Coupé, P.; Eskildsen, SF.; Manjón Herrera, JV.; Fonov, VS.; Collins, DL.; Alzheimer's Dis Neuroimaging (2012). Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's Disease. NeuroImage. 59(4):3736-3747. doi:10.1016/j.neuroimage.2011.10.080

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI).

Simultaneous Segmentation and Grading of Anatomical Structures for Patient's Classification: Application to Alzheimer's Disease

Pierrick Coupé[1], Simon F. Eskildsen[1], José V. Manjón[2], Vladimir S. Fonov[1], D. Louis Collins[1] and the Alzheimer's Disease Neuroimaging Initiative[*]

* Corresponding author: pierrick.coupe@gmail.com

[1] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada University, 3801 University Street, Montreal, Canada H3A 2B4
[2] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

Abstract

In this paper, we propose an innovative approach to robustly and accurately detect Alzheimer's disease (AD) based on the distinction of specific atrophic patterns of anatomical structures such as hippocampus (HC) and entorhinal cortex (EC). The proposed method simultaneously performs segmentation and grading of structures to efficiently capture the anatomical alterations caused by AD. Known as SNIPE (Scoring by Non-local Image Patch Estimator), the novel proposed grading measure is based on a nonlocal patch-based framework and estimates the similarity of the patch surrounding the voxel under study with all the patches present in different training populations. In this study, the training library was composed of two populations: 50 cognitively normal subjects (CN) and 50 patients with AD, randomly selected from the ADNI database. During our experiments, the classification accuracy of patients (CN versus AD) using several biomarkers was compared: HC and EC volumes, the grade of these structures and finally the combination of their volume and their grade. Tests were completed in a leave-one-out framework using discriminant analysis. First, we showed that biomarkers based on HC provide better classification accuracy than biomarkers based on EC. Second, we demonstrated that structure grading is a more powerful measure than structure volume to distinguish both populations with a classification accuracy of 90%. Finally, by adding the ages of subjects in order to better separate age-related structural changes from disease-related anatomical alterations, SNIPE obtained a classification accuracy of 93%.

Keywords: hippocampus, hippocampus volume, hippocampus grading, patient's classification, nonlocal means estimator, Alzheimer's disease, entorhinal cortex.

1. Introduction

The atrophy of medial temporal lobe structures, such as the hippocampus (HC) and entorhinal cortex (EC), is potentially specific and may serve as an early biomarker of Alzheimer's disease

[*] Data used in the 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 contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at www.loni.ucla.edu/ADNI/Collaboration/ADNI Author ship list.pdf).

(AD) (Frisoni et al., 2010). In particular, atrophy of the HC and the EC can be used as a marker of AD progression since changes in these structures are closely related to changes in cognitive performance of the subject (Frisoni et al., 2010). The evaluation of structure atrophy is usually estimated by volumetric studies on anatomical MRI, requiring a segmentation step that can be very time consuming when done manually. This limitation can be overcome by using automatic segmentation methods.

In recent years, numerous methods have been proposed to automatically segment the hippocampus (Barnes et al., 2008; Bishop et al., 2011; Chupin et al., 2009b; Collins and Pruessner, 2010; Coupe et al., 2011b; Gousias et al., 2008; Khan et al., 2011; Lotjonen et al., 2010; Morey et al., 2009; Pohl et al., 2007; van der Lijn et al., 2008; Wang et al., 2011). Among these methods, several have been used to classify AD patients using HC volume (Chupin et al., 2009a; Colliot et al., 2008; Morra et al., 2010; Mueller et al., 2010). Despite the high segmentation accuracy of the new HC segmentation approaches, using the HC volume enables a separation between AD and cognitively normal (CN) subjects with a success rate only around 72-74% over the entire Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Cuingnet et al., 2010). This limited capability to classify AD patients using the HC volume only, may be due to a simplification of the complex atrophy patterns to a volume - a simple scalar. Recently, several shape analysis methods have been proposed (Csernansky et al., 2005; Gerardin et al., 2009; Gutman et al., 2009) to capture detailed HC structural modifications in order to obtain a more accurate classification. At 77% in the comparison proposed by Cuingnet *et al.* (2010), the approach proposed in Gerardin et al. (2009) yields slightly better classification than a volumetric approach. Therefore, development of new methods capable of estimating subtle anatomical modifications of HC appears to be a critical point to obtain better classification rate. Longitudinal approaches to the AD classification problem have also been investigated by estimating the HC atrophy rate over time (Henneman et al., 2009; Wolz et al., 2010). In Wolz et al. (2010), the authors reported a correct classification rate of 82% on 568 images of the ADNI dataset. However, this type of approach requires several time-points for a given patient. Finally, an emerging method is to segment subfields of the hippocampus (Van Leemput et al., 2009; Yushkevich et al., 2010). This approach seems promising since it is potentially able to detect more detailed atrophic patterns. However, ultra-high resolution MRI is required, which is not yet the standard in clinical practice and thus limits the practical applicability of this approach for the moment.

The EC volume has also been investigated as a possible biomarker to detect AD (Devanand et al., 2007; Du et al., 2001; Frisoni et al., 1999; Juottonen et al., 1999; Xu et al., 2000). EC atrophy seems to appear slightly earlier in AD progression than HC atrophy, and thus could be used as a more specific biomarker in the initial stages of the disease (Frisoni et al., 2010). However, the high inter-subject variability of the EC and the difficulty to define EC boundary in anatomical MRI make volumetric studies on EC very challenging (Juottonen et al., 1999; Kenny et al., 2008; Xu et al., 2000). Therefore, studies based on EC volume have been limited to comparison of manual segmentations. Patient classification accuracy using EC volume greatly varies according to the dataset, from 67% (Frisoni et al., 1999) up to 87% (Juottonen et al., 1999). Depending on the study, EC volume can be more sensitive than HC volume to separate CN versus AD (Juottonen et al., 1999), less sensitive (Frisoni et al., 1999) or similarly sensitive (Devanand et al., 2007; Xu et al., 2000). As noticed in Xu et al. (2000), it seems that the theoretical advantage of EC measurements over HC measurements is adversely impacted by the difficulty to segment EC due to the ambiguity in defining its boundary in MRI. The development of automatic methods to segment EC is challenging. However, an accurate and consistent EC segmentation method could have an important

impact on the use of this structure on large datasets and in a more systematic manner within the study of AD.

In this paper, we propose a new approach designed to address the two problems described earlier: *i)* to obtain a more detailed detection of structural changes caused by the disease and *ii)* to perform the automatic segmentation of complex structures such as EC. Inspired by work in image denoising (Buades et al., 2005; Coupe et al., 2008), we have recently proposed a new nonlocal patch-based label fusion method to segment anatomical structures (Coupe et al., 2011b). By taking advantage of pattern redundancy present within the subject's image, as well as the redundancy across training subjects, the nonlocal means scheme enables robust use of a large number of samples during estimation. In Coupé et al. (2011b), we applied this approach to label fusion for the segmentation of anatomical structures such as HC of healthy subjects and lateral ventricles of patients with AD. In this paper, we propose an extension of this patch-based segmentation method in order to evaluate the similarity (in the nonlocal means sense) of the intensity content of one MRI compared to several training populations. By using training populations with different clinical status (e.g., healthy CN subjects and patients with AD), a nonlocal means estimator is used to evaluate the proximity (i.e., the grade of the disease or the degree of anatomical change consistent with disease in the case of AD) of each voxel of the MRI under study compared to the training populations (see Fig. 1). Since the grade estimation and the label fusion steps require the same patch comparison step, simultaneous segmentation and grading of the studied structure can be achieved in one pass without extra computation. In SNIPE, the nonlocal patch-based comparison is used to *i)* efficiently fuse the labels of MRI in a training database in order to segment EC and HC, and simultaneously to *ii)* aggregate the clinical status of the populations constituting the training database in order to detect the presence (or not) of the disease. Finally, the average grading value obtained over the segmented structures is proposed as a new biomarker to estimate the clinical status of the subject under study as a potential computerized aid to diagnosis. The contributions of the paper are: *i)* the introduction of an innovative approach to better characterize the patterns of structural modification caused by the disease (e.g., anatomical changes such as atrophy in case of AD) through the new concept of grading, *ii)* the presentation of a method to automatically and simultaneously perform the segmentation and the grading of EC and HC, and *iii)* the demonstration that the proposed approach can be used as a novel biomarker to efficiently achieve patient classification in the context of AD.

