

EADC-ADNI Benchmark Labels for Harmonized Hippocampal Segmentation



Marina Boccardi¹, Martina Bocchetta^{1,2}, Liana G. Apostolova³, Rossana Ganzola⁴, Gregory Preboske⁵, Dominik Wolf⁶, Patrizio Pasqualetti², Nicolas Robitaille⁴, Simon Duchesne⁴, Clifford R. Jack Jr⁵, Giovanni B. Frisoni¹, for the EADC-ADNI Working Group on The Harmonized Protocol for Hippocampal Volumetry and the Alzheimer's Disease Neuroimaging Initiative

From: ¹LENITEM (Laboratory of Epidemiology, Neuroimaging and Telemedicine) IRCCS – S. Giovanni di Dio – Fatebenefratelli Brescia, Italy (MB, MBocch, GBF); ²AFaR (Associazione Fatebenefratelli per la Ricerca), S. Giovanni Calibita -Fatebenefratelli, Rome, Italy (MB,PP); ³Mary S. Easton Center for Alzheimer's Disease Research and Laboratory of Neuroimaging, David Geffen School of Medicine, University of California, Los Angeles (LA); ⁴Department of Radiology, Université Laval and Centre de Recherche de l'Institut universitaire de santé mentale de Québec, Québec City, Canada (RG, NR, SD); ⁵Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN (GP, CRJ); ⁶Klinik für Psychiatrie und Psychotherapie, Johannes Gutenberg-Universität Mainz, Germany (DW).

Background The use hippocampal volumetry as a biomarker of Alzheimer's Disease (AD) requires standard operating procedures. A Delphi panel of experts converged on a Harmonized Protocol (HP) for manual segmentation from magnetic resonance images (MRIs).

Objective To produce benchmark images of hippocampal segmentation reflecting the HP, to be released publicly as the gold standard for human tracers and algorithms.

Results Two rounds of corrections were asked to tracers. One round of corrections was applied to the HP. Reliability values for the corrected segmentations were: lowest absolute 5-level intra-rater, ICC 0.943 (95% CI 0.335-0.989); lowest inter-rater: ICC 0.943 (0.791-0.986) (Table 2). The mean 5-level inter-rater values were ICC 0.96 (absolute) and ICC 0.98 (consistency). Overlapping reliability among the 5 tracers was 0.73 for 1.5T and 0.75 for 3T images (Figures 1 and 2).

Methods The consensual HP criteria were written in a document, open to corrections based on the feedback received during the project. Based on the HP, 5 expert tracers from independent centres segmented a sample of 40 hippocampi from 10 ADNI (Alzheimer's Disease Neuroimaging Initiative) subjects scanned at both 1.5T and 3T (Table 1). Segmentations were examined slice-by-slice; volume reliability was computed through absolute and consistency 5-level intraclass correlation coefficients (ICC). Dice similarity coefficients were computed based on a formula adapted for 5 raters (5 x intersection of the 5 segmentations / sum of the 5 absolute volumes). Tracers were asked to correct their segmentation when it diverged from the HP. Whenever the tracers' mistakes could be attributed to ambiguities in the HP written description, this was edited, resent to panelists for checking the adherence to the Delphi decisions, and to tracers for improving segmentation.

Table 1: Features of ADNI subjects selected for benchmark labelling

	MTA scale					p-value
	0	1	2	3	4	
Age, years	71 (2.8)	80 (7.8)	75 (2.8)	82 (2.1)	80 (4.2)	0.219
Gender, female	2 (100%)	1 (50%)	0 (0%)	0 (0%)	1 (50%)	0.212
Education, years	17 (1.4)	14 (5.7)	15 (1.4)	19 (1.4)	16 (5.7)	0.567
ApoE ε4 allele, carriers	0 (0%)	0 (0%)	1 (50%)	1 (50%)	1 (50%)	0.582
Diagnosis, CTR/MCIs/MCI/AD	1/1/0/0	2/0/0/0	1/0/1/0	0/1/1/0	0/0/1/1	0.406
CSF Aβ ₁₋₄₂ levels, pg/ml	111 (0)	278 (0)	222 (0)	132 (7.8)	160 (42.4)	0.273
1.5T Scanner Manufacturer, Philips/GE/Siemens	0/1/1	0/2/0	1/1/0	0/1/1	0/2/0	0.448
3T Scanner Manufacturer, Philips/GE/Siemens	0/1/1	0/2/0	1/1/0	1/0/1	0/1/1	0.537

