior commissure (AC-PC) image orientation and the HarP. Extreme variability can be appreciated when segmentations are performed following local protocols, which is greatly reduced when the HarP is used.

within and between human tracers and that HarP segmentations are of much higher agreement than those following local manual segmentation protocols. With the HarP, the Alzheimer's disease community has a largely agreed reference procedure for hippocampal volumetry. The HarP will allow the widespread use of hippocampal volumetry/measures for clinical diagnosis, disease tracking, and qualification of automated segmentation algorithms.

The stability results of the HarP outperformed that of local protocols. Although intra- and inter-rater reliability of published protocols is usually above 0.80, reliability measures are normally estimated with the consistency method, rather than with the more conservative absolute method, and collected from different tracers working in the same laboratory. To our knowledge, this is the first protocol reporting absolute reliability estimates, and measuring inter-rater stability between tracers working in different laboratories. Test-retest and inter-rater reliability of hippocampal manual segmentation protocols published to date have ranged between ICCs of 0.64 and 0.99 [9]. It should be noted, however, that in all cases reliability figures were obtained with the consistency ICC method, whereas the HarP achieved very high ICC values also when stability was assessed with the more conservative absolute method (see section *2.4 Statistical* analysis). In this work, we could estimate absolute inter-rater ICC values for tracers using the same local protocols in Phase I. We could perform this computation only for 6 tracers, of whom four used the same protocol [21] and came from different laboratories, and two used the same protocol [27] within the same laboratory. Absolute ICCs within each protocol were in the same range as HarP, and especially high for the two tracers coming from the same laboratory. We were not able to compute such values for all tracers because the other eight tracers used as many different local protocols. However,