2. Materials and Methods

2.1 Method overview

- *The nonlocal means estimator:*

The nonlocal means filter was first introduced by Buades *et al.* (2005) for the purpose of image denoising. In nonlocal means-based approaches (Buades et al., 2005; Coupe et al., 2008), the patch $P(x_i)$ surrounding the voxel x_i under study is compared with all the patches $P(x_j)$ of the image Ω (or a subpart of the image) whatever their spatial distance to $P(x_i)$ (i.e., this is the meaning of the term "nonlocal"). According to the patch similarity between $P(x_i)$ and $P(x_j)$, estimated by the sum of squared differences (SSD) measure, each patch receives a weight $w(x_i, x_j)$:

$$w(x_i, x_j) = e^{\frac{-\left\| P(x_i) - P(x_j) \right\|_2^2}{h^2}} \quad (1)$$

where $\|.\|_2$ is the L2-norm computed between each intensity of the elements of the patches $P(x_i)$ and $P(x_{s,j})$, and h^2 is the smoothing parameter of the weighting function. This weighting function is designed to give a weight close to 1 when the SSD is close to zero and a weight close to zero with the SSD is high. Finally, all the intensities $u(x_j)$ of the central voxels of the patches $P(x_j)$ are aggregated through a weighted average using the weights $w(x_i, x_j)$. In this way, the denoised intensity $\hat{u}(x_i)$ of the voxel x_i can be efficiently estimated:

$$\hat{u}(x_i) = \frac{\sum_{j \in \Omega} w(x_i, x_j) u(x_j)}{\sum_{j \in \Omega} w(x_i, x_j)} \tag{2}$$

Despite its simplicity, the nonlocal means filter has been demonstrated to have excellent denoising performance. This filter is currently one of the most studied denoising filters and many improvements have been proposed since its introduction (see Buades et al. (2010) for a review of these improvements). The efficiency of the nonlocal means filter relies on two intuitive aspects: the pattern redundancy present in an image (i.e., its self-similarity) and the robust detection of samples derived from the same population by using local context (i.e., patch-based comparison):

- First, to improve the accuracy of an estimator, it is possible to reduce the committed error by increasing the number of involved samples. By using an infinite number of samples derived from the same population, the error theoretically converges to zero. To drastically increase the number of samples used, the nonlocal means filter takes advantage of the redundancy of information by using all the similar voxels present over the entire image.

- Second, to ensure that the used samples are derived from the same population, the surrounding neighbor of a voxel can be used to robustly detect similar realizations of the same process. In the nonlocal means approach, this task is achieved by patch-based comparison using SSD. Two voxels with similar surrounding patches are considered as similar and to belong to the same population. More precisely, the nonlocal means filter performs patch comparison to estimate the degree of the similarity between two voxels. This way, each involved sample has a weight (see Eq. 1) reflecting its relevance.

Finally, a simple weighted average (see Eq. 2) is used to aggregate the samples according to their relevance. This way, the resulting estimator embodies the two interesting qualities described above: to build on a large number of samples and to ensure that the involved samples are derived from the same population.

- *From denoising to segmentation:*

In Coupé et al. (2010, 2011b), we were the first to introduce the nonlocal means estimator in the context of segmentation by averaging labels instead of intensities. By using a training library of N subjects, whose segmentations of structures are known, the weighted label fusion is estimated as follows:

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j}).l(x_{s,j})}{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j})} \tag{3}$$

where $l(x_{s,j})$ is the label (i.e., 0 for background and 1 for structure) given by the expert to the voxel $x_{s,j}$ at location j in training subject s. It has been shown that the nonlocal means estimator $v(x_i)$ provides a robust estimation of the expected label at x_i. With a label set of {0,1} voxels with value $v(x_i) \geq 0.5$ are considered as belonging to the considered structure and the remaining voxels as background.

In Coupé et al. (2010, 2011b), we showed that accurate segmentations of anatomical structures can be obtained using this simple patch-based label fusion framework. In addition, to take advantage of the self-similarity of the image as done for denoising, the nonlocal label fusion also relies on inter-subject anatomical consistency. Therefore, many similar patches (self-similarity) can be found in every training subject (inter-subject consistency), thus improving the final estimation. Finally, compared to atlas-based methods using nonlinear registration, the nonlocal patch-based approach has the advantage of better handling the inter-subject variability problem. Contrary to the one-to-one correspondence assumed by nonlinear warping methods between the source and the target image, the nonlocal means estimator makes it possible to deal with one-to-many mappings, which better captures the link between subjects' anatomies. This interesting aspect of the nonlocal means estimator has been used to improve video super-resolution without explicit estimation of inter-frame motion (Protter et al., 2009; Takeda et al., 2009).

- *From segmentation to grading:*

In this paper, we propose to extend this segmentation method to efficiently aggregate clinical status (CN or AD) in order to estimate the proximity (in the nonlocal means sense) of each voxel compared to both populations constituting the training library (see Fig. 1). To achieve this goal, we introduce the new concept of patch-based grading that reflects the similarity of the patch surrounding the voxel under study with all the patches present in the different training populations. In this way, the neighborhood information is used to robustly drive the search of anatomical patterns that are specific to a given subset of the training library. When the training populations include data from subsets of subjects in different clinical states, this approach provides an estimation of the grade (i.e., degree of closeness to one group or another) for each voxel:

$$g(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j}).p_s}{\sum_{s=1}^{N} \sum_{j \in \Omega} w(x_i, x_{s,j})} \quad (4)$$

where p_s is the clinical status of the training subject s. In our case, $p_s=-1$ was used for AD status and $p_s=1$ for CN status. A negative grading value (respectively, a positive grading value) $g(x_i)$ indicates that the neighborhood surrounding x_i is more characteristic of AD than CN (respectively, of CN than AD) (see Fig. 2). The absolute value $|g(x_i)|$ provides the confidence given to the grade estimation. When $|g(x_i)|$ is close to zero, the method indicates that the patch under study is similarly present in both populations and thus is not specific to one of the compared populations and provides little discriminatory information. When $|g(x_i)|$ is close to 1, the method detects a high proximity of the patch under study with the patches present in one of the training populations and not in the other. Finally, for each subject, an average grading value is computed over all voxels in the estimated structure segmentation (i.e., for all x_i with $v(x_i) \geq 0.5$) for each side (e.g., $\bar{g}_{HC-left}$ or $\bar{g}_{EC-right}$). Since the grading and the segmentation involve the same patch comparison step, these structures are extracted at the same time that their grade is estimated (see Fig. 2).

Several strategies can be used to fuse the average grading of the studied structures. First, each side of the structure can be used separately. Second, it is possible to assign the same weight to the left and right HC and EC (e.g., $\bar{g}_{HC} = (\bar{g}_{HC-left} + \bar{g}_{HC-right})/2$). This strategy of fusing both sides to be more robust to segmentation inaccuracy was used by Chupin et al. in a volumetric study *(Chupin et al., 2009a)*. During our experiments, we found that these two strategies provided similar results for HC and EC. However, for the HC-EC complex, the best strategy was to compute left and right average grading values over HC-EC segmentation (what gives more importance to HC) and then to use the mean of both sides ($\bar{g}_{HCEC} = (\bar{g}_{HCEC-left} + \bar{g}_{HCEC-right})/2$). Therefore, we decided to present all the results using the second strategy.

2.2 Training library construction

- *Datasets*

In this study, the ADNI database (www.loni.ucla.edu/ADNI) was used to validate the proposed approach. This database contains both 1.5T and 3.0T T1-w MRI scans. For our experiments, we randomly selected 120 MRI scans, 60 1.5T MRI baseline scans of CN subjects and 60 1.5T MRI baseline scans of patients with AD.

- *Preprocessing*

All the selected images were preprocessed as follows: 1) correction of inhomogeneities using N3 (Sled et al., 1998), 2) registration to the stereotaxic space using a linear transform to the ICBM152 template (1x1x1 mm^3 voxel size) (Collins et al., 1994) and 3) cross-normalization of the MRI intensity using the method proposed in Nyul and Udupa (2000). After preprocessing, all the MRIs are coarsely aligned (linear registration), tissue intensities are homogeneous within each MRI volume (inhomogeneity correction) and across the training database (intensity normalization) (see Fig. 1).

- *Label propagation*

From the 120 processed MRI scans, 20 scans (10 CN and 10 AD) were randomly selected to be used as **seed dataset** for segmentation. The HC and the EC of this **seed dataset** were manually segmented by following the protocol defined in (Pruessner et al., 2002). The manual segmentations of the **seed dataset** were then propagated to the 100 remaining scans constituting our **test dataset** using the method described in (Coupe et al., 2011b). After the segmentation propagation step, the **test dataset** was composed of 100 MRI (50 CN subjects and 50 patients with AD) with their corresponding automatic segmentations (see Fig. 1). In our **test dataset**, the average age of the populations is 74.8 (±4.8) for CN and 74.9 (±6.4) for AD. The age for the two populations is not significantly different (p=0.36, unpaired t-test). In addition, the Mini Mental State Evaluation (MMSE) is 29.1 (±1.2) for CN and 23.2 (±2.0) for AD.

