Morphological Hippocampal Markers for Automated Detection of Alzheimer Disease and MCI converters in MR Images

Luca Ferrarini[a*] Giovanni B. Frisoni[b] Michela Pievani[b]
Johan H.C. Reiber[a] Rossana Ganzola[b] Julien Milles[a]

[a] LKEB - Department of Radiology, Division of Image Processing, Leiden University Medical Center, Leiden, The Netherlands,

[b] Laboratory of Epidemiology Neuroimaging and Telemedicine, IRCCS San Giovanni di Dio-FBF, Brescia, Italy

[*] Corresponding author: Dr. L. Ferrarini., Leiden University Medical Center, Department of Radiology, Division of Image Processing - Postzone C2S, Postbus 9600, 2300RC Leiden, The Netherlands. Tel/Fax: +31 71 - 5266206/5265342. Email: L.Ferrarini@lumc.nl.

Running head: Morph. MR hippocamp. markers for AD and MCI-c.
Number of Abstract's words: 193.

Abstract

Objective: To investigate the use of hippocampal shape-based markers for automatic detection of Alzheimer Disease (AD) patients and converted mild cognitive impairment (MCI) patients (MCI-c).

Material: Three-dimensional T1-weighted magnetic resonance images of 50 ADs, 50 age-matched controls, 15 MCI-c and 15 MCI-non-converters (MCI-nc). Manual delineations of both hippocampi were obtained from normalized images. Fully automatic shape modeling was used to generate comparable meshes for both structures.

Method/Results: Repeated permutation tests, run over a randomly sub-sampled training set (25 controls and 25 ADs), highlighted shape-based markers, mostly located in the CA1 sector, which consistently discriminated ADs and controls. Support vector machines (SVMs) were trained, using markers from either one or both hippocampi, to automatically classify controls and ADs. Leave-1-out cross-validations over the remaining 25 ADs and 25 controls resulted in an optimal accuracy of 90% (sensitivity 92%), for markers in the left hippocampus. The same morphological markers were used to train SVMs for MCI-c vs. MCI-nc classification: markers in the right hippocampus reached an accuracy (and sensitivity) of 80%. Due to the pattern recognition framework, our results statistically represent the expected performances of clinical setups, and compare favorably to analyses based on hippocampal volumes.

Keywords: Alzheimer's Disease, Mild Cognitive Impairment, Magnetic Resonance Images, Hippocampus, Morphological markers, Support Vector Machines

1 Introduction

In recent years, attempts to detect early markers for Alzheimer Disease (AD) have increased considerably. Therapies to slow down cognitive decline in AD might be more effective when administered during the early stage of the disease ([1]): consequently, markers for early diagnosis are a highly desirable tool to ease the life quality of patients affected by such disease.

Early studies focused mainly on highlighting anatomical or physiological differences between a cohort of patients affected by probable AD and a group of comparable

healthy elderly. Physiological studies focused on CSF sampling for protein examinations ([1–6]) (the latter in combination with volumetric changes of hippocampal atrophy). Alternative methods based on positron emission tomography (PET) metabolic imaging with radioactive glucose ([7]), and molecular imaging with amyloid tracers ([8]) proved to be accurate markers for early diagnosis and tracking of the disease progression. Although these methods have provided encouraging results, the sampling of CSF is highly invasive, requiring lumbar punctures, and the high costs of PET imaging (which is also minimally invasive) prevent its widespread use.

Magnetic resonance imaging (MRI), on the other hand, is non-invasive and largely available in the clinical environment: therefore, it has become a vital tool to monitor the progression of AD and to unveil its mechanisms. Three-dimensional representations of the brain might help researchers assessing subtle changes in the human brain. Changes in cortical thickness were analyzed in probable AD, and correlated with cognitive tests, highlighting cortical loss in several brain regions, and particularly in the parahippocampal gyrus ([9]). It is well known that medial temporal lobes, and in particular the hippocampi and entorhinal cortex, are among the first regions of the brain undergoing atrophy in AD ([10]): MR-based studies investigating either volumetric or morphological changes in the hippocampi and in adjacent regions (e.g. the ventricles), have shown a correlation between changes in these structures and the progression of cognitive decline in AD ([11–15]). Differences in loss of gray matter (GM) were highlighted between early and late onsets of AD: patients with early onset presented greater GM loss in the occipital and parietal lobes, while late onsets were mostly characterized by hippocampal atrophy ([16]).

Studies based on probable AD cohorts are important, since they cast light over the mechanisms with which the disease affects the brain. Nevertheless, in order to detect early markers of the pathology it is necessary to focus on cohorts known to be at risk for developing AD. Perfect candidates are patients affected by amnestic Mild Cognitive Impairment (MCI) ([17]): these patients experience isolated memory deficits (amnesia), but are otherwise cognitively capable and not limited in their daily life. Among these patients, some have shown rapid conversion to AD (MCI-converters) while others have presented a stable behavior (MCI non-converters) or even a regression to normal

cognitive conditions. Thus, the quest for early markers of AD can considerably benefit from studies based on MCI-c and MCI-nc. Greater global gray matter (GM) loss (in particular in the hippocampi, inferior and middle temporal lobes) has been highlighted in MCI-c ([18]). A specific pattern in GM atrophy progression was suggested by [19]: three years before AD diagnosis, GM loss is mainly found in the medial temporal lobes, including structures such as amygdala, anterior hippocampus, and entorhinal cortex. While the disease progresses, atrophy spreads further to the posterior part of the medial temporal lobes, including the entire body of the hippocampus, and finally reaches parietal and frontal lobes. A region-based study ([20]) focused on hippocampal volume showed that hippocampal atrophy of the right hippocampus correlates with AD and MCI, and that both hippocampal volumes correlate with cognitive tests. Spatial patterns of brain atrophy were also investigated in [21], where it was shown that MCI patients often present patterns of atrophy overlapping with those of AD patients. Finally, in [22] mild AD were analyzed with respect to healthy subjects aiming at an early detection of the pathology.

Statistical analyses are needed in order to assess the discriminative performances of brain changes due to AD progression. [23] presented a study based on hippocampal and ventricular area indexes, evaluated both in a cohort of probable AD and in a matched cohort of healthy elderly. The results showed that it was possible to separate controls from AD with high accuracy. The distribution of temporal horn index and temporal horn volume in AD, MCI, and healthy subjects showed statistically significant differences between AD and MCI, and AD and controls, and no significant difference between MCI and controls ([24]). Volumetric and morphological changes in the hippocampi were investigated in MCI-c, MCI-nc, and MCI which regressed to normal conditions, showing significant differences in the CA1 sector in the MCI-c compared to both the other groups ([25]).

Discriminant analyses are important to drive the attention of researchers towards potential AD markers. Nevertheless, it has been increasingly acknowledged that analyses rooted in a pattern recognition framework are more suitable for marker validation. In pattern recognition, one faces the problem of training an intelligent machine (i.e. classifier) to correctly classify instances into two or more classes. Important steps in-

clude: feature selection, to detect salient characteristics discriminating patients from controls; supervised training, to associate specific sets of features to a particular class; independent testing phase, to assess the performances of the classifier. The added value of a pattern recognition framework lies in the independence of these three steps, which provides a statistical estimation on how the classifier will behave in a clinical set-up, when new patients must be diagnosed. Logistic regression has shown that hippocampal morphological changes in the CA1 sector can discriminate mild AD from healthy subjects with an overall accuracy of 84% ([26]). The overall accuracy could be improved by including volumetric changes. Ventricular morphological changes (particularly, changes in the temporal horns) have been used to discriminate AD from healthy subjects with an overall accuracy of 80% ([13]). Automatically detected volumetric changes in a region of interest set around the hippocampi have been used to correctly discriminate AD and controls with an accuracy of 92% ([27]), while automatic classification based on intensity and local volume estimation were used in [28] to automatically detect MCI converters with an accuracy of 81%. Recent methods based on overall changes of gray matter ([21, 22, 29]) or cortical thickness ([30]) have reached even higher performances, especially when combined with genetic information. No final conclusion has yet been made on whether one should consider changes in the entire brain, or rather focus on particular brain regions, while looking for discriminative markers for AD. Ideal MR markers should be easy to detect automatically, sensitive enough to highlight AD, and at the same time specific enough to differentiate between AD and other forms of dementia. Some authors have stated that global brain changes, being able to capture a more global pattern of atrophy, are better suited to discriminate different forms of dementia ([22, 29]). On the other hand, the same argument has been used in favor of a region of interest analysis ([31]), claiming that global cerebral atrophy is not sufficiently specific for AD.

Hippocampal tissue loss in AD has been largely investigated with MRI. The majority of the studies have expressed atrophy in terms of decreased volume ([15, 23, 26, 32–37]). More recent studies, however, have moved towards shape-based analyses to better localize the hippocampus' surface areas mostly affected by the disease ([11, 14, 38]), showing that shape changes due to AD are mostly located on sector CA1 of the hip-

pocampus, while sectors CA2 and CA3 tend to be spared by atrophy. These insights present new challenging questions. Are there markers on the hippocampal surface which can be used for automatic AD detection? Can they be detected in MR images? Are they sufficiently specific and sensitive in differentiating AD subjects from healthy elderly? Can they be applied to discriminate MCI-c from MCI-nc? In a previous work ([13]), we investigated ventricular shape-based markers for AD, showing how the tips of temporal horns could be used to train a classifier in automatically discriminating AD subjects from healthy elderly: although the specificity (healthy elderly detection) was quite high, the sensitivity on independent AD tests reached 76%, leaving space for further improvements.