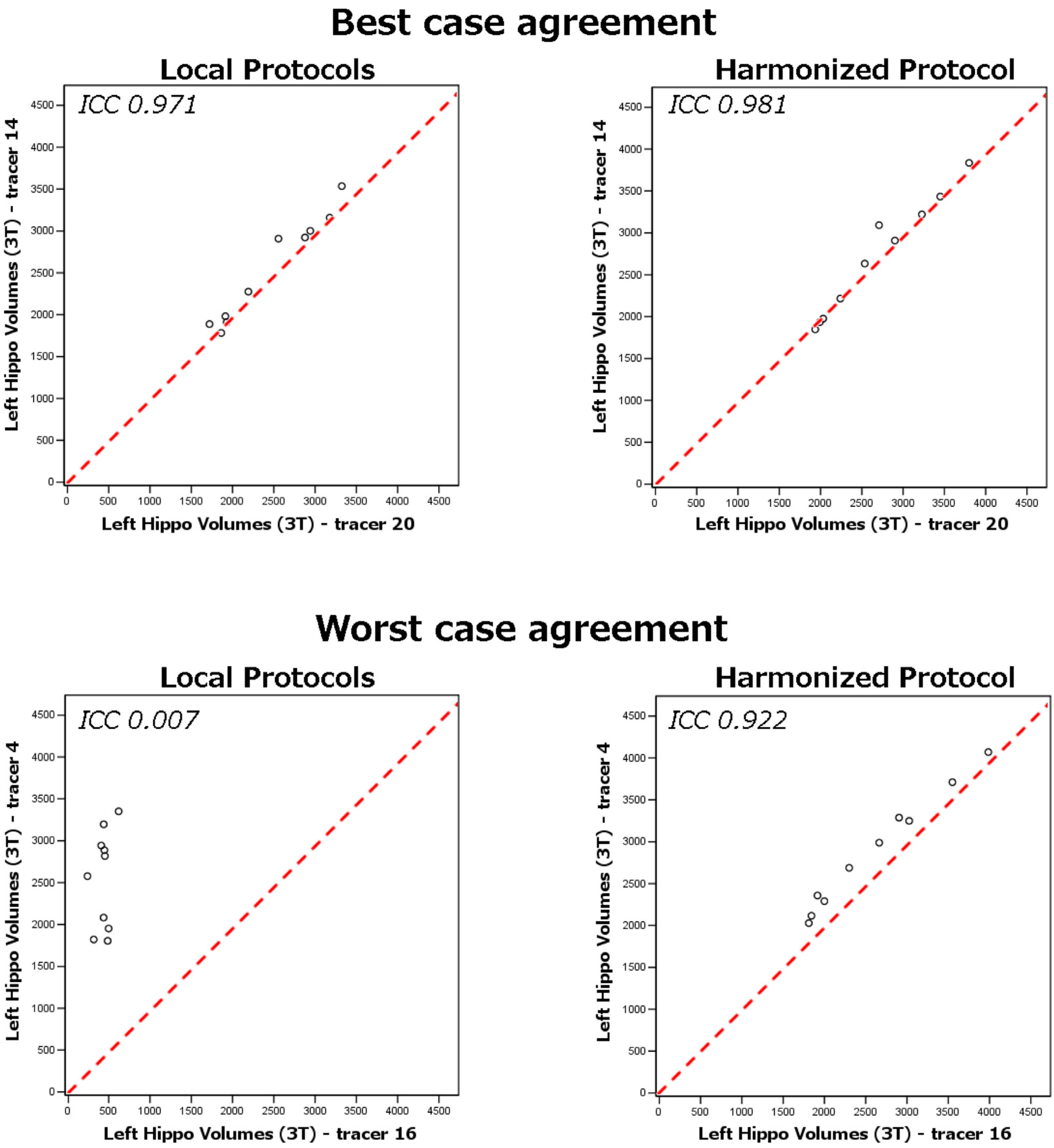


Fig. 5. Phase I: extreme case instances of the stability of the local and harmonized protocols among 14 naïve tracers. Graphs illustrate the best and poorest case agreement of pairs of tracers. In the case of best agreement between local protocols (left hippocampus between tracers #14 and #20), the absolute intra-rater ICC is only marginally lower than using the harmonized protocol. However, in the case of poorest agreement between local protocols (left hippocampus between tracers #4 and #16), the benefit of using the harmonized protocol is much higher. ICC, intraclass correlation coefficient.

considering also that individual tracers accuracy was very high with the newly learned HarP, these data suggest that the variability among "local" segmentations was due more to the difference in the adopted methods than to any heterogeneity in tracers accuracy. These data also suggest that tutorials are needed, to overcome the difficulty of remote learning for the HarP. We may expect that improvements in training conditions, and a longer experience in HarP segmentation may further improve agreement between remote tracers.

The stability results of the HarP also compare favorably with other AD biomarkers. The stability of measurement of CSF biomarkers (Aβ42 and tau) is currently the subject

Table 3
Phase II: stability of the EADC (European Alzheimer's Disease Consortium)-ADNI (Alzheimer's Disease Neuroimaging Initiative) Harmonized Protocol among five expert tracers. Inter-rater values were computed based on 15 + 1 (see Figure 1 and legend to Table 1) ADNI subjects traced at baseline. Intrarater values were computed based on retracings of the same subjects. Figures denote intraclass correlation coefficients and 95% confidence interval computed in 5-level and 2-level random effect absolute and consistency models

Interrater reliability among five tracers

	1.5T		3T	
Side	Consistency	Absolute	Consistency	Absolute
Left	0.958 (0.914–0.984)	0.887 (0.664–0.962)	0.957 (0.913–0.983)	0.896 (0.700–0.965)
Right	0.961 (0.920–0.985)	0.907 (0.728–0.969)	0.972 (0.942–0.989)	0.889 (0.653–0.968)