2.3 Implementation details

In all experiments described here, the optimal parameters empirically found in Coupé et al. (2011b) for HC segmentation have been used and thus the patch size was fixed to 7x7x7 voxels and the pre-selection threshold set to *th=0.95*.

As done in Coupé et al. (2011b), Ω was replaced by a cubic volume V_i centered on x_i . First, this strategy to use a semi-local paradigm instead of a fully nonlocal paradigm makes the processing computationally practical. In the denoising literature, this approach is used in the majority of the papers and has been shown to produce near-optimal or optimal results except for images with repetitive textures (Brox et al., 2008). Second, as shown in Coupé et al. (2011b), in the case of HC segmentation, limitation of the search window provides better results (see left of Fig. 8 in Coupé et al. (2011b)). Since all the images are linearly registered, the patches belonging to HC are located within a restricted area. By using a larger search window, outliers are added that marginally degrade the segmentation and uselessly increases the computational time. While in Coupé et al. (2011b) the search window size was fixed, we used a locally adaptive search window size. The initialization of the search window was set to 9x9x9 voxels as suggested in Coupé et al. (2011b). However, *in the case when* no similar patches can be found in this search window (i.e., none of the patches pass through the pre-selection), its radius is increased by one voxel until at least one similar patch in each population is found (i.e., at least one patch in each population pass through the pre-selection step). For all the studied subjects, the largest search window size found was 15x15x15 voxels.

The automatic local adaptation of the smoothing parameter $h^2(x_i)$ (see Eq. 1) proposed in Coupé et al. (2011b) has been slightly modified. During all the experiments, the squared smoothing parameter was set proportional (with λ=0.5) to the minimal SSD:

$$h^2(x_i) = \lambda^2 \times \arg\min_{x_{s,j}} \left\| P(x_i) - P(x_{s,j}) \right\|_2^2 + \varepsilon \qquad (5)$$

The value of lambda slightly changes the segmentation results. When we validated our segmentation method on the ADNI dataset in Coupé et al. (2011a), using λ=0.5 instead of λ=1 changed the median Dice-Kappa values from 0.882 to 0.883 for CN and from 0.836 to 0.838 for AD.

Finally, a subject selection was also applied to reduce the number of training MRI required as suggested in Aljabar et al. (2009). For each structure, the *N* closest subjects (in terms of SSD over the initialization mask as done in Coupé et al. (2011b)) are equally selected from both populations (*N/2* from the CN population and *N/2* from the AD population) (see Fig. 1). This is done to ensure that the size of the *"patch pool"* from the AD population is coarsely similar to the size of the *"patch pool"* from the CN population.

For a given subject with *N=20 (i.e., 10 AD training templates and 10 CN training templates)*, the segmentation and the grading maps were obtained in less than 4 minutes for left and right HC and less than 2 minutes for left and right EC using a single core of an Intel Core 2 Quad Q6700 processor at 2.66 GHz.

2.4 Validation framework

Our validation framework was designed to compare the capability of different SNIPE-based biomarkers to discriminate between patients and controls. The biomarkers studied were: HC volume, HC grade, EC volume and EC grade as well as their combination.

First, to obtain the segmentation and the grade of the subjects within the **test dataset**, a leave-one-out procedure was performed over the 100 subjects using their corresponding automatic segmentations resulting from the label propagation step (see Fig. 1). For each subject, the *N* closest training subjects were selected from the 99 remaining subjects in the library The average grading value was then estimated over the EC and the HC segmentations (for both left and right sides) obtained at the same time (see an example in Fig. 2). These segmentations were also used to measure the HC and EC volumes in the stereotaxic space.

Once all the subjects had a volume and a grade for each structure, a quadratic discriminant analysis (QDA) was performed. Each subject was classified by performing a QDA over the 99 remaining subjects. This approach was applied to volume-based classification, grade-based classification and the combination of both for HC, EC and HC + EC. We found that QDA slightly improved the results compared to linear discriminant analysis, especially when the subject's age was used as an additional parameter. The success rate (SR), the specificity (SPE), the sensitivity (SEN), the positive predictive value (PPV) and negative predictive value (NPV) are presented for each of the tested biomarkers (see (Cuingnet et al., 2010) for details on these quality metrics).

3. Results

Figure 2 shows the grading maps obtained for 2 test subjects (1 CN and 1 AD). The corresponding average grading values and the estimated volumes are also provided for left and right HC and for left and right EC. Visually, the CN subject clearly appears closer to the CN population (mainly red color related to values close to 1) while the AD patient is visually closer to the AD population (mainly purple and black colors related to values close to -1). In addition, Fig. 2 also provides a visual assessment of the quality of the segmentation and grading.

3.1 Volumetric study

The left column of Fig. 3 shows the volumes for the 100 subjects of the **test dataset** for HC and EC for *N=80* (i.e., 40 CN and 40 AD). The volumetric approach provided a classification success rate of 80% for HC and 69% for EC. The use of both structures at the same time produced a success rate of 78% through our QDA-based classification. This result indicates that the estimated HC volume is more powerful than the EC volume to identify patients with AD. This observation is in accordance with Frisoni et al. (1999). Our result using only HC volume is slightly superior to a recently published method comparison (Cuingnet et al., 2010). This might come from differences in the test dataset used here or due to a higher accuracy and consistency of the segmentation method used compared to Chupin et al. (2009b). The success rate obtained with EC volume is similar to the results reported in Frisoni et al. (1999) but lower than the values reported in other studies (Devanand et al., 2007; Xu et al., 2000), all using manual segmentations. Figure 3 shows the higher variability of EC volume compared to HC volume. As mentioned in the introduction, this range of volumes comes from the high inter-subject variability of EC, but may also be due to the difficulty to

distinguish EC structure boundaries on anatomical MRI (e.g., identification of the collateral sulcus and the sulcus semiannularis). Due to this last point, less accurate segmentations may be obtained for this structure and thus the introduction of segmentation errors may negatively impact the patient's classification. The use of both structures at the same time did not improve the result compared to the method based on HC only, while improvements have been observed in Devanand et al. (2007) doing similar experiments on manual segmentations.

3.2 Grading study

The right column of Fig. 3 shows the grading values for the 100 subjects of the **test dataset** for HC and EC for *N=80*. The success rate of the classification was 89% for HC, 78% for EC and 90% for the combination of both structures. For HC, the success rate obtained by using QDA is similar to thresholding the grading value at zero (4 false positives CN and 7 false negatives AD). In fact, in the perfect case, the 50 first subjects (CN) should have positive average grading values and the 50 last (AD) should have negative average grading values. This result indicates that the HC grade estimator is not biased and thus that the sign of the final grading value can be used directly to classify the patient. On the other hand, the EC grade estimator is biased in the sense that the optimal threshold obtained using QDA is superior to zero. As shown on Fig 3, the EC grades of AD are frequently superior to zero, thus indicating a higher similarity with the patches present in CN population. As we will show later, the normal age-related structural changes in the EC may disturb the detection of the disease-related anatomical changes. However, this bias, which depends on the training library used, can be partially compensated for by using QDA, yielding a success rate of 78%. Finally, by computing the average grade value over the HC and the EC improved the HC results and leads to a very high success rate of 90%.

3.3 Comparison of SNIPE anatomical biomarkers

In Tab. 1, the SEN, SPE, PPV and NPV obtained by the different SNIPE-based biomarkers considered are presented. These results show that for both structures studied, the classification based on grading provides significantly better results than the volumetric approach (89% vs. 80% for HC and 78% vs. 69% for EC). Moreover, while the combination of HC+EC tends to spoil the results of volumetric analysis, the combination of both slightly improves the results of the grading study. Three different combinations of biomarkers obtained a success rate of 90% during our experiments: HC volume and grade, HC + EC grade, HC + EC volume and grade. In the three cases, the HC grading was used, indicating a potential key role of this new imaging biomarker.