In this work, we apply the same methodology to detect shape-based markers in the hippocampi. Compared to previously published results, this work presents several novelties: from a technical point of view, the modeling of hippocampal surface (based on either manually or automatically segmented volumes) is fully automatic; moreover, the features are extracted from an independent set, removing the bias during the validation phase. Support vector machines (SVMs) ([39]) are used for the classification: the use of SVMs for diagnostic tool has increased remarkably, due to their excellent performances ([13, 22, 27, 29]). Another novelty of our approach is that features extracted for controls and AD are used to train the classifier for MCI-c and MCI-nc. A set of MCI-nc might always include subjects which will eventually develop AD: therefore, features extracted by MCI-nc might limit the final performances of a classifier. Finally, we present classifiers based on different sets of features: (1) using the entire hippocampal surface, (2) using approximately half of the hippocampal surface, as detected by our feature extraction method with a rather relaxed threshold, and (3) using specific areas of the hippocampal surface previously reported in literature as discriminative for AD (particularly the CA1 sector and subiculum); these sets are tested for either the right, left, or a combination of both hippocampi, investigating the potential advantages of combining information from both hemispheres. To conclude, we present similar analyses performed on volumetric markers for both right, left, and combined hippocampi, investigating their performances in comparison to shape-based classifiers.

2 Material and Method

2.1 Subjects and MRI acquisition

Fifty patients with probable AD (35 females, mean age 71.3 $\pm$ 7.7 years, range 52-84), and 50 volunteers with normal cognitive functions (31 females, mean age 70.8 $\pm$ 5.8 years, range 62-84) were included in the study. The patients with probable AD had been consecutively referred to the outpatient memory clinic of the Laboratory of Epidemiology, Neuroimaging, and Telemedicine, at the IRCSS San Giovanni di Dio-FBF in Brescia (Italy) between 2002 and 2008. Healthy subjects were selected among those enrolled in a study on normal brain structure with MRI, as described in detail elsewhere ([40]). Additionally, 30 patients with mild cognitive impairment (MCI) were included: during the period of under examination, 15 of them had turned into probable AD (7 females, mean age 72.4 $\pm$ 7.3, range 59-85, mean interval to conversion: 17 months), and 15 stable MCI were then age- and sex-matched to MCI converters (7 females, mean age 71.4 $\pm$ 5.2, range 62-83, mean interval of examination: 33 months).

All subjects were evaluated for memory loss using a standardized dementia screening including a detailed medical history, a general internal and neurological exam, laboratory tests, and an extensive neuropsychological battery including the Mini Mental State Examination (MMSE, [41]), and MRIs of the brain (the MRIs for the MCI patients included in this study are those at baseline). Diagnoses were made in a multidisciplinary consensus meeting based on the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimers Disease and Related Disorders Association (NINCDS-ADRDA) criteria for probable AD, while MCI patients were included when presenting (1) memory or other cognitive disturbances, (2) MMSE score between 24 and 28, (2) preserved general cognitive performance, (4) intact activity of daily living, and (5) were not demented. Patients and controls were included if they had no other neurologic or psychiatric illness, and had no abnormalities on MRI other than white matter hyperintensities or an incidental small lacunar lesion ($\leq$ 5 mm diameter). The study was approved by the local Medical Ethical Committee. Written informed consent was obtained from all subjects or from a close relative if a patient was demented. The socio-demographic and clinical features of all subjects are reported in

Table 1, together with the results of statistical tests applied to controls vs. AD and MCI converted vs. MCI non-converted.

MRI scans were performed on a 1.0 Tesla Philips Gyroscan at the Neuroradioloy Unit of the Cittá di Brescia hospital, Brescia. High-resolution sagittal T1-weighted MRIs were acquired with a gradient echo 3D technique using the following parameters: TR=20ms, TE=6ms, flip angle=30°, field of view=250mm, acquisition matrix=256x256, slice thickness=1.3mm.

2.2 Preprocessing

The MR images were linearly registered to a standard template (colin27, [42]) in the stereotaxic space using a combination of scripts written in Perl (http://www.perl.com) based on software developed at the McConnel Brain Imaging Center (Montreal Neurological Institute, McGill University, Montreal, Canada). Manual tracing of left and right hippocampal volumes was performed by an expert tracer (R.G.) following a standardized and validated protocol ([43]) on contiguous coronal 1.5mm thick images. Hippocampi were delineated with the software program DISPLAY (McGill University, Montreal, Canada) that allows to check segmentation accuracy simultaneously on the sagittal and axial planes. Tracing included the hippocampus proper, dentate gyrus, subiculum (subiculum proper and presubiculum), alveus, and fimbria. Test-retest reliability on 20 patients and controls was good, intra-class correlation coefficients were 0.92 for the left and 0.86 for the right hippocampus. Each binary volume was finally converted into a set of surface locations by extracting the surface voxels.

In order to highlight shape differences between subjects, it was first necessary to transform each set of surface points to a standard space, in order to correct for differences in position and orientation between the hippocampi in the different image spaces. As a standard space we chose to select the healthy subject which most resembled the other healthy individuals and AD patients (MCIs were not included in this analysis, since they represent an intermediate situation between these extremes). For each possible pair of subjects' MR images (considering both controls and AD subjects), a six-parameter rigid transformation was applied (FLIRT, [44, 45]): the co-registration error for the particular transformation was evaluated as the Euclidian distance between the

corresponding matrix and the identity matrix. For a given subject, the co-registration errors were averaged over all the 99 possible rigid transformations: the healthy subject with the minimum average co-registration error was finally selected as the most representative subject, and its image space considered as standard space.

Transformation matrices of all subjects (including the MCIs) to the chosen standard space were evaluated. Subsequently, for each subject the set of left and right hippocampal surface points was extracted from the manual delineation in subject-space and transformed into standard space. Thus, the final results of the preprocessing phase were sets of co-registered surface points, each representing the left and right hippocampi of a subject (see Fig. 2). It is worth mentioning that any pairwise approach might be biased towards the initial choice for the representative subject. However, in previous studies on brain ventricles ([12, 46]) we have already shown that choosing a particular individual as starting point only slightly decreases the overall accuracy of the final statistical model.

2.3 Modeling of the hippocampal shape

The first step towards a quantitative analysis of hippocampal shape differences is the modeling of all the instances' shapes. In a previous work, we introduced a novel method for shape modeling and analysis, GAMEs ([46]), and used it to highlight ventricular shape differences between controls and AD subjects ([12]) (the method is schematically represented in Fig. 1, applied to a synthetic shape). The accuracy of the method in detecting local shape changes was shown on tubular synthetic structures [46], by adding artificial variations which were then correctly picked up by GAME. The same method was applied to both the left and right hippocampi separately. First, a mesh is grown (adding nodes and edges) until convergence to the set of surface points of the healthy subject chosen to represent the standard space. Subsequently, the mesh basic characteristics (i.e., its number of nodes and edges) are frozen. In the second phase, the mesh is adapted (moving the nodes and edges in space, and without adding nor removing any of them) to all the other similar instances of the individuals included in the study. The adaptation is performed by applying the Kohonen self-organizing map algorithm ([47]), which is known to preserve topology. The final meshes (also called a

point distribution model, PDM) are locally comparable, since each node in a mesh is uniquely associated with a specific node in any other mesh: this is an essential property for local shape analysis, although not sufficient. For the shape analysis to be meaningful, corresponding nodes in the PDM have to be representative of similar anatomical locations. In previous works we have shown how the GAME algorithm successfully satisfies these conditions ([46, 12]), qualifying as a sound shape modeling technique. When applied to the populations in our study, GAMEs provided us with 130 meshes for the left hippocampus (50 controls, 50 AD, 30 MCIs: 431 nodes in each mesh), and 130 meshes for the right hippocampus (387 nodes per mesh) (see Fig. 2). The algorithm decides upon the best number of nodes (sub-sampling of the surface points) depending on the desired overall accuracy of the model: in our study, considering that the resolution of the original data was approximately $1mm^3$ isotropic per voxel, we set the accuracy per node at 2mm isotropic. Higher resolutions would simply being meaningless, since they would approach the limiting resolution of the original data.

Thus, each subject could be represented in three high dimensional feature spaces: a 3 (coordinates per node)x431-dimensional feature space for the left hippocampus, a 3x387-dimensional feature space for the right hippocampus, and a 3x(431+387)-dimensional feature space for both hippocampi.

2.4 Feature extraction

The goal of this study was to perform both volumetric and morphometrical analyses of the hippocampi, in order to assess the best set of features to automatically discriminate between healthy subjects and AD patients, and between MCI converters and MCI non-converters. Volumetric features could directly been obtained from the manual delineations of the hippocampi, and correcting them for intra-cranial volume: thus, for each subject one can easily evaluate the left, right, and total (left plus right) normalized hippocampal volumes (see Table 1).

Shape-based features (or markers) are more challenging to determine. In a previous work, we already presented a method to obtain consistent sets of markers by means of repeated permutation tests ([13]). In this work, we applied the same procedure to extract consistent shape-based markers to discriminate healthy subjects from AD.

Subsequently, these markers were used to train intelligent machines to automatically discriminate controls from AD, and MCI-converters from MCI-non converters.

2.4.1 Consistent shape-based markers for AD

The detection of markers was carried on as previously reported in [13]. The main goal is to detect locations of the hippocampal surface which consistently discriminate between healthy elderly and AD. A desired characteristic of any marker is that it should reflect differences between populations which are due to the investigated disease, rather than the particular datasets being investigated. Permutation tests, when applied to two populations, provide us with surface locations significantly different among the two groups: but how can we be sure that these differences are due to the disease? To answer this question, we performed repeated permutation tests, following the procedure reported in Fig. 3 (*left*). Starting from the complete set of 50 controls and 50 ADs, we created a *feature-extraction* set by randomly sub-sampling 25 controls and 25 AD. Marker selection was performed on this set, as follows:

1. For $N_{iter} = 30$ randomly sub-sample the feature-extraction set into $CTRL_1 = 13$ controls and $AD_1 = 13$ AD. Apply permutation tests to $CTRL_1$ and AD_1, and store the results (see Appendix for details on the permutation tests). This generates N_{iter} p-value maps: each node in a p-value map is given a p value associated with the null hypothesis that the two groups have a similar local shape at the location indicated by the node.