Test-retest reliability

		1.5T		3T	
Tracer#	Side	Consistency	Absolute	Consistency	Absolute
19	Left	0.984 (0.953–0.995)	0.984 (0.954–0.994)	0.993 (0.978–0.997)	0.993 (0.979–0.998)
	Right	0.987 (0.962–0.996)	0.988 (0.965–0.996)	0.991 (0.974–0.997)	0.992 (0.976–0.997)
16	Left	0.991 (0.974–0.997)	0.992 (0.976–0.997)	0.981 (0.944–0.994)	0.982 (0.948–0.994)
	Right	0.993 (0.979–0.998)	0.993 (0.979–0.998)	0.989 (0.967–0.996)	0.989 (0.969–0.996)
8	Left	0.976 (0.930–0.992)	0.963 (0.805–0.989)	0.994 (0.983–0.998)	0.988 (0.868–0.997)
	Right	0.987 (0.961–0.995)	0.975 (0.795–0.994)	0.986 (0.958–0.995)	0.980 (0.905–0.994)
18	Left	0.969 (0.911–0.990)	0.970 (0.916–0.990)	0.979 (0.940–0.993)	0.980 (0.943–0.993)
	Right	0.967 (0.904–0.989)	0.966 (0.905–0.988)	0.978 (0.936–0.993)	0.979 (0.940–0.993)
4	Left	0.952 (0.865–0.984)	0.937 (0.764–0.980)	0.958 (0.880–0.986)	0.955 (0.874–0.985)
	Right	0.957 (0.877–0.985)	0.947 (0.825–0.983)	0.977 (0.933–0.992)	0.965 (0.818–0.990)

of keen interest and international efforts. It has recently been estimated that the coefficient of variation of different batches of reagents or across different laboratories is between 13 and 36%, i.e. approximately one order of magnitude greater than that of the HarP [32]. The stability of plasma biomarkers of amyloidosis is even lower [33].

4.1. *What factors improve the measurement stability of the HarP?*

The landmark definition procedure of the HarP disambiguates structures in greater detail than before. Detailed instructions are provided for example to segment the transition tissue between the hippocampus and amygdala, resulting in lower heterogeneity between tracers. Specifically, tracers are guided in excluding the cortical and accessory basal nuclei of the amygdala, the entorhinal cortex, and in including the vertical digitation of the hippocampus. In all these regions discrepancies among tracers are frequently observed. By utilizing 3D visualization tools, the definition of the caudal hippocampal boundaries are specified as far as the *isthmus* and the *indusium griseum*, decreasing the wide variability due to the otherwise heterogeneous inclusion of gray matter in the large slices of the hippocampal tail. In the HarP, a 34 page-long user manual provides detailed instructions on the segmentation of individual structures on a 1 mm-by-1 mm slice basis (a summary schematic diagram is shown in Figure 2). An additional factor improving stability may consist in the AC-PC orientation of images. This may appear counterintuitive: manual hippocampal segmentations have traditionally been carried out on images orthogonal to the long hippocampal axis, because this orientation was considered to be associated with lesser partial volume effects.

Table 4
Phase II: sources of variability of hippocampal volumes with the EADC (European Alzheimer's Disease Consortium)-ADNI (Alzheimer's Disease Neuroimaging Initiative) Harmonized Protocol. An analysis of variance model shows that the variance due to different tracers is significantly lower than that due to the physiologic variability among individuals ("subject") and hippocampal atrophy at baseline; variance is lower, albeit nonsignificantly, than that due to field strength, hemispheric asymmetry (laterality), atrophy progression over 2 years (time point), and manufacturer

Anova model		Source	Mean	d.f.	Variance	Percentage of variance	CV (%)	*P* (versus tracer)
Effect	Pure random	Subject	3064	15	256,984	49.2	16.5	<.001
	Within	Tracer	2901	4	4899	0.9	2.4	–
		Field strength	2821	1	22,789	4.4	5.3	.90
		Laterality	2752	1	19,333	3.7	5.1	.88
		Atrophy rate	2850	2	17,719	3.4	4.7	.87
	Between	Atrophy	3367	4	174,746	33.5	12.4	<.001
		Manufacturer	2767	2	6316	1.2	2.9	.37
Residual error			–	870	19,257	3.7	–	.09

Abbreviation: CV, coefficient of variation.