Table 1: *Results of the patient classification (AD vs CN) for the different SNIPE-based biomarkers under investigation. These results were obtained by using discriminant analysis through a leave-one-out procedure on the* ***test dataset*** *with N = 80 (i.e., 40 CN and 40 AD).*

AD vs. CN	*SR*	*SEN*	*SPE*	*PPV*	*NPV*
HC volume	80%	78%	82%	81%	79%
HC grading	89%	86%	92%	91%	87%
HC volume and grading	**90%**	**88%**	92%	92%	**88%**
EC volume	69%	66%	72%	70%	68%
EC grading	78%	74%	82%	80%	76%
EC volume and grading	78%	74%	82%	80%	76%
HC + EC volume	78%	76%	80%	79%	77%
HC + EC grading	**90%**	86%	**94%**	**93%**	87%
HC + EC volume and grading	**90%**	**88%**	92%	92%	**88%**

3.4 Impact of the number of selected best training subjects

This experiment presents the impact of the number of selected best training subjects on the studied biomarkers. Figure 4 presents the success rate for all the biomarkers from *N=20* (10 CN and 10 AD) to *N=80* (40 CN and 40 AD). As some authors have noted that using the age of subjects could increase the classification accuracy (Chupin et al., 2009a; Devanand et al., 2007), we used the subject's age as supplementary information during QDA.

Volume (see top of Fig. 4): For HC, the classification accuracy was quite stable from *N=40* to *N=80*. In (Coupé et al., 2011b), we showed that a plateau in terms of segmentation accuracy was reached around *N = 30*. For EC, the best results were obtained for *N=80*. This result seems to indicate that a large library is required to achieve consistent segmentation of EC. Indeed, increasing the size of the *"patch pool"* and better address issues related to inter-subject variability. The addition of the age as parameter in QDA improved the results of the classification, especially for EC and HC+EC biomarkers. By performing the QDA only with age provided a success rate of 48% in the classification. Finally, at *N=60*, the HC volume combined with the age provided a success rate of 82%.

Grade (see middle of Fig. 4): For HC, the best classifications were obtained by using high *N* values (*N=60* and *N=80*). For EC, the best classification rate was obtained for the smallest value of *N=20*, a result that was not expected. However, by also using age, the best results were obtained for *N=80* for EC. For HC and for HC+EC, using the age improved the results of the classification. In these cases, HC-based classification yielded a success rate of 92% and HC+EC a success rate of 93% at *N=40* and *N=60*.

Volume + Grade (see bottom of Fig. 4): By combining the volume and the grade of the SNIPE biomarkers, we obtained slightly better results than by using only the grade, except for EC. By using the age of the subjects, the volume and the grade over HC (with *N=60*) provided 92% classification accuracy. Combining all the parameters (i.e., volume, grade and age) slightly decreased the results for the biomarkers involving EC compared to use only grade and age.

3.5 Relationship between SNIPE grade and age

As shown in the previous experiment, using the subject's age improved the classification based on the grading measure, except for EC with *N=20*. This supplementary information seems to help distinguish age-related MRI changes from those related to AD pathology. Figure 5 shows the grade values as a function of age on HC + EC with *N = 60 (*the case with the highest classification accuracy: 93%). It appears that the grading values decrease with age in both populations. This variation indicates that the grading measure captures the age-related anatomical changes (possibly related to atrophy), and thus this observation may explain the better results obtained using age for all the biomarkers except for EC with *N=20*. As previously mentioned, QDA provides slightly better results than LDA during classification (between 0 to 2% depending on the biomarker studied). This slightly better fitting is assessed by Pearson's coefficient and corresponding *p*-value of the linear and quadratic regressions presented in Fig. 5. While for CN, the traditional linear model and quadratic provided similar results, it seems that for AD a quadratic model fits better than a linear model. The nonlinear nature of the atrophy related to AD has recently been studied (Frisoni et al., 2010; Frisoni et al., 2009; Jack et al., 2008). As noticed in (Frisoni et al., 2010), while the majority of studies are based on the linear assumption of the AD progression, brain atrophy during AD is not a linear process. In addition, the grade measure is correlated with age while the volume does not appear to be statistically correlated with age since similar regressions provided correlation of $r = 0.31$ for CN and $r = 0.34$ for AD with respective *p*-values of 0.09 and 0.06.

3.6 Relationship between SNIPE grade and MMSE score

Finally, the link between the mini mental state examination (MMSE) score and the grade is studied. The MMSE is a test evaluating the cognitive function of the patient. As noticed in (Black, 1999), a useful imaging biomarker should have a link with the cognitive decline of the patient with AD usually estimated by using MMSE. Several studies have investigated the relationship between the MMSE score and the volume or the shape of key structures such as HC (Devanand et al., 2007; Du et al., 2001; Thompson et al., 2004) and EC (Devanand et al., 2007; Du et al., 2001; Kenny et al., 2008). As done in (Du et al., 2001), we investigated the correlation between MMSE score and anatomical measurements (i.e., volume and grade) for HC and EC. Fig. 6 shows the plots of the grade and the volume as functions of the MMSE score. For both structures, the coefficient of correlation for grade was higher ($r = 0.75$ for HC and $r = 0.58$ for EC) than for the volume ($r = 0.55$ for HC and $r = 0.28$ for EC). A statistically significant correlation has been found in all cases, consistent with previous literature (Du et al., 2001; Kenny et al., 2008; Thompson et al., 2004). Another trend was that the HC measurements were more consistent with MMSE scores than EC measurements (see Fig. 6). Finally, the HC grade was the biomarker most consistent with MMSE with a high coefficient of correlation ($r = 0.75$).

In Du et al. (2001), the authors obtained a correlation coefficient of $r = 0.48$ for HC and $r = 0.48$ for EC volume based on manual segmentations with a p-value less than 0.001 in both cases. In our experiment, slightly higher correlation was obtained for HC, but a significantly lower value was obtained for EC as assessed by our higher p-value=0.005. However, our correlation coefficient between EC volume and MMSE score is similar to the correlation presented in Kenny et al. (2008) (r = 0.34). It should be noted that the estimation of correlation on discrete functions such as MMSE can bias the significance of correlation. However, we wanted to compare our results with previously published studies using this metric.

4 Discussion

During our experiments we showed that: *i)* HC-based measures were more discriminant than EC-based measures, *ii)* the grading had a higher discriminatory capability than the volume, *iii)* by adding the age, the classification rate improved, especially when using the HC-grade-based metrics, *iv)* by computing the grade over a larger area (HC+EC) tended to slightly improve results, especially when the subjects' ages were used within the classification model, and *v)* the optimal size of the number of selected training subjects were *N=60* (60% of the full library) in the majority of the situations studied. A balance appears to be required between using a large enough training population and potentially introducing outlier subjects by using all the available subjects. According to the structure of interest, a different number of training subjects could be used. Moreover, by using a larger library, it could be possible to select a higher number of subjects without introduction of outliers. The difficult segmentation of EC due to inter-subject variability could be partially compensated by using non-linear registration of training subjects instead of linear registration. However, this type of approach is more computational intensive. The introduction of shape priors (Hu et al., 2011) could also be a possibility to deal with ambiguity of the EC boundaries.

The SEN, SPE, PPV and NPV obtained by SNIPE are competitive compared to the ten methods compared in Cuingnet et al. (2010) involving voxel-based morphometry (VBM) (Ashburner and Friston, 2000), cortical thickness (Fischl et al., 1999), HC volume (Chupin et al., 2009b) and HC shape (Gerardin et al., 2009). In that comparison paper, the best VBM-based approach obtained 89% accuracy; the best method based on cortical thickness obtained 85% accuracy, the best approach using HC volume 74% accuracy and the method using HC shape 77% accuracy. However, during our experiment, only a subset of the entire ADNI database has been used, contrary to the experiments done in (Cuingnet et al., 2010). Moreover, the classification algorithm used in Cuingnet et al. (2010) was a support vector machine while we used a quadratic discriminant analysis approach. Despite these differences, the classification results obtained by using grade only are competitive to the best results reported in Cuingnet et al. (2010). Moreover, by adding the subjects' age yielded an accuracy of 93%. This result is similar to the highest classification accuracy 93.3% reported on a similar sized subset of ADNI (51 AD and 52 CN) in Zhang et al. (2011). However, Zhang *et al.* (2011) used a multimodal approach involving positron emission tomography (PET) and cerebro-spinal fluid (CSF) markers to reach this degree of accuracy. By using only MRI, their method based on volumetric features provided an accuracy of 86.2%.

It appears that using a larger area of analysis by grading several structures tended to improve the grading estimation. The extension of grading to other key structures impacted by AD seems to be an interesting path to follow for further research. Structures such as parahippocampal cortex and perihinal cortex (Dickerson et al., 2009) or fornix and mammillary body (Copenhaver et al., 2006) could be valuable anatomical structures to improve AD detection. Moreover, further work should investigate the spatial distribution of grade maps over the populations. This information could help to detect more discriminant areas for classification and might provide information on the AD progression. Finally, the application of the proposed grading measure to other diseases has a great potential. Moreover, the difficult problem of clinical differentiation (such as AD and frontal lobe dementia for instance) should also be investigated.