2. For each iteration, a node on the surface is considered significantly discriminative at a certain threshold P_{thr} if:

   - its $P \leq P_{thr}$, or
   - one of its neighbors, up to two links away, satisfies the previous condition in order to account small registration errors in the adaptation process.

3. Evaluate the consistency index for each node: i.e., the percentage out of the N_{iter} tests in which it was found to be significantly different.

4. For each node, evaluate the median p value through the N_{iter} iterations.

A node was considered consistent if its consistency index was either below 20% (the node is consistently not significantly different) or higher than 80% (the node is consistently significantly different). Nodes which are found significantly different less than 20% of the times, are thus *not* significantly different 80% of the times: one can think of these locations as areas which are not affected by the disease, regardless which sub-sampled set one considers. Conversely, consistent nodes with an index higher than 80% are locations affected by the disease, regardless which sub-sampled set one considers. Thus, within the group of consistent nodes, we finally define as markers those which are significantly different between controls and AD: these nodes are the features used for the subsequent training of the intelligent machines. Obviously, different P_{thr} threshold might lead to different feature sets: ideally, one aims at a high percentage of consistent nodes. In this work, we used four different values of P_{thr} (0.05, 0.01, 0.001, and 0.0001) to assess the best feature sets for the classifiers. Results are reported in Table 2, and presented in section 3.1.

2.5 Support vector machine for AD and MCI-converters detection

The final goal of our study is to build intelligent machines (i.e., classifiers) able to automatically differentiate AD subjects from healthy controls, and MCI-converters from MCI-non converters. The design of a classifier undertakes two important phases: a training phase, during which the classifier learns how to discriminate between subjects belonging to two classes, and a testing phase, in which the classifier is tested over previously unseen and independent testing subjects. The performances of a classifier can be assessed in terms of:

- *Separability* in the feature space: successful rate in classifying the objects in the training set.
- *Specificity and Sensitivity*: successful rate while classifying previously unseen healthy subjects and AD (or MCI non-converters and MCI converters). The average of these two measurements is referred to as *Generalization* or *overall accuracy*.

A good separability in the feature space is an indication of the discriminative performances of the selected features. Nevertheless, the goal is to obtain a classifier with high sensitivity and specificity when presented with previously unseen cases: hence, good generalization properties.

The training and testing of the *controls vs. AD* classifier was performed over an independent *training/testing* set made of those subjects (25 AD and 25 controls) not previously included in the *feature-extraction* test for shape-based markers: the independence of the training/testing set is essential to guarantee that the results are not biased by the previously detected features. The *training/testing* set for the MCI-converters vs. MCI-non converters was the entire set of 15 + 15 MCIs: this is acceptable since the MCIs were not involved in the feature extraction phase. Both for *controls vs. AD* and for *MCI-converters vs. MCI-non converters*, the training and testing procedure was based on the leave-1-out test: given a set of N subjects, the classifier is trained on $N-1$ subjects, and tested on the one subject left out. The entire procedure is then repeated, permuting over all the possible subjects to be left out. Overall generalization, specificity and sensitivity measures can be assessed.

The classifiers used in this study were Support Vector Machines (SVMs) based on radial basis functions (RBF) (e.g., [48, 39, 49]). To the scope of our work, it is sufficient to describe a RBF-SVM as a well-known pattern recognition tool for data classification, fully characterized by two parameters: the error cost for misclassifications C, and the kernel width γ. These parameters are tuned in order to minimize the classification error: given a set of n objects $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ and a set of corresponding labels $\{\mathbf{l_1}, \mathbf{l_2}, ..., \mathbf{l_n}\}$ (with $l_i \in [1, -1]$), the performances of a particular (C, γ) pair is evaluated with the leave-1-out test (see Fig. 3, *right*). By systematically searching over several values of C and γ, the best (C, γ) pair can be detected (for more details on the tuning, the reader is referred to the Appendix).

Support vector machines based on non-linear kernels, such has radial basis functions, can be subject to over-training: the classifier may become biased towards one of the two possible labels, and classify most of the subjects as belonging to one particular class. When this occurs, it might be useful to introduce linear SVMs, trying to reduce the over-training effects. Linear SVMs are fully characterized by only one parameter,

the error cost C: the tuning procedure is a simplified version of the one presented in the Appendix for the RBF-SVM (in the linear case, only C needs to be tuned).

3 Results

3.1 Hippocampal shape-based markers for AD

Markers selection was performed at different P_{thr} thresholds, both for the left and right hippocampal surface. The results are reported in Table 2: for the right hippocampus, the first three thresholds (0.05, 0.01, and 0.001) provided large sets of consistent nodes, of which respectively 98%, 66%, and 15% could be used as markers; the last threshold ($P_{thr} = 0.0001$) provided a very limited set of markers. The left hippocampus presented valid marker sets only at the first three thresholds, and no markers for the last one. Thus, for the subsequent analyses, we limit our investigation to the first three thresholds. In Fig. 4 we show p-values maps, for both the right and left hippocampi, reporting the local median p-value as evaluated from the $N_{iter} = 30$ repeated permutation tests. The color codes the corresponding threshold on the p values. When the strict threshold of 0.001 is applied, markers are found mostly in the frontal part of sector CA1 (color-coded in red), while sectors CA2 and CA3 are rather spared by atrophy: these areas have previously been reported in literature as highly characteristic for AD ([50]). The second threshold, 0.01, presents considerably larger areas of both left and right hippocampi as markers (more than half of the surface, color-coded in yellow). Finally, a threshold set a 0.05 shows almost the entire surface of both hippocampi as markers: this explains why relaxing the threshold even further (e.g., higher than 0.05) would not bring any benefits.

3.2 Automatic AD detection based on shape markers

Three different classifiers were trained to automatically discriminate AD from healthy subjects: one based on markers from the right hippocampus, one based on markers from the left hippocampus, and one based on combined markers from both hippocampi. In pattern recognition, it is not always obvious that adding features will automatically improve the performances of the classifier: indeed, combined features might confound

each other, causing the overall accuracy to decrease. Therefore, it is meaningful to test the combination of right and left hippocampi. Results are reported in Table 3: for the three scenarios, we report the overall accuracy, specificity and sensitivity obtained with the leave-1-out test over an independent (from feature extraction) testing set. The best performances were obtained with the markers of the left hippocampus, both at threshold 0.05 and 0.01, and with the combination of both left and right hippocampi at thresholds 0.01 and 0.001: the overall accuracy reached 90%, with specificity and sensitivity also set at 88% and 92% respectively. Markers on the right hippocampus could only reach an overall accuracy of 84% (specificity and sensitivity both set at 84%). The risk of overtraining is well-known in pattern recognition, especially when non-linear classifiers are used (as in this case, where radial basis functions are used): to rule out the possibility that overtraining is degrading the classifier performances, we performed the same analysis using a Linear-SVM, which is far less prone to overtraining. The best performances did not change, both for the left hippocampus at 0.05 and for the combination of both hippocampi at 0.001. The other cases, although not reported in the text, all presented slightly worse performances.

3.3 Automatic AD detection based on volume

The performances of volumetric features in discriminating AD from healthy subjects were assessed by building a RBF-SVM over the same training/testing set previously used for shape-based markers. Similarly, three classifiers were built, based on the either the right, left, or combined volumes. The results of the leave-1-out test are reported in Table 3. Clearly, the results obtained by volumetric features (best overall accuracy 74% for the combination of left and right hippocampal volumes) are inferior to those obtained by shape-based features. Moreover, the results provided by the RBF-SVM might indicate overtraining and bias towards the AD group (e.g. for the left hippocampal volume we have a sensitivity of only 52% and a specificity of 88%). When linear classifiers were used, results showed less bias towards AD, although the overall best accuracy did not improve (see Table 3).

3.4 Automatic discrimination of MCI converters and non-converters based on shape markers

An important question behind this study is whether those shape-based features which characterize AD with respect to healthy subjects might also be used to discriminate MCI-converters from MCI- non converters. The three thresholds (0.05, 0.01, and 0.001) represent respectively the entire hippocampal surface, half of the hippocampal surface, and well-specific areas of sector CA1. Following the same analysis presented in section 3.2, and using the features extracted from controls and AD, we trained and tested three RBF-SVMs for each threshold (right, left, and combination of both hippocampi). Results are reported in Table 4. Contrary to what previously seen for controls and AD, the best feature sets to discriminate MCI-converters from non-converters are those from the right hippocampus, particularly for thresholds 0.05 and 0.01: the overall accuracy reaches 80%, with both specificity and sensitivity set at 80% as well. The combination of both right and left hippocampi does not improve the performances. Finally, markers on the left hippocampus could only reach an overall accuracy of 77% at the threshold of 0.01. It is interesting to note how the lowest threshold, 0.001, being more specific for controls vs. AD, provides here the worst case scenario, both for the right-based, left-based, and combined classifiers. When using linear classifiers, the performances slightly degenerated: the best accuracy was still found for the right hippocampus (at threshold 0.05), but it only reached 73% (specificity/sensitivity of 67% and 80%). All the other cases presented worse performances.

The problem of over-determination was also taken into account, by repeatedly ($N = 100$) and randomly mixing the MCI-converters and MCI-non converters in two equal groups, and performing the training and testing again using the same sets of features. Results, reported in Table 5 for few key cases, show performances close to chance (50%), proving that the classifier is not affected by over-determination.

3.5 Automatic discrimination of MCI converters and non-converters based on volume

Similarly to what was done for controls and AD, we trained and tested both linear and non-linear(RBF) SVMs based on the volumetric distribution of MCI-converters and non converters. Results are reported in Table 4. The best overall accuracy was obtained with the RBF-SVMs trained on the combination for right and left hippocampal volume: specificity and sensitivity were respectively 80% and 73% respectively. Thus, also for MCIs volumetric information has less discriminative power than shape-based analyses.