However, further investigation into this issue, required for the consensus definition of the HarP, denoted that volume ICCs were non significantly higher for the hippocampi segmented on the AC-PC images, and that the overlapping values were significantly higher for segmentations on AC-PC images [34]. Thus, the AC-PC orientation, previously used by a minority of protocols, is associated with better agreement among tracers, possibly due to richer anatomical information allowing to discriminate the hippocampal head boundaries from amygdala in the axial plane [34].

4.2. Extremely good stability, but time consuming

The learning and qualification procedure to segment the hippocampus following the HarP is time consuming. The naïve tracers of Phase I of this study underwent three rounds of training including segmentation, feedback and correction, and a fourth round segmenting a sample of 20 hippocampi. As the segmentation of a single hippocampus takes about 40 minutes, the examination of the visual feedback 10 to 15 minutes and correction of segmentation for a single hippocampus between 10 and 30 minutes, the total time for learning and qualification can be estimated at 5 to 8 person-days. This is not a negligible effort, but one which we believe feasible in many situations. For instance, a clinical trial with 1000 cases scanned twice (total of 4000 hippocampi) would require 2.5 to 3.1 persons/year for complete segmentation. Feasibility in a clinical diagnostic setting is admittedly lower however, due to the current lack of reimbursement of most biomarkers in the diagnostic workup of dementia cases.

Of course, the cost of hippocampal volumetry quantification based on the HarP must be weighed against its diagnostic informative value. We believe that hippocampal volumetry is typically indicated in the etiologic diagnosis of MCI. Here, cognitive impairment may be due to AD or normal ageing with almost equal a priori probability, and finding hippocampal atrophy militates in favor of the former. On the contrary, in the large majority of dementia cases where typical AD is obvious based on history, medical and neurological examination, and neuropsychological testing, hippocampal volumetry may not add significant incremental diagnostic information.

4.3. Impact on the community

Perhaps the greatest value of the HarP will be in validation and qualification of automated segmentation algorithms which in turn will follow a single accepted standard within the field. A number of automated hippocampal segmentation algorithms have been developed to date [7], some of which very popular such as FreeSurfer. Virtually all have been validated against the "gold standard" of manual segmentation. However, the manual segmentation procedures used to date differ from each other, preventing reliable comparison of algorithms. The HarP provides algorithm developers with a single, internationally recognized true gold standard standard (all information, links and instructions for download and use are available at www.hippocampal-protocol.net). Indeed, a recent expansion of the original HarP project has produced an extended set of HarP-compliant hippocampal labels that will be used to train segmentation algorithms [35]; once trained, algorithms will be qualified by comparison to the "benchmark labels" segmented by master tracers [13,35].

Importantly, the HarP has been developed by and for the community of Alzheimer's scientists. Even so, we believe that the definition of hippocampus in the HarP is not disease-specific and it might be suited to study conditions unrelated to dementia, such as epilepsy and psychiatric disorders. In contrast to currently available protocols that exclude large parts of the hippocampal formation such as the head and/or tail [36], or that fail to separate the hippocampus from adjacent non-hippocampal structures [37,38], the HarP captures 100% of the hippocampus proper. Future developments such as segmentation of subfields with ultra-high field strength MR may break down the hippocampus into smaller structures with potentially greater disease-specific effects (http://www.hippocampalsubfields.com/).

4.4. Limitations

The HarP has been validated versus local protocols based on the availability of tracers and laboratories in the years 2012 to 2013. Although we were not able to include all protocols reported in the literature (over 70 according to [9]), we included 9 of the 12 most frequently used protocols reported in the AD literature [10].

The use of the same segmentation software (MultiTracer) as the only tool in the HarP project may have led tracers to adapt their protocols and habits to a different tool than they were used to, with possible lower segmentation performance. However, this difficulty should have affected in a similar way both local and HarP segmentations.

We selected scans to control the effect of relevant confounders, but others were not taken into account. Motion artifacts were not controlled for: case selection did not take image quality into account, such that the HarP was developed on scans representative of the general ADNI population. The high stability of the HarP despite lack of exclusion of motion artifacts suggests that tracers can easily account for motion artifacts, based on a priori knowledge of brain anatomy. Indeed, the availability of images with and with no motion artifacts in the HarP data set will allow to empirically test their effect on automated algorithms' accuracy.