Using SSD as similarity metric, our approach is sensitive to inaccuracy in inter-subject intensity normalization. In Coupé et al. (2011a; 2011b), we demonstrated that the proposed preprocessing pipeline involving (Nyul and Udupa, 2000) provides a sufficiently robust normalization to obtain accurate segmentations. In this paper, we also showed that the preprocessing pipeline used yields high classification accuracy. Nevertheless, any improvements on the inter-subject normalization should yield further improvements in grading estimation. The use of other similarity metrics less sensitive to intensity normalization should be studied in future work. However, according to our experiments, there is no trivial solution since cross-correlation or correlation ratio cannot distinguish constant areas with different means (e.g, in CSF and white matter), mutual information requires a higher number of samples (bigger patch) and introduces the binning problem for histogram construction, and finally the SSIM index also requires matching of intensity. The use of hybrid metrics based on intensity and derivatives could be further investigated.

As for VBM-based approaches, SNIPE requires several scans of each population to be usable. The construction of a large enough training library might be an issue for trials based on a small number of subjects. However, the number of training subjects required by SNIPE is similar to the number required by VBM studies. As noted by Pell et al. (2008), a group size of 30 to 50 subjects per population is typical in a VBM study while a group size of 70-90 subjects per population is optimal for detection of HC volume loss. In our experiment, we found that 30 subjects from each population is sufficient to provide very high classification rates.

In this proof of concept study of our proposed grading technique, we focused on the problem of AD vs. CN classification. However, the prediction of conversion from prodromal AD (also known as mild cognitive impairment or MCI) to clinically definite AD is more useful from a clinical and diagnostic point of view. The prediction of patients with MCI who will convert to AD and those who will stay stable is an extremely complex task for which no method has yet provided satisfactory classification results (Cuingnet et al., 2010; Davatzikos et al., 2010). Proposed methods based on structural MRI have been focusing on gray matter loss as markers for prediction. Our proposed grading and segmentation method SNIPE may add valuable information for the problem of prediction.

5 Conclusion

In this paper, a new method is proposed to robustly detect the patterns of anatomical change in the hippocampus and entorhinal cortex caused by AD. Based on a nonlocal means estimation framework, the proposed novel grading measure (i.e., anatomical change possibly related to atrophy in the context of AD) enables an accurate distinction between CN subjects and patients with AD leading to a classification success rate of 90%. When the subject's age is combined with the grading measure, a success rate of 93% was obtained. These results are competitive compared to the AD detection performance of VBM, cortical thickness, HC volume and HC shape methods extensively compared in (Cuingnet et al., 2010). In contrast to these approaches, SNIPE has the advantage of: *i)* simplicity (it can be coded in few hundred lines of code), *ii)* low computational cost (as it does not require non-rigid registration), *iii)* robustness of the process (all the subjects get final grading maps) and *iv)* the possibility to achieve *individual* classifications based on a MRI data from a single time point (contrary to *group* classifications or longitudinal studies). These first results are promising and indicate that this new structure grading approach could be a useful biomarker to efficiently detect AD. Further work will investigate the possibility to discriminate populations of patients with MCI compared to AD or CN and furthermore, even the possibility of predicting AD.

Acknowledgments

This study was supported by the Canadian Institutes of Health Research (CIHR, MOP-84360 and MOP-111169) and CDA (CECR)-Gevas-OE016. This work was also partially supported by the Spanish Health Institute Carlos III through the RETICS Combiomed, RD07/0067/2001. This work benefited from the use of ITK-SNAP from the Insight Segmentation and Registration Toolkit (ITK) for 3D rendering.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46, 726-738.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry--the methods. Neuroimage 11, 805-821.

Barnes, J., Foster, J., Boyes, R.G., Pepple, T., Moore, E.K., Schott, J.M., Frost, C., Scahill, R.I., Fox, N.C., 2008. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. Neuroimage 40, 1655-1671.

Bishop, C.A., Jenkinson, M., Andersson, J., Declerck, J., Merhof, D., 2011. Novel Fast Marching for Automated Segmentation of the Hippocampus (FMASH): method and validation on clinical data. Neuroimage 55, 1009-1019.

Black, S.E., 1999. The search for diagnostic and progression markers in AD: so near but still too far? Neurology 52, 1533-1534.

Brox, T., Kleinschmidt, O., Cremers, D., 2008. Efficient nonlocal means for denoising of textural patterns. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 17, 1083-1092.

Buades, A., Coll, B., Morel, J.M., 2005. A non-local algorithm for image denoising. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 2, Proceedings, 60-65.

Buades, A., Coll, B., Morel, J.M., 2010. Image Denoising Methods. A New Nonlocal Principle. Siam Review 52, 113-147.

Chupin, M., Gerardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehericy, S., Benali, H., Garnero, L., Colliot, O., 2009a. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579-587.

Chupin, M., Hammers, A., Liu, R.S., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., Garnero, L., Lemieux, L., 2009b. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. Neuroimage 46, 749-761.

Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J Comput Assist Tomogr 18, 192-205.

Collins, D.L., Pruessner, J.C., 2010. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 52, 1355-1366.

Colliot, O., Chetelat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., Lehericy, S., 2008. Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. Radiology 248, 194-201.

Copenhaver, B.R., Rabin, L.A., Saykin, A.J., Roth, R.M., Wishart, H.A., Flashman, L.A., Santulli, R.B., McHugh, T.L., Mamourian, A.C., 2006. The fornix and mammillary bodies in older adults with Alzheimer's disease, mild cognitive impairment, and cognitive complaints: a volumetric MRI study. Psychiatry Res 147, 93-103.

Coupé, P., Fonov, V., Eskildsen, S., Manjon, J., Arnold, D., Collins, L., 2011a. Influence of the training library composition on a patch-based label fusion method: Application to hippocampus segmentation on the ADNI dataset. Alzheimer's and Dementia 7, S24-S24.

Coupé, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2010. Nonlocal patch-based label fusion for hippocampus segmentation. Med Image Comput Comput Assist Interv 13, 129-136.

Coupé, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011b. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54, 940-954.

Coupé, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans Med Imaging 27, 425-441.

Csernansky, J.G., Wang, L., Swank, J., Miller, J.P., Gado, M., McKeel, D., Miller, M.I., Morris, J.C., 2005. Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage 25, 783-792.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2010. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2010. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of Aging.

Devanand, D.P., Pradhaban, G., Liu, X., Khandji, A., De Santi, S., Segal, S., Rusinek, H., Pelton, G.H., Honig, L.S., Mayeux, R., Stern, Y., Tabert, M.H., de Leon, M.J., 2007. Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease. Neurology 68, 828-836.

Dickerson, B.C., Feczko, E., Augustinack, J.C., Pacheco, J., Morris, J.C., Fischl, B., Buckner, R.L., 2009. Differential effects of aging and Alzheimer's disease on medial temporal lobe cortical thickness and surface area. Neurobiology of Aging 30, 432-440.

Du, A.T., Schuff, N., Amend, D., Laakso, M.P., Hsu, Y.Y., Jagust, W.J., Yaffe, K., Kramer, J.H., Reed, B., Norman, D., Chui, H.C., Weiner, M.W., 2001. Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. J Neurol Neurosurg Psychiatry 71, 441-447.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9, 195-207.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6, 67-77.

Frisoni, G.B., Laakso, M.P., Beltramello, A., Geroldi, C., Bianchetti, A., Soininen, H., Trabucchi, M., 1999. Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer's disease. Neurology 52, 91-100.

Frisoni, G.B., Prestia, A., Rasser, P.E., Bonetti, M., Thompson, P.M., 2009. In vivo mapping of incremental cortical atrophy from incipient to overt Alzheimer's disease. J Neurol 256, 916-924.

Gerardin, E., Chetelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. Neuroimage 47, 1476-1486.

Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A., 2008. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. Neuroimage 40, 672-684.

Gutman, B., Wang, Y., Morra, J., Toga, A.W., Thompson, P.M., 2009. Disease classification with hippocampal shape invariants. Hippocampus 19, 572-578.

Henneman, W.J., Sluimer, J.D., Barnes, J., van der Flier, W.M., Sluimer, I.C., Fox, N.C., Scheltens, P., Vrenken, H., Barkhof, F., 2009. Hippocampal atrophy rates in Alzheimer disease: added value over whole brain volume measures. Neurology 72, 999-1007.

Hu, S., Coupé, P., Pruessner, J.C., Collins, D.L., 2011. Appearance-based modeling for segmentation of hippocampus and amygdala using multi-contrast MR imaging. Neuroimage.

Jack, C.R., Jr., Weigand, S.D., Shiung, M.M., Przybelski, S.A., O'Brien, P.C., Gunter, J.L., Knopman, D.S., Boeve, B.F., Smith, G.E., Petersen, R.C., 2008. Atrophy rates accelerate in amnestic mild cognitive impairment. Neurology 70, 1740-1752.