4 Discussion

In this work, we have investigated the use of classifier, based on MR markers, to aid the diagnosis of AD and the detection of MCI at high risk of conversion. Both volumetric and morphological features of the hippocampi have been studied, proving that the latter are more accurate in discriminating AD and MCI converters. The results, following and improving on our previous work on MR ventricular markers ([13]), represent a further step towards computer-aided diagnostic tools for AD. Before compare our results with previously published studies, some methodological considerations are needed.

4.1 Methodological considerations

The methodology presented in this study is rooted in a pattern recognition framework. Support vector machine classifiers were implemented as a diagnostic tool to detect AD and MCI converters. Earlier studies usually adopted a different technique: a certain set of features (let them be anatomical, physiological, or of any other nature) would be detected in all subjects, and statistical tests would be applied to prove significant differences between the distribution of such features within each cohort. Such studies are useful to highlight changes in the brain which might be adopted as predictors of AD progression: volumetric changes in hippocampi and amygdalae ([15, 20, 23–25, 36, 37]), concentration of t-tau in CSF ([1, 2]), changes in cortical thickness ([9]), volumetric changes of deep gray-matter structures ([51]). However, these statistical analyses do not address the generalization capabilities of the findings: given that a sig-

nificant difference exists between hippocampal volume distribution in healthy elderly and ADs, how reliable is the hippocampal volume of a new patient when a diagnosis needs to be made? This question can only be answered within a pattern recognition framework. The independence of the testing dataset guarantees that the performances of the classifier are a statistically valid estimation of its success rate within a clinical set-up.

Within this context, several studies have recently been published using different kinds of classifiers: logistic regression and discriminative analysis ([26, 30]); multivariate analysis ([21]); support vector machines ([22, 27, 29]). We opted for SVMs, since they have been proved to provide excellent classification results ([27, 39, 48]). Moreover, different SVM's kernel (e.g., radial basis function, linear, etc.) directly implement different kinds of classifiers. The independence of training/testing from feature extraction was guaranteed in our work by using separate independent sets. [29] highlighted the importance of performing the feature extraction phase within the cross-validation loop: if feature extraction was performed on the same dataset subsequently used for testing, the results would be heavily biased and not reliable. In agreement with [29], we decided to further increase the independence between feature extraction and training, by splitting the feature extraction and training into two separate sets. The advantage of our approach is that only one final set of feature is selected and used for training and testing. The disadvantage is that less subjects are available for the training and testing of the controls vs. AD classifier. Thus, both for controls vs. AD and for MCI-c vs. MCI-nc we performed a leave-1-out cross validation test. However, [29] have already shown that leave-1-out tests are good representative of results obtained on larger sets.

A final methodological issue concerns the nature of the features to be extracted. Both physiological studies ([1–6]) and studies based on proton emission tomography ([7]) and molecular imaging ([8]) have shown promising results. However, these techniques are either highly invasive (CSF sampling) or rather expensive, and therefore not sufficiently widespread. Thus, several studies have looked for markers in MR structural images. Within this context, different approaches have been taken. It is well known that early stages of AD present atrophy in the hippocampi and entorhinal cor-

tex: subsequently, volumetric and morphological changes of the hippocampi have been investigated as potential markers for AD progression and prediction of MCI conversion ([11, 14, 20, 25–27, 37, 52, 53]). However, some criticisms have been risen: on one side, a fully automatic and validated method for hippocampal segmentation is still needed (although promising methods have been presented ([54]); thus, manual segmentation is often needed to delineate the hippocampi, introducing potential subjectivity in the studies. Moreover, it has been proven that hippocampal atrophy, although characteristic of AD, is also characteristic of other forms of dementia ([15, 36]). In order to overcome this limitation, some authors have suggested a more global approach, investigating changes in overall cortical thickness ([30]), or tissue density changes in both gray and white matter ([18, 19, 21, 22, 29]). Studying carefully the results of these investigations, a consistent pattern emerges. Baseline MCI-c present significantly reduced gray matter tissue, particularly in the hippocampi and parahippocampal cortex ([18]), whose atrophy increases at follow-up. In AD, hippocampal gyrus and limbic structures are found to be highly discriminative ([30]). The pattern of grey matter loss was also address by [19], showing that atrophy in MCI begins in the head of the hippocampi and in the entorhinal cortex, to later spread through the entire body of the hippocampi and medial temporal gyrus. Interestingly, [22] showed that classification performances between controls and AD could be improved considerably by constraining the search for makers to the medial temporal lobe; however, frontal areas were essential to discriminate AD from frontal temporal lobe dementia (FTDL). In conclusion, even studies based on global brain analysis point to regions in the medial temporal lobe as mostly discriminative for AD or MCI-c. What appears to be an intermediate solution is to combine features from several specific brain regions. [15] suggested that although hippocampal atrophy is present in both AD and FTDL, severe or asymmetrical amygdala atrophy would suggest FTDL; [55] presented a study based on Bayesian networks to model the relation between several brain regions, showing that atrophy of hippocampi and right thalamus correlates with the presence of MCI. Finally, volumetric and morphological ventricular changes might reflect atrophy in several periventricular structures ([56]). In previous works we had already applied shape modeling and pattern recognition to the study of ventricular shape changes in AD ([12, 13]): in this work,

we have focused on hippocampal shape changes, proving that diagnostic tools can be built both for AD and for MCI converters. Moreover, we have shown that classifiers based on volumetric information are less accurate. It is worth noting that the features used to train the MCI-c vs. MCI-nc classifiers were detected on controls vs. AD: this is justified by the fact that MCI-nc might always include subjects who will eventually develop AD and might therefore bias the analysis.

A final remark about over-determination is needed. The feature sets used throughout our study are quite large compared to the sample sizes available for training and testing. In such a situation, the good performances of the classification might not be due to appropriate feature extraction and training, but rather to a large set of features which could equally discriminate any two given groups of subjects (regardless their medical condition). To rule out this possibility, we have randomly and repeatedly mixed MCI-converters and MCI non converters in two equal groups, each time performing the training and testing again using the same sets of features. Classification results were proximal to chance (50%). Thus, it was not possible to build an accurate classifier to predict AD by simply having a large number of features: this ruled out the possibility of over-determination for our classifiers.

4.2 Comparison with related studies

The SVM created for the classification of AD and healthy elderly reached an overall accuracy of 90% (specificity and sensitivity of 88% and 92% respectively): these results were obtained for the left hippocampus (while using almost the entire surface) and for the combination of right and left hippocampi (while focusing on very specific locations, primarily the CA1 sector and subiculum). The overall accuracy drops slightly in the other cases, but remains higher than those found for volume-based classifiers. The anatomical findings for CA1 and subiculum are in agreement with previously reported studies ([11, 14, 25]). [26] also analyzed morphological and volumetrical changes of the hippocampi in AD: their results indicated atrophy in the CA1 areas and the discriminant analysis based on shape changes resulted in an overall accuracy of 84% (specificity and sensitivity of 85% and 84%). Adding volumetric information did

not improve the overall accuracy, but rather changed the specificity and sensitivity to 96% and 72%. [54] assessed hippocampal volumes in a semi-automatic way in both health elderly, ADs and MCIs: individual classification based on bootstrap methods provided an overall accuracy of 84% (same specificity and sensitivity) for controls vs. AD, and of 73% (specificity, 70%; sensitivity 75%) for controls vs. MCI. [22] reported a thorough analysis based on support vector machines for different scenarios: subjects with definitive AD (confirmed histopathologically by either cerebral biopsy or autopsy) could correctly been separated by matched healthy elderly with an overall accuracy of 96% (specificity, 93%; sensitivity, 100%); however, when dealing with probably AD (as compared to our study), the overall accuracy reached a maximum of 86% (specificity, 91%; sensitivity 76%). [29] trained three different classifiers: the first one (comparable to our study), based solely on MR-based information of tissue density, discriminated AD from healthy elderly with an overall accuracy of 86% (equal specificity and sensitivity); when additional information was added, the overall accuracy increased to approximately 89% both for the second classifier (based on MR information and age), and the third classifier (based on MR information, age, and genetic information). Finally, two recent studies showed accuracy higher than those obtained by our method. [27] trained a support vector machine to correctly classify controls and AD with an overall accuracy of 92%: the method, fully automatic, is based on MR information from a specific volume of interest set in the medial temporal lobe. However, it is not clear how this approach could be extended to other brain regions, in order to increase the specificity in discriminating different forms of dementia. [30] based their discriminant analysis on cortical thickness of several brain regions: the parahippocampal gyrus alone could be used to train a classifier with an overall accuracy of approximately 94%. Multi-variate analysis, including the parahippocampal gyrus, lead to full correct classification (100%). Our results outperform all the studies previously reported, except for the last two.