Our effort to develop the HarP covered only one, albeit pivotal, step in the harmonization of hippocampal volumetry, i.e. the segmentation protocol. However, other actions involved in hippocampal volumetry need to be harmonized such as preprocessing (from normalization to image inhomogeneity correction); tools and settings used for segmentation; measurement of intracranial volume to correct for head size, and the specific statistical method by which head size adjustment of raw hippocampal volume is performed [2].

Although very high reliability values were observed across tracers, further analyses will need to clarify whether tracers are indeed segmenting the same voxels.

The hippocampal labels produced in this project will be used to qualify human tracers or algorithms. In this study, qualification of tracers was achieved based on the validity of their segmentations and showed extremely high test-retest and inter-rater reliability. However, thresholds will need to be set for the general qualification of human tracers or automated algorithms. This limitation is of particular relevance considering the pressure to qualify algorithms for hippocampal segmentation to be used in clinical trials.

Atrophy rate was taken into consideration in the error source estimate of this study, however additional validation may be required to accurately estimate the stability of the HarP in longitudinal scans. As well, a larger set of MR scans may have allowed to evaluate the effect of additional confounds, such as different receiver coils within manufacturer or magnetic field strengths. Similarly, inclusion of scans from a larger number of subjects would have increased the value of this study. Indeed, an expansion of the original design has allowed to segment with the HarP as many as 135 different ADNI AD cases and healthy controls [35], providing evidence of known group validity.

Despite the great detail of the HarP user manual, more user-friendly tutorials and interactive tools for training may be welcome, to shorten the time and effort in the learning phase. Finally, the HarP has been validated only in AD to date; additional studies are needed to ascertain its validity for diagnosis or tracking of other brain diseases.

Acknowledgments

The "Harmonized Protocol For Manual Hippocampal Segmentation: An EADC-ADNI Effort" has been funded thanks to an unrestricted grant from the Alzheimer's Association (grant number 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); University of California, Los Angeles (UCLA), CA (LG Apostolova); Laboratory of NeuroImaging (LoNI), 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); Northwestern University, IL (JG 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: 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). 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.

Preparatory work (reported in Boccardi et al., J Alzheimers Dis, 2011; 26:61-75 and in Boccardi et al, Alzheimer's and Dementia 2013, doi:pii: S1552-5260(13)00078-2) was funded by unrestricted grants from Lilly International and Wyeth International (a part of the Pfizer group).

The Statistical Working Group is co-funded through an international collaborative grant from the Ministère du Développement Économique, de l'Innovation et de l'Exportation of Quebec.

Josephine Barnes is an Alzheimer's Research UK senior research fellow based at the Dementia Research Centre, Department of Neurodegenerative Disease, UCL, Institute of Neurology. The Dementia Research Centre is an Alzheimer's Research UK Coordinating Centre and has also received equipment funded by Alzheimer's Research UK and Brain Research Trust. Part of this work was supported by the NIHR Queen Square Dementia BRU.

Contribution by Tim Swihart, Lisa C Silbert and Jeffrey Kaye was supported by NIH/NIA ADC P30 grant (P30 AG008017). Contribution by Stephane Lehéricy was supported by the grant: "Investissements d'avenir" (grant number ANR-10-IAIHU-06).

The authors of the 12 original segmentation protocols were: George Bartzokis, Mony deLeon, Leyla deToledo-Morrell, John G Csernansky, Clifford R Jack, Ronald J Killiany, Stephane Lehéricy, Nikolai Malykhin, Johannes Pantel, Jens C Pruessner, Hilkka Soininen, Craig Watson.

Delphi panelists were: Liana G Apostolova, Josephine Barnes, George Bartzokis, Charles DeCarli, Leyla deToledo-Morrell, Michael Firbank, Lotte Gerritsen, Wouter Henneman, Clifford R Jack, Ronald J Killiany, Nikolai Malykhin, Jens C Pruessner, Hilkka Soininen, Lei Wang, Craig Watson, Henrike Wolf.