Juottonen, K., Laakso, M.P., Partanen, K., Soininen, H., 1999. Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. AJNR Am J Neuroradiol 20, 139-144.

Kenny, E.R., Burton, E.J., O'Brien, J.T., 2008. A volumetric magnetic resonance imaging study of entorhinal cortex volume in dementia with lewy bodies. A comparison with Alzheimer's disease and Parkinson's disease with and without dementia. Dement Geriatr Cogn Disord 26, 218-225.

Khan, A.R., Cherbuin, N., Wen, W., Anstey, K.J., Sachdev, P., Beg, M.F., 2011. Optimal weights for local multi-atlas fusion using supervised learning and dynamic information (SuperDyn): validation on hippocampus segmentation. Neuroimage 56, 126-139.

Lotjonen, J.M., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D., 2010. Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49, 2352-2365.

Morey, R.A., Petty, C.M., Xu, Y., Hayes, J.P., Wagner, H.R., 2nd, Lewis, D.V., LaBar, K.S., Styner, M., McCarthy, G., 2009. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. Neuroimage 45, 855-866.

Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M., 2010. Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE Trans Med Imaging 29, 30-43.

Mueller, S.G., Schuff, N., Yaffe, K., Madison, C., Miller, B., Weiner, M.W., 2010. Hippocampal atrophy patterns in mild cognitive impairment and Alzheimer's disease. Human Brain Mapping 31, 1339-1347.

Nyul, L.G., Udupa, J.K., 2000. Standardizing the MR image intensity scales: making MR intensities have tissue specific meaning. Medical Imaging 2000: Image Display and Visualization 1, 496-504

606.

Pell, G.S., Briellmann, R.S., Chan, C.H., Pardoe, H., Abbott, D.F., Jackson, G.D., 2008. Selection of the control group for VBM analysis: influence of covariates, matching and sample size. Neuroimage 41, 1324-1335.

Pohl, K.M., Bouix, S., Shenton, M.E., Grimson, W.E., Kikinis, R., 2007. Automatic Segmentation Using Non-Rigid Registration. Med Image Comput Comput Assist Interv 26, 1201-1212.

Protter, M., Elad, M., Takeda, H., Milanfar, P., 2009. Generalizing the nonlocal-means to super-resolution reconstruction. IEEE Trans Image Process 18, 36-51.

Pruessner, J.C., Kohler, S., Crane, J., Pruessner, M., Lord, C., Byrne, A., Kabani, N., Collins, D.L., Evans, A.C., 2002. Volumetry of temporopolar, perirhinal, entorhinal and parahippocampal cortex from high-resolution MR images: considering the variability of the collateral sulcus. Cereb Cortex 12, 1342-1353.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. Ieee Transactions on Medical Imaging 17, 87-97.

Takeda, H., Milanfar, P., Protter, M., Elad, M., 2009. Super-resolution without explicit subpixel motion estimation. IEEE Trans Image Process 18, 1958-1975.

Thompson, P.M., Hayashi, K.M., De Zubicaray, G.I., Janke, A.L., Rose, S.E., Semple, J., Hong, M.S., Herman, D.H., Gravano, D., Doddrell, D.M., Toga, A.W., 2004. Mapping hippocampal and ventricular change in Alzheimer disease. Neuroimage 22, 1754-1766.

van der Lijn, F., den Heijer, T., Breteler, M.M., Niessen, W.J., 2008. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. Neuroimage 43, 708-720.

Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L.L., Augustinack, J., Dickerson, B.C., Golland, P., Fischl, B., 2009. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus 19, 549-557.

Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A., 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmentation. Neuroimage 55, 968-985.

Wolz, R., Heckemann, R.A., Aljabar, P., Hajnal, J.V., Hammers, A., Lotjonen, J., Rueckert, D., 2010. Measurement of hippocampal atrophy using 4D graph-cut segmentation: application to ADNI. Neuroimage 52, 109-118.

Xu, Y., Jack, C.R., Jr., O'Brien, P.C., Kokmen, E., Smith, G.E., Ivnik, R.J., Boeve, B.F., Tangalos, R.G., Petersen, R.C., 2000. Usefulness of MRI measures of entorhinal cortex versus hippocampus in AD. Neurology 54, 1760-1767.

Yushkevich, P.A., Wang, H., Pluta, J., Das, S.R., Craige, C., Avants, B.B., Weiner, M.W., Mueller, S., 2010. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. Neuroimage 53, 1208-1224.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55, 856-867.

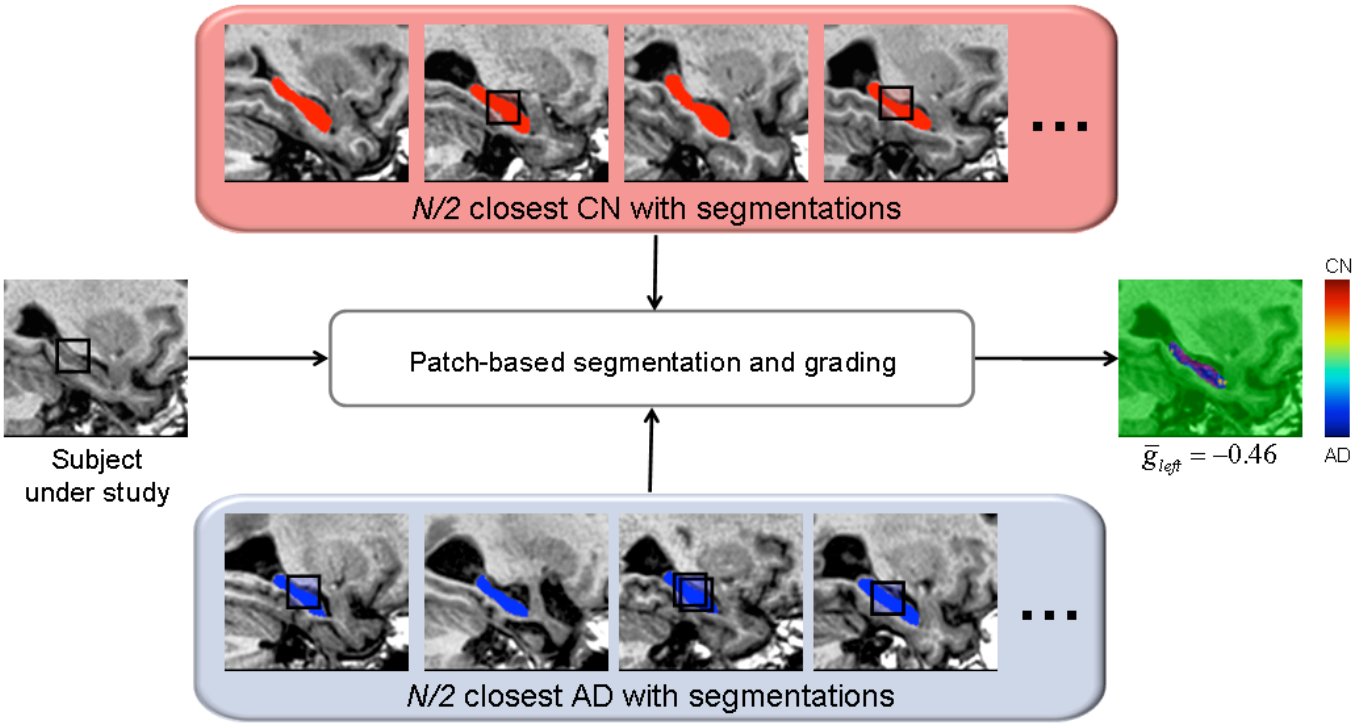
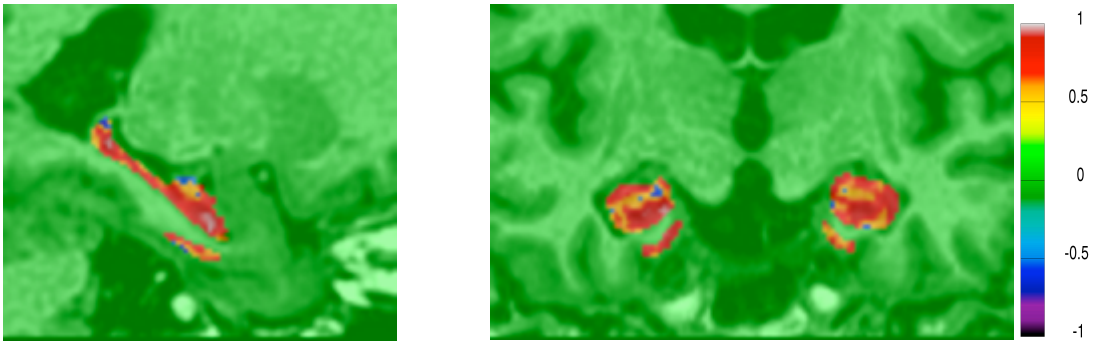


Figure 1: **Global overview of the proposed grading and segmentation method SNIPE.** The *N/2* closest subjects in the training library are selected from both populations. All these subjects have automatic segmentations, resulting from the label propagation step, of the structures studied (the left HC segmentations are displayed in this example). For each patch of the subject under study (an AD patient in this example), a comparison is performed with all the patches constituting the selected training subjects (some examples of similar patches are displayed). In this way, the simultaneous segmentation and grading is obtained for the structure studied. The final grading value corresponds to the average value over the estimated segmentation. This procedure is carried out for each structure studied: left and right HC and left and right EC.