Differently from the studies mentioned above, we also applied the morphological and volumetric features, obtained with controls and AD, to the classification of MCI converters and MCI non-converters. From a clinical point of view, it is essential to discriminate between these cohorts, in order to detect patients with high risk of de-

veloping AD at early stages. The best accuracies were obtained for either the right hippocampus or for the combination of right and left (in both cases, overall accuracy of 80%). Interestingly, the right hippocampus appears to be more discriminative than the left one in separating MCI-c and MCI-nc; conversely, the left hippocampus lead to better performances for controls vs. AD. This is in agreement with previous findings: [16] showed that early onsets of AD present higher hippocampal grey-matter loss in the right hemisphere (16%) than in the left one (9%). Previous studies have addressed a similar problem, both using neurological tests, physiological information, or MR-based analyses. [57] showed that the degree of cognitive impairment, as assessed via the clinical dementia rating (CDR), correlates with the progression to AD within 5 years from the assessment; [58] reported a predictive accuracy in detecting MCI-c of 86% by using two neuropsychological tests; [59] showed a predictive accuracy of approximately 86% by using both neuropsychological tests and demographic information. MR-based techniques have also been presented. [60] reported a statistically significant correlation between elevated diffusivity of the left hippocampal and conversion to dementia; [61] investigated hippocampal and entorhinal cortex atrophy, showing that at a specificity of 80%, the sensitivity for the detection MCI-c was approximately 67% when using both hippocampal and entorhinal atrophy, and could be increased to 84% by adding demographic information; [62] applied deformation-based analysis to detect brain atrophy in MCI-c and MCI-nc. Principal component analysis was first used to detect salient features in controls vs. ADs: subsequently, the MCI groups were projected over this feature space and statistical analyses performed to detect separability of MCI-c and MCI-nc. Results showed an accuracy of 80% (based on CSF maps) and 73% (based on whole brain maps). [63] showed a statistically significant correlation between the baseline hippocampal volume and conversion to AD, suggesting that the left hippocampus has a better predictive power. Similar correlations were highlighted also in [64] and [65]. Most of the studies mentioned above were limited to statistical analysis, and therefore did not report accuracy in terms of specificity or sensitivity: their goal was to highlight the potential predictive power of specific brain characteristics: the lack of a pattern recognition framework allowed no further claims. The same holds true for those studies who reported specificity and sensitivity ([58, 59, 61, 62]): in all

these cases, the specificity and sensitivity were assessed as separability of the cohorts, without assessing the real generalization performances of the methods. Therefore, no direct comparison is possible with the results reported in our study: our accuracy of 80% represents a statistical estimation of how the classifier would perform in a clinical set-up. Only recently another study has presented results on automatic detection of MCI converters rooted in a pattern recognition framework [28]: intensity features and local volume estimation were obtained from a region of interest surrounding the hippocampi; feature extraction was performed on a cohort of healthy subjects and AD patients; subsequently, a classifier was trained to separate MCI-c converters from MCI-nc, reaching an overall accuracy of 81% (specificity and sensitivity of 100% and 70% respectively). These results are fully in line with ours, although based on a different approach, and support even further the idea that baseline MR features are important tools for early detection of Alzheimer disease.

4.3 Caveats

Few caveats related to our work should be mentioned. The manual delineation of the hippocampi can introduce subjectivity in a study, and can therefore cast doubts over the use of hippocampal markers. Other brain structures, such as brain ventricles, can easily be segmented automatically, due to the higher contrast between CSF and the rest of the parenchyma. In our work, we checked the test-retest reliability in order to reduce any possible bias in the manual delineation of the hippocampi. Moreover, the manual outlining is performed by experts blinded to the final diagnosis: therefore, the only bias could come from different MR contrast between groups. However, most of the structures surrounding the hippocampi (with the exception of the amygdalae) present a high contrast (e.g., ventricular CSF), considerably lowering the possibility of biased segmentation errors ([14]). A recent study has presented convincing steps towards an automatic segmentation of the hippocampi ([54]): thus, in the future more automatic methods could be used to delineate the hippocampal volume.

The hippocampal markers reported in our study should eventually be combined with morphological changes in other structures, in order to increase the specificity in discriminating AD from other forms of dementia. Future investigations should consider

covariant morphological changes in ventricles, hippocampi and deep grey-matter structures, and focus on more cohorts, including different kinds of dementia (e.g. FTLD). Moreover, the use of volumetric information, demographic data, and genetic information might increase the performances even further.

5 Conclusions

In conclusion, we have shown that morphological hippocampal shape-based markers for AD and MCI converters can be detected in 3D T1-weighted MR images. We have used these markers to train different classifiers to automatically discriminate AD from healthy elderly, and MCI converters from MCI non converters. Specificity and sensitivity for the classifiers were assessed within a pattern recognition framework, using independent sets, and are therefore statistically valid approximation of the performances to be expected in a clinical set-up. Furthermore, we have shown that morphological markers can reach higher performances than volumetric ones. The results obtained with our method are in line with (or better than) previously published results. Finally, we have reported some limitations of our approach and suggested ways to improve clinical diagnosis: future research should focus on incorporating morphological changes from different brain structures, and investigating the markers' discriminative performances on different kinds of dementia.

Acknowledgements

This work was supported by the Technology Foundation STW (project number LNN. 6122), and by Medis medical imaging systems bv, Leiden, The Netherlands (www.medis.nl).

Appendix

Permutation test

Permutation tests have been successfully used for the analysis of brain images [66, 11]: they require limited assumptions and are corrected for multiple comparison.

Given two populations G_1 and G_2, can we localize statistically significant differences on the average surface? Outcome of the permutation tests is, in the first place, a *p value* for the omnibus hypothesis "the two groups G_1 and G_2 are drawn from the same population". Moreover, we obtain the *p values* for each node in the model, telling us whether the distribution of that node in space is the same in G_1 and G_2 or not. Since all meshes are co-registered to a standard space, significant differences in space distribution of a certain location indicates either a significant enlargement or shrinking of one population with respect to the other (hence pointing to atrophy).

Permutation tests can be summarized as follows:

1. considering two groups G_1 and G_2:
   (a) for each node in the model, build up two clouds of points, C_1 and C_2, considering the positions the node assumes through all the shapes in G_1 and G_2;
   (b) C_1 and C_2 are compared via a Hotelling's T2 statistic test: outcome of the test is the *t value* for the node comparison (is the node distributed in space in significantly different ways?); such a test tests both the average positions in space for the two clouds, and their variances.
2. for $N_{perm} = 10000$ times, two groups of shapes A and B are built up by randomly mixing G_1 and G_2, and point 1 is performed on them. Only the highest *t value* is stored for each iteration;
3. a critical *t value* t_c is evaluated as the k^{th} highest value of all the N_{perm} *t values* previously stored (plus the t_{Max} for the original division in G_1-G_2), where

$$k = \lfloor \alpha * N_{perm} \rfloor + 1, \quad \alpha = 0.05; \tag{1}$$

4. the *p value* for the omnibus hypothesis "G_1 and G_2 are the same" is evaluated as:

$$p_{value} = \frac{N}{N_{tests}}, \quad where \tag{2}$$

$$N = \#\{stored \quad t_{values} | t_{value} > t_c\}, \tag{3}$$

$$N_{tests} = N_{perm} + 1; \tag{4}$$

5. finally, point 4 is applied to each single node, counting how many *t values* are higher than the *t value* associated with a particular node in the original G_1-G_2 grouping of shapes, and dividing the number for N_{perm}+1; this leads to a *p value* (corrected for multi-tests) for each node in the model.

Training of the SVM

A support vector machine based on radial basis function is characterized by two parameters: the error cost C and the kernel width γ. The procedure and values used for tuning these parameters are the same as suggested by Hsu et al.: *A Practical Guide to Support Vector Classification* (*citeseer.ist.psu.edu/689242.html*).

Given a training set T, a grid search is performed over the C x γ space:

1. $\forall$ (C,γ) $\in [2^{-5}, 2^{-3}, 2^{-1}, .., 2^{15}]x[2^{-15}, 2^{-13}, 2^{-11}, .., 2^{3}]$
   (a) assess the generalization capability of the SVM(C,γ) with leave-1-out over T;
   (b) store the success rate for (C, γ): $succRate_{(C,\gamma)}$;
2. Select the best $(\overline{C}, \overline{\gamma}) = argmax_{(C,\gamma)}$ $(succRate_{(C,\gamma)})$;
3. Perform a refined grid search (as in point 1) in the surrounding of $(\overline{C}, \overline{\gamma})$: e.g., if $\overline{C} = 2^{-3}$ and $\overline{\gamma} = 2^{-13}$, search within $[2^{-5}, 2^{-1}]$x$[2^{-15}, 2^{-11}]$ with steps equal to one tenth of the intervals;
4. Select the final best $(\overline{C}, \overline{\gamma})$;
5. Train the SVM($\overline{C}, \overline{\gamma}$) on the entire training set.

When a linear SVM is considered, only C needs to be tuned.

References

[1] Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., and Minthon, L. (2006) Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impariment: a follow-up study *Lancet Neurol* **5**, 228–234.

[2] Stefani, A., Martorana, A., Bernardini, S., Panella, M., Mercati, F., Orlacchio, A., and Pierantozzi, M. (2006) CSF markers in Alzheimer disease patients are not related to the different degree of cognitive impairment *J. Neurological Sicences* **251**, 124–128.

[3] Clark, C., Davatzikos, C., Borthakur, A., Newberg, A., Leight, S., Lee, V., and Trojanowski, J. (2008) Biomarkers for early detection of Alzheimer pathology *Neurosignal* **16(1)**, 11–18.

[4] Zetterberg, H., Pedersen, M., Lind, K., Svenson, M., Rolstad, S., Eckerström, C., Syversen, S., Mattsson, U., Ysander, C., Mattsson, N., Nordlund, A., Vandersichele, H., Vanmechelen, E., Jonsson, M., Edman, A., Blennow, K., and Walling, A. (2007) Intra-individual stability of CSF biomarkers for Alzheimer's disease over two years *J. Alzheimers Dis.* **12(3)**, 255–260.

[5] Ewers, M., Buerger, K., Teipel, S., Scheltens, P., Schröder, J., Zinkowski, R., Bouwman, F., Schönknecht, P., Schoonenboom, N., Andreasen, N., Wallin, A., DeBernardis, J., Kerkman, D., Heindl, B., Blennow, K., and Hampel, H. (2007) Multicenter assessment of CSF-phosphorylated tau for prediction of conversion of mci *Neurology* **69(24)**, 2205–2212.

[6] Leon, M., DeSanti, S., Zinkowski, R., Mehta, P., Pratico, D., Segal, S., Clark, C., Kerkman, D., Debernardis, J., Li, J., Lair, L., Reisberg, B., Tsui, W., and Ruisnek, H. (2004) MRI and CSF studies in the early diagnosis of Alzheimer's disease *Journal of Internal Medicine* **256**, 205–223.

