MANUAL SEGMENTATION QUALIFICATION PLATFORM FOR THE EADC-ADNI HARMONIZED PROTOCOL FOR HIPPOCAMPAL SEGMENTATION PROJECT

Simon Duchesne[1], Fernando Valdivia[2], Nicolas Robitaille[2], Abderazzak Mouiha[2], Abiel F. Valdivia[2], Martina Bocchetta[3], Liana G. Apostolova[4], Rossana Ganzola[2], Greg Preboske[5], Dominik Wolf[6], Marina Boccardi[3], Clifford R. Jack Jr. [5], and Giovanni B. Frisoni[3]

[1]Universite Laval, Quebec, Quebec, Canada, [2]Universite Laval, Quebec, Canada, [3]IRCCS Fatebenefratelli, Brescia, Italy, [4]UCLA, Los Angeles, CA, United States, [5]Mayo Clinic, Rochester, Minnesota, United States, [6]Johannes Gutenberg-Universität, Mainz, Germany

TARGET AUDIENCE - Methodological and clinical researchers alike interested in MRI-based hippocampal volumetry as a neurodegeneration biomarker.

PURPOSE - 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]. Measuring the hippocampus reliably requires, on the one hand, high-contrast images of the human brain, such as those obtained via T1-weighted magnetic resonance images (**MRI**); and on the other, a sound neuroanatomical protocol for manual delineation of the structure on MRI. Heterogeneity in anatomic definitions and segmentation guidelines have hampered comparisons among different studies using hippocampal volumetry for diagnosis or as a surrogate marker for disease progression. An effort has been undertaken by European Alzheimer's Disease Consortium (**EADC**) and Alzheimer's Disease Neuroimaging Initiative (**ADNI**) centers to develop a harmonized protocol for the manual segmentation of the hippocampus on MR scans [2]. This project conducted evidence-based Delphi panels to facilitate a consensual definition of a Harmonized Protocol (**HarP**) [3], using which a small group of "master tracers" segmented a set of *benchmark images*[4]. This article describes our work towards 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. Our first objective was to demonstrate that the training process embedded in the platform led to improved performance (i.e. increased compliance) with the HarP.

METHODS - *Subjects/Labels* - For this study, we selected 10 subjects from the ADNI database according to visual atrophy ratings of the medial temporal lobe [5], in order 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 [6]. 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 [4]. 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 1.5T and 3T MRIs. *New tracers* - For the purposes of this study, new tracers were required to segment the same set of 20 ADNI images following the same settings/procedures as Master Tracers [4]. Specifically, 10 images were assigned to a "training" set (for a total of 20 hippocampi), and the remaining assigned to the "qualification" set. *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. To validate segmentation accuracy, we measured the following elements: (A) **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 to the average masters' volume for that hippocampus on a pairwise basis; (B) **Spatial overlap:** while segmentations may have similar volumes, in order to be accurate they must significantly overlap. To capture this variability, we calculated the Jaccard similarity index as a metric of spatial overlap. (C) **Spatial distance:** to ensure further compliance with the definitions set forth in the HarP, we required a distance metric to assess whether or not the new tracers more or less espoused the same contour than 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 values are bound between [0, 1]. 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 equal to the distance from the minimum or maximum contour. The final statistic 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. *Statistical analysis* - Our objective was to assess the increased compliance of tracers that had gone through the training phases. We segregated the training set in three phases, whereby users segmented two images (four hippocampi) in Phase I, six images (twelve 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. To test the increase in compliance between phases, we performed a repeated measures analysis of the Jaccard overlap statistic, averaged over all tracers that completed Phases II and III, and tracers that completed Phases I, II and III.

RESULTS - The Qualification Platform came online on 5 Oct 2012. In this project, 16 users registered on the platform, and 13 completed all three steps of the training. 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 ($p < 0.0001$) (i.e. Jaccard overlap increased significantly between phases for all images, on average for all raters), as well as a significant effect for SIDE ($p < 0.001$) for all variables except one (i.e., there was a difference between performance between the left and right hippocampi). Testing for those raters that performed all three phases for those selected images which were present in each phase again showed a significant effect for Jaccard overlap ($p < 0.0001$), but SIDE fell below significance ($p > 0.05$).

DISCUSSION - Statistical analysis has shown that the effect of training positively increased the compliance with the HarP and therefore served to reduce between-rater variance. It is therefore recommended to maintain all three phases of training to increase the rater's chance of complying with the HarP. A thorough statistical validation must be performed to determine metric qualification thresholds for new users, based on the results from the current group of tracers. This will determine the specificity and sensitivity of the current metrics. The platform as presented is geared towards measuring compliance of manual raters with the HarP; it does not provide for the training and testing of results from automated algorithms. For that purpose, the training set will need to be substantially expanded; and the testing of voxellized labels will require adaptation of the current platform metrics, from 2D to 3D and contours to objects.

CONCLUSION - We have developed a prototype web-based Qualification Platform for training of new tracers on the HarP for the segmentation of the hippocampus on MRI, including automated feedback and qualification features. This on-line system is available at www.hippocampal-protocol.net.

REFERENCES - 1.McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging and the Alzheimer's Association workgroup. Alzheimer's & dementia : the journal of the Alzheimer's Association 2011. **-** 2.Frisoni GB, Jack CR. Harmonization of magnetic resonance-based manual hippocampal segmentation: A mandatory step for wide clinical use. Alzheimer's & dementia : the journal of the Alzheimer's Association 2011;7:171-174. - 3.Boccardi M, Bocchetta M, Apostolova L, et al. Delphi consensus on landmarks for the manual segmentation of the hippocampus on MRI: preliminary results from the EADC-ADNI Harmonized Protocol Working Group. Neurology 2012;78:003. - 4.Bocchetta M, Boccardi M, Ganzola R, et al. Harmonized benchmark labels of the hippocampus on MR: the EADC-ADNI project. Alzheimer's & Dementia: The Journal of the Alzheimer's Association 2013. - 5.Scheltens P, Leys D, Barkhof F, 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;55:967-972. **-** 6.Boccardi M, Bocchetta M, Ganzola R, et al. Operationalizing protocol differences for EADC-ADNI manual hippocampal segmentation. Alzheimers Dement 2013.