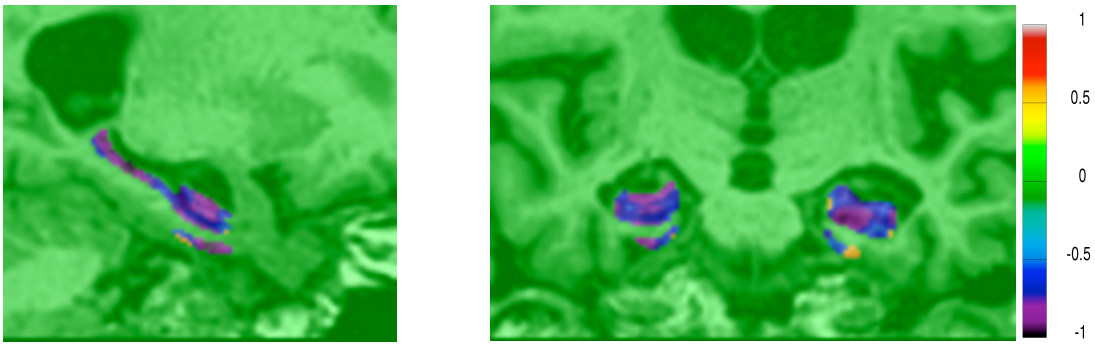
CN subject (ID 34)

Left HC: $\overline{g}_{left} = 0.65$ / volume = 3.73 cm^3

Right HC: $\overline{g}_{right} = 0.51$ / volume = 3.87 cm^3

Left EC: $\overline{g}_{left} = 0.74$ / volume = 0.36 cm^3

Right EC: $\overline{g}_{right} = 0.58$ / volume = 0.26 cm^3



AD patient (ID 73)

Left HC: $\overline{g}_{left} = -0.72$ / volume = 2.89 cm^3

Right HC: $\overline{g}_{right} = -0.63$ / volume = 2.73 cm^3

Left EC: $\overline{g}_{left} = -0.62$ / volume = 0.29 cm^3

Right EC: $\overline{g}_{right} = -0.26$ / volume = 0.32 cm^3

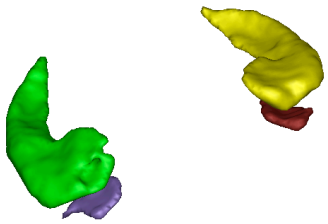


Figure 2. The SNIPE grading maps obtained for (top) one CN subject (ID 34) and (bottom) one AD patient (ID 73), plotted with a color scale from -1 (AD) to +1 (CN). The subject IDs are the same as those used in Fig. 3. The image slices of both subjects have the same position in stereotaxic space. The volumes in stereotaxic space and the average grade values for each structure are provided. For each subject, 3D renderings of the segmentations are presented.

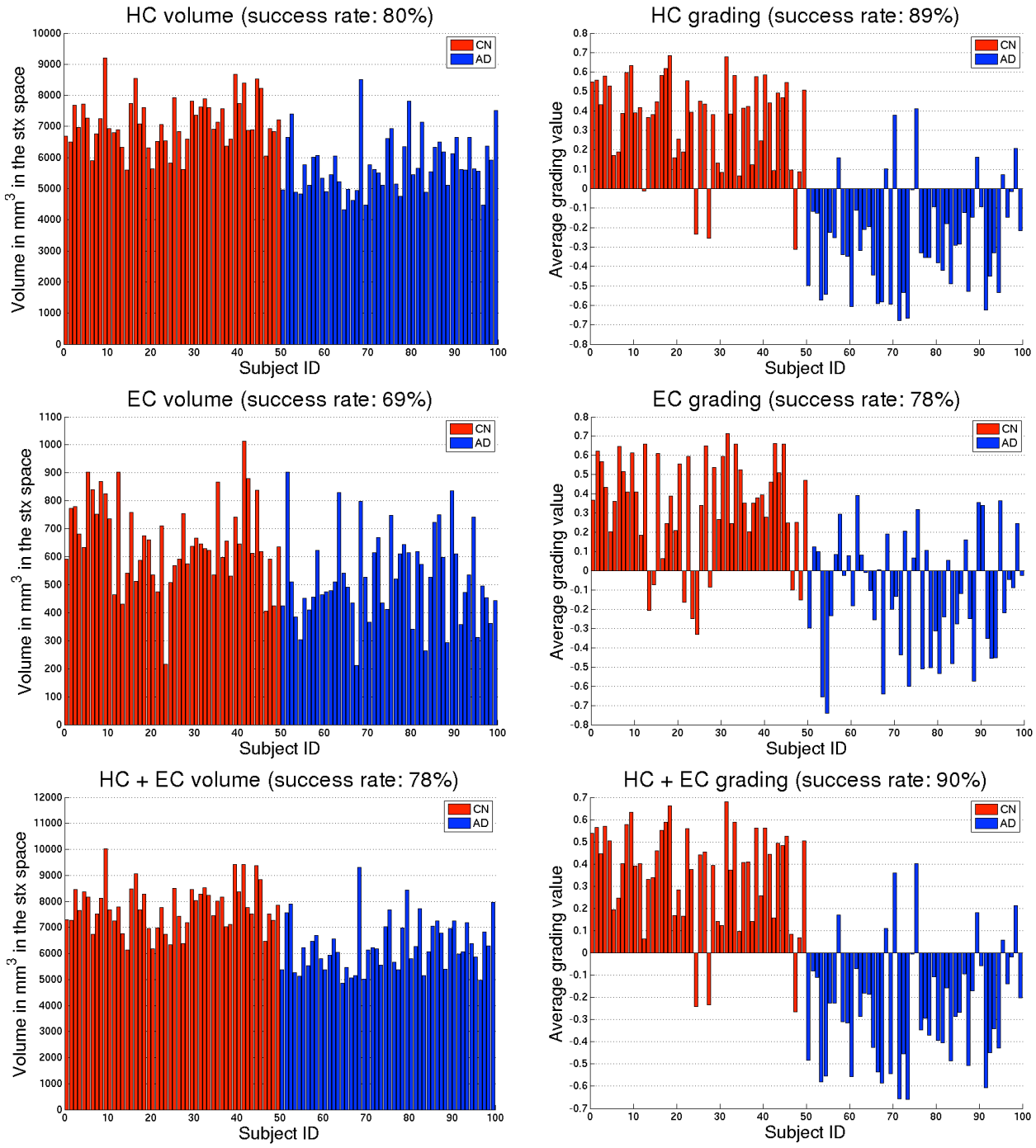


Figure 3: Volumes and grading values of the studied structures for all the subjects (with $N = 80$). The success rates are obtained using quadratic discriminant analysis between both populations in a leave-one-out fashion.