[7] Herholz, K., Salmon, E., Perani, D., Baron, J., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., Zündorf, G., Delbeuck, X., Pelati, O., Anchisi, D., Fazio, F., Kerrouche, N., Desgranges, B., Eustache, F., Beuthien-Baumann, B., Menzel, C., Schröder, J., Kato, T., Arahata, Y., Henze, M., and Heiss, W. (2002) Discrimination between Alzheimer dementia and controls by automated analysis of multicenter fdg pet *NeuroImage* **17(1)**, 302–316.

[8] Frisoni, G. (2008) Interactive neuroimaging *Lancet Neurol.* **7(3)**, 204.

[9] Lerch, J., Priessmer, J., Zijdenbos, A., Hamperl, H., Teipel, S., and Evans, A. (2004) Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy *Cerebral Cortex* **15**, 995–1001.

[10] Convit, A., deAsis, J., deLeon, M., Tarshish, C., Santi, S. D., and Rusinek, H. (2000) Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease *Neurobiolg. Aging* **21(1)**, 19–26.

[11] Thompson, P., Hayashi, K., deZubicaray, G., Janke, A., Rose, S., Semple, J., Hong., M., Herman, D., Gravano, D., Doddrell, D., and Toga, A. (2004) Mapping hippocampal and ventricular change in Alzheimer disease *NeuroImage* **22(4)**, 1754–1766.

[12] Ferrarini, L., Palm, W., Olofsen, H., vanBuchem, M., Reiber, J., and Admiraal-Behloul, F. (2006) Shape differences of the brain ventricles in Alzheimer's disease *NeuroImage* **32(2)**, 1060–1069.

[13] Ferrarini, L., Palm, W., Olofsen, H., van derLanden, R., vanBuchem, M., Reiber, J., and Admiraal-Behloul, F. (2008) Ventricular shape biomarkers for Alzheimer's disease in clinical MR images *Magnetic Resonance in Medicine* **59(2)**, 260–267.

[14] Frisoni, G., Sabattoli, F., Lee, A., Dutton, R., Toga, A., and Thompson, P. (2006) In vivo neuropathology of the hippocampal formation in AD: a radial mapping MR-based study *NeuroImage* **32(1)**, 104–110.

[15] Barnes, J., Whitwell, J., Frost, C., Josephs, K., Rossor, M., and Fox, N. (2006) Measurements of the amygdala and hippocampus in pathological confirmed Alzheimer disease and frontotemporal lobar denegeneration *Arch Neurol* **63**, 1434–1439.

[16] Frisoni, G., Pievani, M., Testa, C., Sabattoli, F., Bresciani, L., Bonetti, M., Beltramello, A., Hayashi, K., Toga, A., and Thompson, P. (2007) The topography of grey matter involvment in early and late onset Alzheimer's disease *Brain* **130**, 720–730.

[17] Petersen, R., Doody, R., Kurz, A., Mohs, R., Morris, J., Rabins, P., Ritchie, K., Rossor, M., Thal, L., and Winblad, B. (2001) Current concepts in mild cognitive impairment *Arch. Neurol.* **58**, 1985–1992.

[18] Chételat, G., Landeau, B., Mezenge, F., Viader, F., de laSayette, V., Desgranges, B., and Baron, J. (2005) Using voxel-based morphometry to map structural changes associated with rapid conversion in mci: a longitudinal MRI study *NeuroImage* **27**, 934–946.

[19] Withwell, J., Przybelski, S., Weigand, S., Knopman, D., Boeve, B., Petersen, R., and Jr., C. J. (2007) 3d maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease *Brain* **130**, 1777–1786.

[20] Yavuz, B., Ariogul, S., Cankurtaran, M., Oguz, K., Halil, M., Dagli, N., and Cankurtaran, E. (2007) Hippocampal atrophy correlates with the severity of cognitive decline *International psychogeriatrics* **19(4)**, 767–777.

[21] Fan, Y., Batmanghelich, N., Clark, C., Davatzikos, C., and ADNI (2008) Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline *NeuroImage* **39**, 1731–1743.

[22] Klöppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., Rohrer, J., Fox, N., Jr, C. J., Ashburner, J., and Frackowiak, R. (2008) Automatic classification of MR scans in Alzheimer's disease *Brain* **131**, 681–689.

[23] Kodama, N., Shimada, T., and Fukumoto, I. (2002) Image-based diagnosis of Alzheimer-type dementia: Measurements of hippocampal and ventricular areas in MR images *Magn Reson Med Sci* **1(1)**, 14–20.

[24] Giesel, F., Hahn, H., Thomann, P., Widjaja, E., Wignall, E., vonTengg-Kobligk, H., Pantel, J., Griffiths, P., Peitgen, H., Schroder, J., and Essig, M. (2006) Temporal horn index and volume of medial temporal lobe atrophy using a new semi-automated method for rapid and precise assessment *AJNR* **27**, 1454–1458.

[25] Apostolova, L., Dutton, R., Dinov, I., Hayashi, K., Toga, A., Cummings, J., and Thompson, P. (2006) Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps *Arch Neurol* **63**, 693–699.

[26] Wang, L., Swank, J., Glick, I., Gado, M., Miller, M., Morris, J., and Csernansky, J. (2003) Changes in hippocampal volume and shape across time distinguish dementia of the Alzheimer type from healthy aging *NeuroImage* **20**, 667–682.

[27] Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G., and Collins, D. L. (2008) MRI-based automated computer classification of probable AD versus normal controls *IEEE Transactions on Medical Imaging* **27(4)**, 509–520.

[28] Duchesne, S., Bocti, C., Sousa, K. D., Frisoni, G., Chertkow, H., and Collinds, D. (2008) Amnestic mci future clinical status prediction using baseline MRI features *Neurobiology of Aging* **doi:10.1016/j.neurobiolaging.2008.09.003**, .

[29] Vemuri, P., Gunter, J., Senjem, M., Whitwell, J., Kantarci, K., Knopman, D., Boeve, B., Petersen, R., and Jr, C. J. (2008) Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies *NeuroImage* **39**, 1186–1197.

[30] Lerch, J., Pruessner, J., Zijdenbos, A., Collins, D., Teipel, S., Hampel, H., and Evans, A. (2008) Automated cortical thickness measurements from MRI can accurately separate Alzheimer's disease patients from normal elderly controls *Neurology of Aging* **29**, 23–30.

[31] Kantarci, K. and Jack, C. (2004) Quantitative magnetic resonance technique as surrogate markers of Alzheimer's disease *NeuroRX: The Journal of the American Society for Experimental NeuroTherapeutics* **1**, 196–205.

[32] Pruessner, J., Collins, D., Pruessner, M., and Evans, A. (2001) Age and gender poredict volume decline in the anterior and posterior hippocampus in early adulthood *The journal of neuroscience* **21(1)**, 194–200.

[33] Barns, J., Scahill, R., Boyes, R., Frost, C., Lewis, E., Rossor, C., Rossor, M., and Fox, N. (2004) Differentiating AD from aging using semiautomated measurement of hippocamapal atrophy rates *Neuroimage* **23**, 574–581.

[34] van dePol, L., Hensel, A., van derFlier, W., Visser, P., Pijnenburg, Y., Barkhof, F., Gertz, H., and Scheltens, P. (2006) Hippocampal atrophy on MRI in frontotemporal lobar degeneration and Alzheimer's disease *J Neurol Neurosurg Psychiatry* **77**, 439–442.

[35] Lehericy, S., Baulac, M., Chiras, J., Pierot, L., Martin, N., Pillon, B., Deweer, B., Dubois, B., and Marsault, C. (1994) Amygdalohippocampal MR volume measurements in the early stage of Alzheimer's disease *AJNR Am J Neuroradiol.* **15(5)**, 929–937.

[36] Barnes, J., Godbolt, A., Frost, C., Boyes, R., Jones, B., Scahill, R., Rossor, M., and Fox, N. (2007) Atrophy rates of the cingulate gyrus and hippocampus in AD and FTLD *Neurobiology of Aging* **28**, 20–28.

[37] Tarroun, A., Bonnefoy, M., Bouffard-Vercelli, J., Gedeon, C., Vallee, B., and Cotton, F. (2007) Could linear MRI measurements of hippocampus differentiate normal brain aging in elderly persons from Alzheimer disease? *Surg Radiol Anat* **29(1)**, 77–81.

[38] Frisoni, G., Ganzola, R., Canu, E., Rueb, U., Pizzini, F., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C., and Thompson, P. (2008) Mapping local structural hippocampal changes in Alzheimer's disease at 3 tesla *Brain* **131(12)**, 3266–3276.

[39] Vapnik, V. (1995) The nature of statistical learning theory *Springer-Verlag, New York*.

[40] Riello, R., Sabattoli, F., Beltramello, A., Bonetti, M., Bono, G., Falini, A., Magnani, G., Minonzio, G., Piovan, E., Alaimo, G., Ettori, M., Galluzzi, S., Locatelli, E., Noiszewska, M., Testa, C., and Frisoni, G. (2005) Brain volumes in healthy adults aged 40 years and over: a voxel-based morphometry study *Aging Clin Exp Res.* **17(4)**, 329–336.

[41] Folstein, M., Folstein, S., and McHugh, P. (1975) Mini-mental state *J Psych Res* **12**, 189–198.

[42] Holmes, C., Hoge, R., Collins, L., Woods, R., Toga, A., and Evans, A. (1998) Enhancement of MR images using registration for signal averaging *J Comput Assist Tomogr.* **22(2)**, 324–333.

[43] Pruessner, J., Serles, M., Pruessner, M., Collins, D., Kabani, N., Lupien, S., and Evans, A. (2000) Volumetry of hippocampus and amygdala with high-resolution MRI and three-dimensional analysis software: minimizing the discrepancies between laboratories *Cereb. Cortex* **10**, 433–442.

[44] Jenkinson, M. and Smith, S. (2001) A global optimisation method for robust affine registration of brain images *Medical Image Analysis* **5(2)**, 143–156.