Table 2: Reliability of the expert tracers computed on benchmark labels

	Left Hippocampus	Right Hippocampus
Intra-rater 1.5T vs 3T (n=10)		
MB	0.981 (0.928-0.995)	0.986 (0.776-0.997)
RG	0.968 (0.879-0.992)	0.974 (0.902-0.994)
GP	0.943 (0.335-0.989)	0.968 (0.541-0.994)
LA	0.966 (0.819-0.992)	0.971 (0.818-0.993)
DW	0.981 (0.930-0.995)	0.986 (0.944-0.997)
Inter-rater (n=10)		
1.5T	0.957 (0.881-0.988)	0.971 (0.916-0.992)
Inter-rater (n=10)		
3T	0.943 (0.791-0.986)	0.962 (0.863-0.990)

Figure 1: Box-plots of similarity coefficients denoting spatial overlap of segmentations among the 5 expert tracers for the benchmark labels

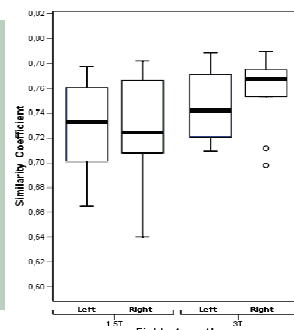
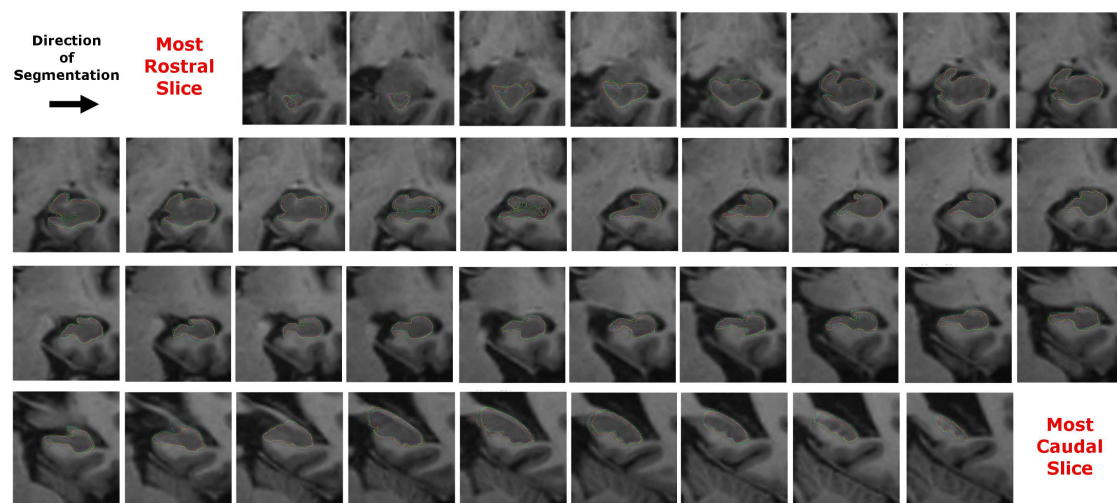


Figure 2: Segmentations of the 5 expert tracers on a sample benchmark image



Conclusions The HP showed to produce very reliable manual segmentations. The obtained hippocampal segmentations appear as an appropriate benchmark set for certification of tracers who will carry out the validation of the HP.

References Web-site: www.hippocampal-protocol.net mail: hippocampal.protocol@gmail.com
 Boccardi et al; J Alzheimers Dis. 2011;26 Suppl3:61-75.
 Boccardi et al., Alzheimer's & Dementia, 2013, in press