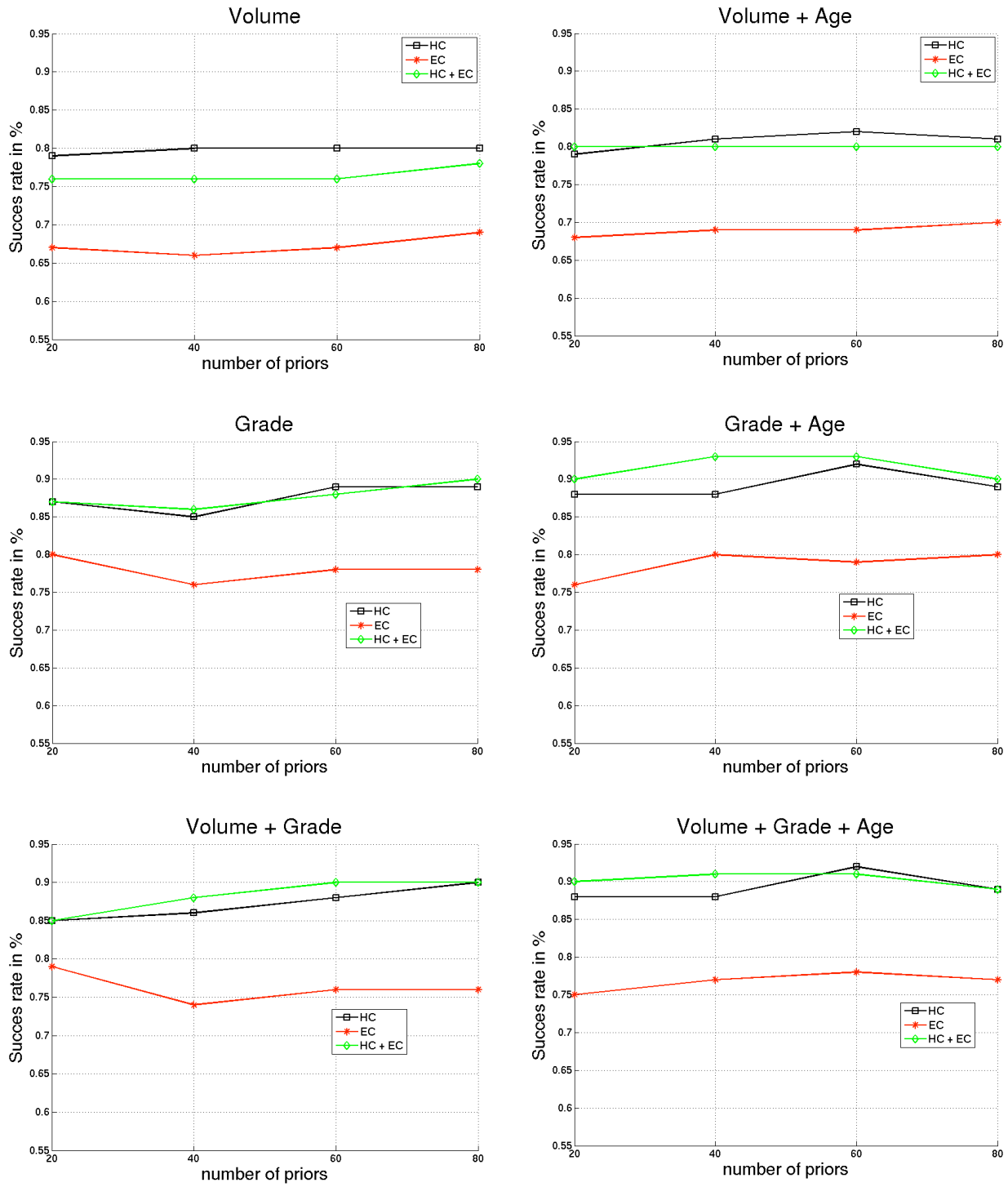


Figure 4: Impact of N on the classification accuracy using volume, grade and age as well as their combination. The success rates are obtained using QDA between both populations in a leave-one-out fashion.

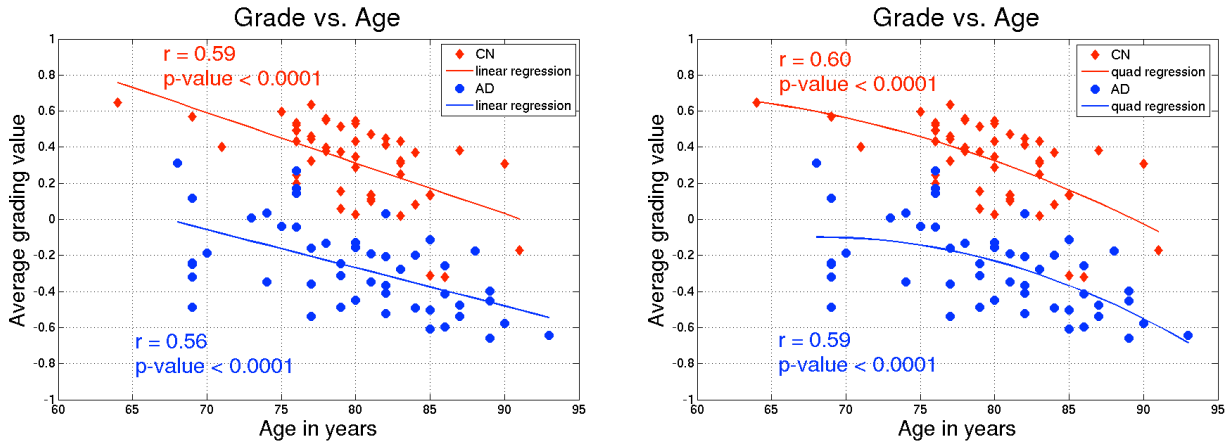


Figure 5: Analysis of the relation between age and grading values. The grading values were obtained on HC + EC with $N = 60$, the combination that provided the highest classification accuracy (93%). Linear and quadratic fitting were compared for both populations. The correlation coefficient r and the corresponding p-values are provided on the graphs.

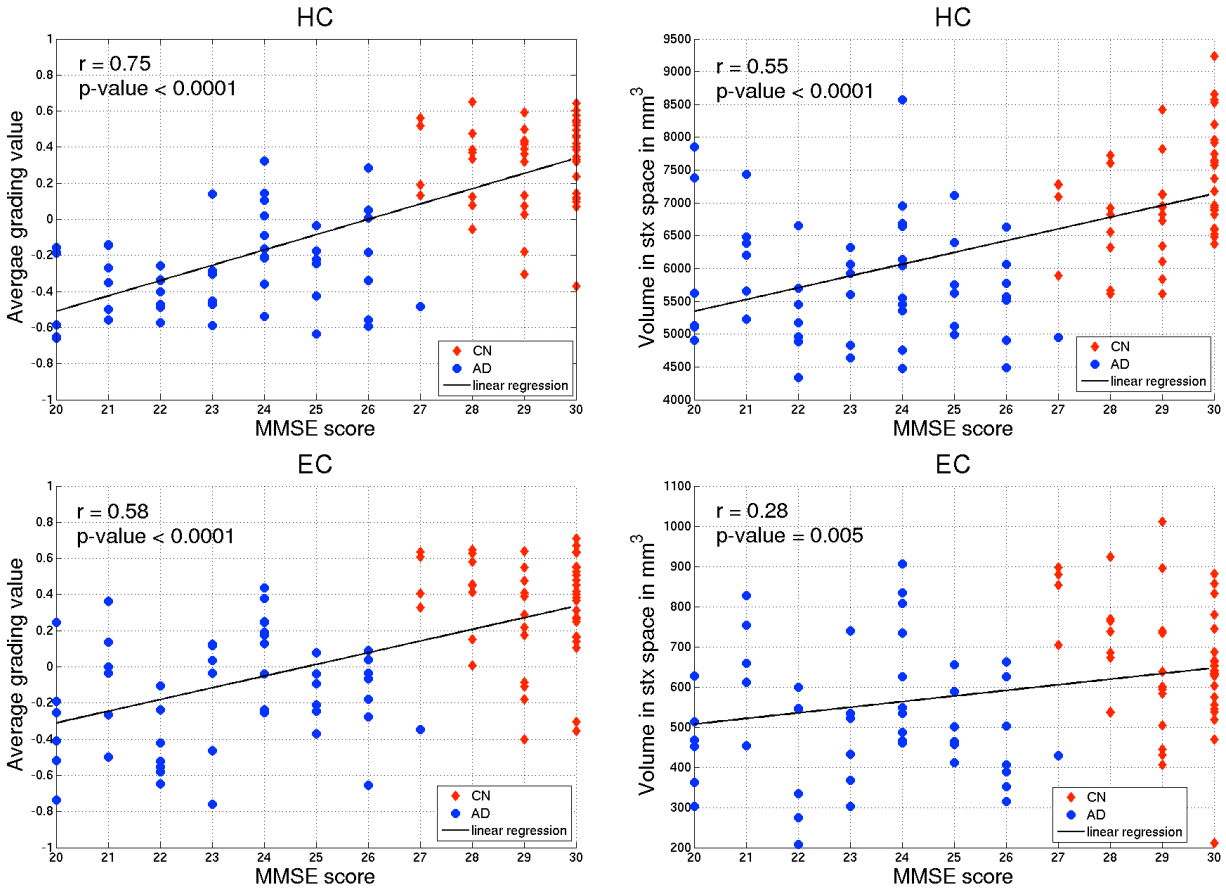


Figure 6: Analysis of the relation between the MMSE score and anatomical measurements (i.e., grade and volume) for HC and EC. The correlation coefficient *r* and the corresponding *p*-values are provided on the graphs.