[45] Jenkinson, M., Bannister, P., Brady, J., and Smith, S. (2002) Improved optimisation for the robust and accurate linear registration and motion correction of brain images *Neuroimage* **17(2)**, 825–841.

[46] Ferrarini, L., Olofsen, H., Palm, W., vanBuchem, M., Reiber, J., and Admiraal-Behloul, F. (2007) Games: Growing and adaptive meshes for fully automatic shape modeling and analysis *Medical Image Analysis* **11(3)**, 302–314.

[47] Kohonen, T. (1990) The self-organizing map *Proceedings of the IEEE* **78(9)**, 1464–1480.

[48] Boser, B., Guyon, I., and Vapnik, V. (1992) A training algorithm for optimal margin classifiers *In Fifth Annual Workshop on Computational Learning Theory - Pittsburgh, ACM* pp. 144–152.

[49] Burges, C. (1998) A tutorial on support vector machines for pattern recognition *Data Mining and Knowledge Discovery* **2**, 121–167.

[50] vanHoesen, G. and Hyman, B. (1990) Hippocampal formation: anatomy and the patterns of pathology in Alzheimer's disease *Prog. Brain Res.* **83**, 445–457.

[51] deJong, L., van derHiele, K., Veer, I., Houwing, J., Westendorp, R., Bollen, E., deBruin, P., Middelkoop, H., vanBuchem, M., and van derGrond, J. (2008) Strongly reduced volumes of putamen and thalamus in Alzheimers disease: an MRI study *Brain* **131(12)**, 3277–3285.

[52] Modrego, P. (2006) Predictors of conversion to dementia of probably Alzheimer type in patients with mild cognitive impairment *Curr. Alzheimer Res.* **3(2)**, 161–170.

[53] Miller, M., Priebe, C., Qiu, A., Fischi, B., Kolasny, A., Brown, T., Park, Y., Ratnanather, J., Busa, E., Jovicich, J., Yu, P., Dickerson, B., Buckner, R., and BIRN, M. (2008) Collaborative computational anatomy: an MRI morphometry study of the human brain via diffeomorphic metric mapping *Human Brain Mapping* **(in press: DOI 10.1002/hbm.20655)**, .

[54] Colliot, O., Chételat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., Dubois, B., Garnero, L., Eustache, F., and Lehéricy, S. (2008) Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus *Radiology* **248(1)**, 194–200.

[55] Chen, R. and Herskovits, E. (2006) Network analysis of mild cognitive impairment *NeuroImage* **29**, 1252–1259.

[56] Nestor, S., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J., Fogarty, J., Bartha, R., and ADNI (2008) Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging intiative database *Brain* **131**, 2443–2445.

[57] Dickerson, B., Sperlind, R., Hyman, B., Albert, M., and Blacker, D. (2007) Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment *Arch. Gen. Psychiatry* **64(12)**, 1443–1450.

[58] Tabert, M., Manly, J., Liu, X., Pelton, G., Rosenblum, S., Jacobs, M., Zamora, D., Goodkind, M., Bell, K., Stern, Y., and Devanand, D. (2006) Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment *Arch. Gen. Psychiatry* **63**, 916–924.

[59] Randall, H., Netson, K., Harrell, L., Zamrini, E., Brockington, J., and Marson, D. (2006) Amnestic mild cognitive impairment: diagnostic outcomes and clinical prediction over a two-year time period *J. Int. Neuropsychological Soc.* **12**, 166–175.

[60] Fellgiebel, A., Dellani, P., Greverus, D., Scheurich, A., Stoeter, P., and Müller, M. (2006) Predicting conversion to dementia in mild cognitive impairment by volumetric and diffusivity measurements of the hippocampus *Psychiatriy research neuroimaging* **146**, 283–287.

[61] Devanand, D., Pradhaban, G., Liu, X., Khandji, A., DeSanti, S., Segal, S., Rusinek, H., Pelton, G., Honig, L., Maueux, R., Stern, Y., Tabert, M., and deLeon, M. (2007) Hippocampal and antorhinal atrophy in mild cognitive impairment: prediction of Alzheimer disease *Neurology* **68**, 828–836.

[62] Teipel, S., Born, C., Ewers, M., Bokde, A., Reiser, M., Möller, H.-J., and Hampel, H. (2007) Multivariate deformation-based analysis of brain atrophy to predicty Alzheimer's disease in mild cognitive impairment *NeuroImage* **38**, 13–24.

[63] Eckerström, C., Olsson, E., Borga, M., Ekholm, S., Ribbelin, S., Rolstad, S., Starck, G., Edman, A., Wallin, A., and Malmgren, H. (2008) Small baselin volume of left hippocampus is associated with subsequent conversion of mci into dementia: the göteborg mci study *J. of Neurological Sciences* **272**, 48–59.

[64] Karas, G., Sluimer, J., Goedkoop, R., van derFlier, W., Rombouts, S., adnP. Scheltens, H. V., Fox, N., and Barkhof, F. (2008) Amnestic mild cognitive impairment: structural MR imageing findings predictive of conversion to Alzheimer disease *AJNR* **29**, 944–949.

[65] Smith, E., Egorova, S., Blacker, D., Killiany, R., Muzikansky, A., Dickerson, B., Tanzi, R., Albert, M., Greenberg, S., and Guttmann, C. (2008) Magnetic resonance imaging white matter hyperintensities and brain volume in the prediction of mild cogntive impairment and dementia *Arch. Neurol.* **65(1)**, 94–100.

[66] Nichols, T. and Holmes, A. (2001) Nonparametric permutation tetsts for functional neuroimaging: A primer with examples *Human Brain Mapping* **15**, 1–25.

Table 1: Socio-Demographic and clinical features of all subjects included in the study.

	Controls	AD	p	MCI conv.	MCI non-conv.	p
N.	50	50		15	15	
Age (years, $\mu(\sigma)$)	70.8 (5.8)	71.3 (7.7)	0.735	72.4(7.3)	71.4 (5.2)	0.656
N. females (%)	31 (62%)	35 (70%)	0.398	7 (47%)	7 (47%)	1.000
Education (years, $\mu(\sigma)$)	8.2 (4.1)	7.0 (3.5)	0.121	8.5 (4.6)	7.3 (4.3)	0.456
MMSE ($\mu(\sigma)$)	28.1 (1.4)	18.5 (3.4)	< 0.001	26.4 (1.5)	27.1 (1.0)	0.126
R. norm. hipp. vol. ($\mu(\sigma)$, mm^3)	2786(453)	2317(396)	< 0.001	2151(530)	2570(501)	0.03
L. norm. hipp. vol. ($\mu(\sigma)$, mm^3)	2669(421)	2225(451)	< 0.001	2074(533)	2361(475)	0.13

p denotes significance on t-tests for continuous variables, and on chi-square tests for dichotomous variables.

Table 2: Consistent nodes and Marker nodes. For four different P_{thr} thresholds, the % of consistent nodes and markers are reported.

	% of Consistent Nodes		% of Marker Nodes	
	Right	Left	Right	Left
$p \leq 0.05$	98%	99%	98%	99%
$p \leq 0.01$	66%	55%	66%	55%
$p \leq 0.001$	34%	36%	15%	8%
$p \leq 0.0001$	68%	79%	3%	-

Table 3: Performances in discriminating healthy subjects and AD: accuracy, specificity and sensitivity (based on leave-1-out on the testing set) for the three classifiers based on the right (R), left (L), or both (R+L) hippocampi. The first three rows show the results for RBF-SVMs trained on shape-based markers at three P_{thr} thresholds. The last two rows report the results for SVMs based on volumetric features of the same testing set, both for non-linear (RBF-based) and linear kernels.

	R	L	R+L
	Accuracy(%)	Accuracy(%)	Accuracy(%)
	(Spec./Sens.)	(Spec./Sens.)	(Spec./Sens.)
$p \leq 0.05$	84(84/84)	90(88/92)	86(88/84)
$p \leq 0.01$	84(84/84)	88(88/88)	88(88/88)
$p \leq 0.001$	84(88/80)	82(84/80)	90(88/92)
RBF-SVM on vol. feats.	72(76/68)	70(88/52)	74(76/72)
linear-SVM on vol. feats.	72(76/68)	66(72/60)	74(76/72)

Table 4: Performances in discriminating MCI-non converters and MCI-converters: accuracy, specificity and sensitivity (based on leave-1-out on the testing set) for the three classifiers based on the right (R), left (L), or both (R+L) hippocampi. The first three rows show the results for RBF-SVMs trained on shape-based markers at three P_{thr} thresholds. The last two rows report the results for SVMs based on volumetric features of the same testing set, both for non-linear (RBF-based) and linear kernels.

	R	L	R+L
	Accuracy(%)	Accuracy(%)	Accuracy(%)
	(Spec./Sens.)	(Spec./Sens.)	(Spec./Sens.)
$p \leq 0.05$	80(80/80)	73(93/53)	80(80/80)
$p \leq 0.01$	80(80/80)	77(93/60)	80(87/73)
$p \leq 0.001$	77(80/73)	73(87/60)	77(67/87)
RBF-SVM on vol. feats.	73(80/68)	63(73/53)	77(80/73)
linear-SVM on vol. feats.	70(73/67)	60(47/73)	73(80/67)

Table 5: Tests for over-determination. For $N = 100$ times, two groups were randomly created mixing MCI-converters and MCI non-converters: training and testing (leave-1-out) were performed with RBF-SVMs for some key scenearios. Results show accuracy, specificity, and sensitivity as average (standard deviation).

		R		L	
$p \leq$	0.05	0.01	0.001	0.01	0.001
Accuracy(%) $\mu(\sigma)$	54(10)	55(11)	56(11)	56(11)	57(10)
Spec.(%) $\mu(\sigma)$	57(16)	53(17)	54(15)	57(16)	57(17)
Sens.(%) $\mu(\sigma)$	51(17)	57(19)	59(16)	55(16)	56(15)

Fig. 1: GAME applied to a synthetic shape. (*top*) Growing phase: a mesh is grown by adding nodes and edges until it converges to the desired shape. (*bottom*) Adaptive phase: the topology (nodes and edges) of the previously created mesh is frozen, and the mesh is adapted to a new instance.