Master tracers were: Liana G Apostolova, Martina Bocchetta, Rossana Ganzola, Gregory Preboske, Dominik Wolf.

Naïve tracers of Phase I were: Corinna Bauer, Claire Boutet, Emma Burton, Adam Christensen, Melanie Blair, Enrica Cavedo, Kristian S Frederiksen, Michel J Grothe, Sarah Hollander, Mariangela Lanfredi, Yawu Liu, Oliver Martinez, Masami Nishikawa, Marileen Portegies, Gregory Preboske, Margo Pronk, Travis Stoub, Tim Swihart, Mat Tinley, Felix van Dommelen, Chadwich Ward.

Naïve tracers of Phase II were: Corinna Bauer, Kristian S Frederiksen, Yawu Liu, Gregory Preboske, Tim Swihart.

Supervisors of naïve tracers were: Frederik Barkhof, George Bartzokis, Charles DeCarli, John C. Csernansky, Leyla deToledo-Morrell, Andreas Fellgiebel, Nick Fox, Giovanni B Frisoni, Mirjam Geerlings, Clifford R Jack, Jeffrey Kaye, Ronald J Killiany, Stephane Lehéricy, Hiroshi Matzuda, John O'Brien, Lisa Silbert, Philip Scheltens, Hilkka Soininen, Stefan Teipel, Gunhild Waldemar.

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

RESEARCH IN CONTEXT

1. Systematic review: Hippocampal volumetry is a useful biomarker for Alzheimer's disease (AD), but the heterogeneities among different segmentation protocols provide exceedingly different volume estimates. The definition of standard operating procedures for hippocampal volumetry is required for its concrete use as a biomarker. In this work, we have evaluated the reliability of the Harmonized Protocol for Hippocampal Volumetry, defined by a panel of international experts in the field of AD.
2. Interpretation: The protocol proved to be very reliable, and to provide hippocampal volume estimates that can be considered as standard measures, enabling the use of hippocampal volumetry as a proper biomarker for AD.
3. Future directions: This protocol will enable results from different studies to be compared or pooled and to provide standard hippocampal volumetry for the diagnosis of individual patients. This will boost pharmacological research and the everyday use of scientific knowledge.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jalz.2014.05.1756.

References

[1] Frisoni GB, Fox NC, Jack CR Jr, Scheltens P, Thompson PM. The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 2010; 6:67–77.

[2] 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;7:474–85.e4.

[3] Committee for Medicinal Products for Human Use (CHMP). Qualification opinion of low hippocampal volume (atrophy) by MRI for use in clinical trials for regulatory purpose - in pre-dementia stage of Alzheimer's disease. EMA/CHMP/SAWP/809208/2011. 17 November 2011.

[4] Jack CR Jr, Shiung MM, Gunter JL, O'Brien PC, Weigand SD, Knopman DS, et al. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. Neurology 2004; 62:591–600.

[5] Ridha BH, Barnes J, Bartlett JW, Godbolt A, Pepple T, Rossor MN, et al. Tracking atrophy progression in familial Alzheimer's disease: a serial MRI study. Lancet Neurol 2006;5:828–34.

[6] Chetelat G, Baron JC. Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage 2003;18:525–41.

[7] 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.

[8] Bosscher L, Scheltens P. MRI of the medial temporal lobe for the diagnosis of Alzheimer disease. In: Qizilbash N, Schneider LS, Chui H, Tarriot P, Brodaty H, Kaye J, et al., eds. Evidence-based dementia practice. Oxford, United Kingdom: Blackwell Science; 2002:154–62. p. II. 4.7.

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

[10] 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.

[11] Boccardi M, Bocchetta M, Ganzola R, Robitaille N, Redolfi A, Duchesne S, et al. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimers Dement 2015; 11:184–94.

[12] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi Definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimers Dement 2015;11:126–38.

[13] 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.