Fig. 2: (*left*) MR of the healthy subject chosen as representative. (*middle*) Close-up of the MR images, with superimposed the hippocampal segmentation. (*right*) Surface reconstruction of the right hippocampus.

Fig. 3: (*left*) Marker Selection. Repeated permutation tests were applied to randomly sub-sampled instances of the training set. After N_{iter} runs, a consistency index was evaluated for each node, providing the $\%$ of runs in which the node was found to be significantly different. Median p values of the runs were also assessed. (*right*) Design of the SVM. The same training set used for feature selection is used to tune and train a RBF-SVM. After training, specificity and sensitivity are tested on a completely independent testing set.

Fig. 4: P-value maps. Color-coded map showing, for each location on the hippocampal surfaces, the *median* p value evaluated over the N_{iter} runs.

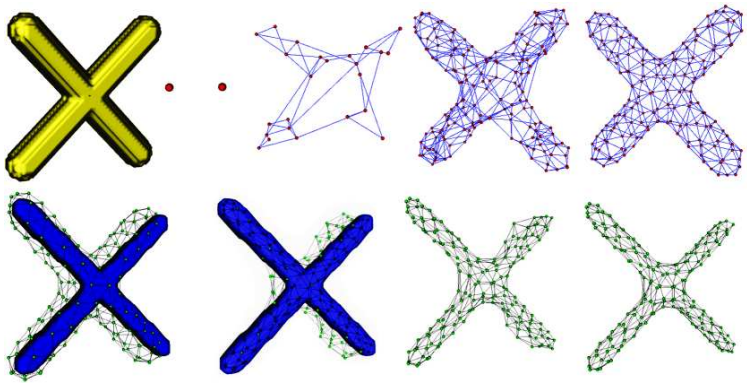

Figure 1:

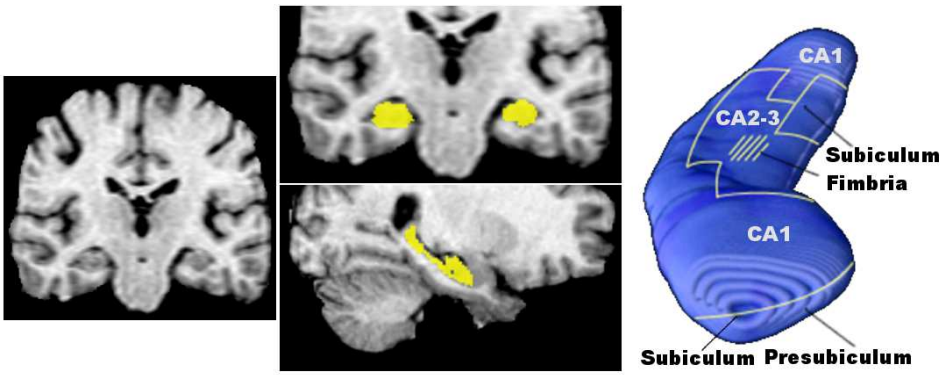

Figure 2:

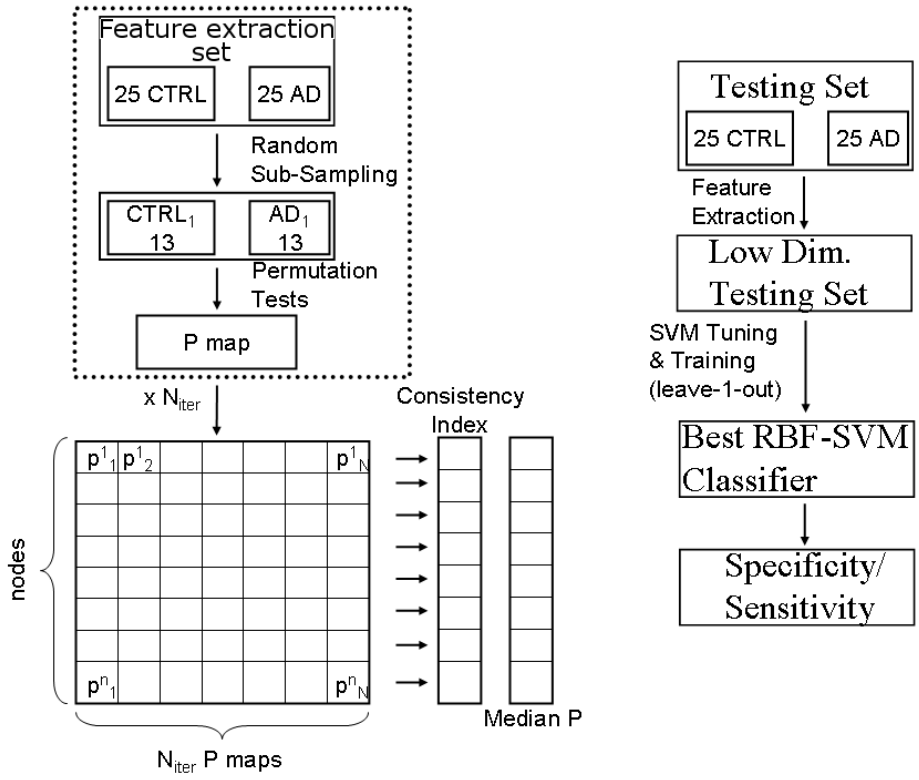
Feature extraction set
25 CTRL
25 AD
Random Sub-Sampling
CTRL1 13
AD1 13
Permutation Tests
P map
x Niter
Consistency Index
nodes
Median P
Niter P maps
Testing Set
25 CTRL
25 AD
Feature Extraction
Low Dim. Testing Set
SVM Tuning & Training (leave-1-out)
Best RBF-SVM Classifier
Specificity/ Sensitivity

Figure 3:

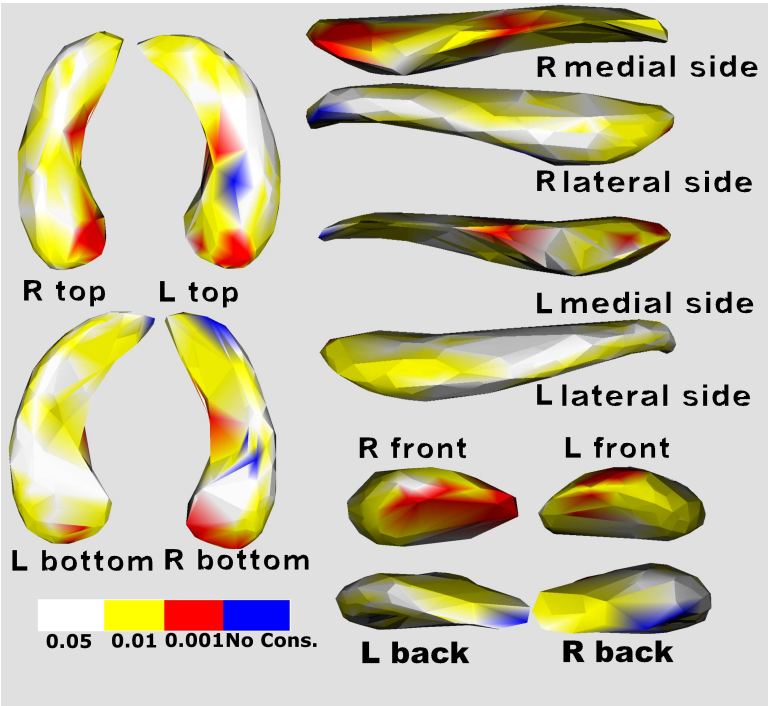
R medial side
R lateral side
L medial side
L lateral side
R top
L top
L bottom
R bottom
R front
L front
L back
R back
0.05
0.01
0.001
No Cons.

Figure 4:

Answers to the Reviewers – Second Revision

We would like to thank both Reviewers for their useful suggestions and positive feedbacks. We have considered the last points risen by the Reviewers and improved the manuscript accordingly to them. Changes in the manuscript are highlighted in **bold** to better guide the Reviewers through the reviewing process. Moreover, each point is answered in this letter, preceded by quoted versions of the Reviewers' comments.

Reviewer 1

1 –
"I thank you for your well-detailed response to my previous comments. I agree with most of your answers, however before accepting for publication, over-determination is still an issue, and I would like the following to be added/commented upon in the manuscript. It is necessary to point it out in the limitations, at the very least. It remains that a space of 276+ dimensions is massively over-determined to solve a problem of at most (25+25) variables (e.g. CTRL vs AD). A simple experiment with 276 random variables would probably yield results that are not too far from those presented. It would be better to: a) merge the 3 coordinates into a distance feature; and b) increasing the threshold for significance, in order to limit the number of features. Ideally, for the purposes of this manuscript, one could run a simple series of experiments (e.g. 100) whereby subjects are randomly assigned to groups A and B and then the methodology is used to attempt to separate the two groups. If correct, then the results should be equal to chance. This is not too expensive to perform, and would go a long way in increasing the validity of the technique."

Following the suggestion of the Reviewer, we have performed 100 tests by randomly mixing the MCI-converters and MCI non-converters in two groups. Then, RBF-SVMs were used, with the same feature sets used for the original groups, to train and test new classifiers. The results were next to chance level (50%), proving that a large dataset is *per se* not sufficient to discriminate any two given groups. We have added the results in section 3.4, discussed them at the end of section 4.1, and reported them in a new Table (Table 5).

Reviewer 2

1 –
The authors have considered and correctly answered the two reviewers' comments. The mscr. can be accepted for publication as it stands. Ref. 38 can be completed in proof (Brain 2008; 131: 3266-76).

All references have now been updated, with either their full information or with the DOI number when still ahead of print.