[14] Duchesne S, Valdivia F, Robitaille N, Mouiha A, Abiel Valdivia F, Bocchetta M, et al. Manual segmentation qualification platform for the EADC-ADNI harmonized protocol for hippocampal segmentation project. Alzheimers Dement 2015;11:161–74.

[15] Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, Vermersch P, et al. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatry 1992;55:967–72.

[16] Haller JW, Banerjee A, Christensen GE, Gado M, Joshi S, Miller MI, et al. Three-dimensional hippocampal MR morphometry with high-dimensional transformation of a neuroanatomic atlas. Radiology 1997;202:504–10.

[17] Killiany RJ, Moss MB, Albert MS, Sandor T, Tieman J, Jolesz F. Temporal lobe regions on magnetic resonance imaging identify patients with early Alzheimer's disease. Arch Neurol 1993;50:949–54.

[18] Chupin M, Mukuna-Bantumbakulu AR, Hasboun D, Bardinet E, Baillet S, Kinkingnéhun S, et al. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: method and validation on controls and patients with Alzheimer's disease. Neuroimage 2007;34:996–1019.

[19] Scahill RI, Frost C, Jenkins R, Whitwell JL, Rossor MN, Fox NC. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch Neurol 2003;60:989–94.

[20] Schott JM, Fox NC, Frost C, Scahill RI, Janssen JC, Chan D, et al. Assessing the onset of structural change in familial Alzheimer's disease. Ann Neurol 2003;53:181–8.

[21] Pruessner JC, Li LM, Serles W, Pruessner M, Collins DL, Kabani N, et al. Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories. Cereb Cortex 2000;10:433–42.

[22] van de Pol LA, van der Flier WM, Korf ES, Fox NC, Barkhof F, Scheltens P. Baseline predictors of rates of hippocampal atrophy in mild cognitive impairment. Neurology 2007;69:1491–7.

[23] Convit A, De Leon MJ, Tarshish C, De Santi S, Tsui W, Rusinek H, et al. Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. Neurobiol Aging 1997;18:131–8.

[24] Watson C, Jack CR Jr, Cendes F. Volumetric magnetic resonance imaging. Clinical applications and contributions to the understanding of temporal lobe epilepsy. Arch Neurol 1997;54:1521–31.

[25] Soininen HS, Partanen K, Pitkanen A, Vainio P, Hanninen T, Hallikainen M, et al. Volumetric MRI analysis of the amygdala and the hippocampus in subjects with age-associated memory impairment: correlation to visual and verbal memory. Neurology 1994;44:1660–8.

[26] Knoops AJ, Gerritsen L, van der Graaf Y, Mali WP, Geerlings MI. Loss of entorhinal cortex and hippocampal volumes compared to whole brain volume in normal aging: the SMART-Medea study. Psychiatry Res 2012;203:31–7.

[27] Jack CR Jr. MRI-based hippocampal volume measurements in epilepsy. Epilepsia 1994;35(Suppl 6):S21–9.

[28] deToledo-Morrell L, Stoub TR, Bulgakova M, Wilson RS, Bennett DA, Leurgans S, et al. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. Neurobiol Aging 2004;25:1197–203.

[29] 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.

[30] Bartzokis G, Altshuler LL, Greider T, Curran J, Keen B, Dixon WJ. Reliability of medial temporal lobe volume measurements using reformatted 3D images. Psychiatry Res 1998;82:11–24.

[31] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

[32] 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;7:386–95.e6.

[33] 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.

[34] 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.

[35] 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.

[36] Bremner JD, Randall P, Scott TM, Bronen RA, Seibyl JP, Southwick SM, et al. MRI-based measurement of hippocampal volume in patients with combat-related posttraumatic stress disorder. Am J Psychiatry 1995;152:973–81.

[37] Noga JT, Vladar K, Torrey EF. A volumetric magnetic resonance imaging study of monozygotic twins discordant for bipolar disorder. Psychiatry Res 2001;106:25–34.

[38] Cook MJ, Fish DR, Shorvon SD, Straughan K, Stevens JM. Hippocampal volumetric and morphometric studies in frontal and temporal lobe epilepsy. Brain 1992;115:1001–15.