How early can we predict Alzheimer’s disease using computational anatomy?

Stanislaw Adasiewski a,b,1, Juergen Dukart a,b,*,1, Ferath Kherifa, Richard Frackowiak a, Bogdan Dragnaski a,c, for the Alzheimer’s Disease Neuroimaging Initiative2

*Département des Neurosciences Cliniques, Laboratoire de Recherche en Neuroimagerie, Centre Hospitalier Universitaire Vaudois, Université de Lausanne, Lausanne, Switzerland
1Department of Neurology, Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland
2Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

A R T I C L E   I N F O
Article history:
Received 2 April 2013
Received in revised form 24 May 2013
Accepted 20 June 2013
Available online 26 July 2013

Keywords:
Structural magnetic resonance imaging
Alzheimer’s disease
Mild cognitive impairment
Biomarker

A B S T R A C T
Computational anatomy with magnetic resonance imaging (MRI) is well established as a noninvasive biomarker of Alzheimer’s disease (AD); however, there is less certainty about its dependency on the staging of AD. We use classical group analyses and automated machine learning classification of standard structural MRI scans to investigate AD diagnostic accuracy from the preclinical phase to clinical dementia. Longitudinal data from the Alzheimer’s Disease Neuroimaging Initiative were stratified into 4 groups according to the clinical status—(1) AD patients; (2) mild cognitive impairment (MCI) converters; (3) MCI nonconverters; and (4) healthy controls—submitted to a support vector machine. The obtained classifier was significantly above the chance level (62%) for detecting AD already 4 years before conversion from MCI. Voxel-based univariate tests confirmed the plausibility of our findings detecting a distributed network of hippocampal-temporoparietal atrophy in AD patients. We also identified a subgroup of control subjects with brain structure and cognitive changes highly similar to those observed in AD. Our results indicate that computational anatomy can detect AD substantially earlier than suggested by current models. The demonstrated differential spatial pattern of atrophy between correctly and incorrectly classified AD patients challenges the assumption of a uniform pathophysiological process underlying clinically identified AD.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction
Recent advances in computer-based diagnosis making use of structural magnetic resonance imaging (sMRI) and machine learning methods provide evidence of sufficient accuracy in discriminating Alzheimer’s disease (AD) patients not only from healthy controls but also from other common types of dementia (Davatzikos et al., 2008b; Dukart et al., 2011a, 2012; Fan et al., 2008; Koppel et al., 2008a, 2008b). For the clinically and neuroscientifically pertinent case of early AD detection, support vector machine (SVM) classification and other machine learning studies taping into the preclinical phase of AD convincingly demonstrate the potential for reliable early diagnosis (Casanova and Hsu, 2012; Davatzikos et al., 2008a; Devanand et al., 2007; McEvoy et al., 2009; Misra et al., 2009; Modrego, 2006). Two limitations applying to most of the previous studies are the tuning of the classifiers to specific cohorts and with respect to cross-validation. Both restrict their generalizability to the general population. Tuning of a classifier to achieve a high accuracy for detection of mild cognitive impairment (MCI) patients might result in a substantial drop in accuracy when applying the same classifier to AD patients. Similarly, the tuning of a classifier to achieve high cross-validation accuracies might substantially increase the risk of overfitting the classification model to the particular dataset used in the study therewith providing an overoptimistic estimation for the accuracy of the method when applied to a general population.

Despite the progress in the field of computer-based AD detection, our knowledge about the capability of sMRI for early diagnosis even before the first manifestation of clinical signs is still very limited. The most recent model of brain anatomy—derived biomarker in AD (Jack et al., 2010) suggested a protracted progression of atrophy compared with functional changes as observed by [F18]fluorodeoxyglucose positron emission tomography. In contrast, other prospective studies provided evidence that sMRI measurements may contain information of ongoing disease-related process already before clinical...
manifestation of cognitive decline (Dickerson and Wolk, 2012; Quiroz et al., 2012; Smith et al., 2008). However, these studies did not test the predictive power of sMRI information to detect AD in untested subjects or cross-validation. Therefore, there is a pressing need to investigate the timescale of disease-related structural brain changes in AD not only to advance our understanding of the disorder but also to provide better tools for early diagnosis when neuro-protection is possible. Another related aspect is that the application of multivariate pattern classification techniques for early AD detection produces a high proportion of erroneous predictions for conversion from MCI to AD (Ewers et al., 2010; Misra et al., 2009). Thus, the secondary aim of our study is to investigate if false predictions are because of random noise or deterministic atrophy pattern.

We systematically address the questions of timescale of disease detection and potential causes of erroneous prediction while aiming to overcome the aforementioned limitations. We first adopt a pragmatic strategy testing whether AD-related atrophy is already detectable several years before conversion followed by in-depth investigation of atrophy patterns comparing incorrectly and correctly diagnosed AD, MCI converting to AD during the follow-up (MCI converters [cMCI]), MCI nonconverters (ncMCI), and healthy control subjects. To this end, we apply classical mass-univariate voxel-based analysis paralleled by machine learning classification using SVMs.

2. Methods

2.1. Subjects

To evaluate temporal sensitivity of sMRI data for early detection of AD, we used 1.5-T T1-weighted images from the Alzheimer’s Disease Neuroimaging Initiative (ADNI, http://www.adni-info.org/) of all available AD, cMCI, and ncMCI patients and healthy controls who had baseline and at least 2 years of follow-up MRI scans.

The AD patient and control subject data were split into a dataset used for SVM classifier training and another for diagnosis (Tables 1 and 2). Critically, the cMCI (Table 3) and ncMCI (Table 2) data were used for diagnosis only. Baseline and follow-up scans after 6, 12, 24, 36, 48, and 60 months, if available, were downloaded from the ADNI database along with the corresponding clinical information. All the data available in ADNI1 and ADNI-GO studies were used for subsequent evaluation. The diagnosis of AD was based on NINCDS/ADRDA (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s Disease and Related Disorders Association) criteria (McKhan et al., 1984). Exclusion criteria were the presence of any significant neurologic disease other than AD, history of head trauma followed by persistent neurologic deficits or structural brain abnormalities, psychotic features, agitation or behavioral problems within the previous 3 months, or history of alcohol or substance abuse. The study was conducted according to the Declaration of Helsinki. Written informed consent was obtained from all participants before protocol-specific procedures were performed.

The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether sMRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments, monitor their effectiveness, and lessen the time and cost of clinical trials. The Principal Investigator of this initiative

### Table 1
Subject group characteristics

<table>
<thead>
<tr>
<th>Training set</th>
<th>Testing set</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD Control subjects</td>
<td>AD Control subjects</td>
<td>ncMCI cMCI</td>
</tr>
<tr>
<td>n</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>74.4 ± 8.3</td>
<td>75.8 ± 6.9</td>
</tr>
<tr>
<td>Gender (M/F)</td>
<td>31/23</td>
<td>25/29</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>23.1 ± 1.9</td>
<td>23.4 ± 2.1</td>
</tr>
<tr>
<td>Follow-up time (y) (mean ± SD)</td>
<td>2.1 ± 0.2</td>
<td>2.8 ± 0.4</td>
</tr>
</tbody>
</table>

Key: AD, Alzheimer’s disease; ANOVA, analysis of variance; cMCI, mild cognitive impairment (converters); df, degree of freedom; F, female; M, male; MMSE, mini-mental state examination; ncMCI, mild cognitive impairment (nonconverters); SD, standard deviation.

### Table 2
Follow-up testing group characteristics

<table>
<thead>
<tr>
<th>Control subjects</th>
<th>Baseline</th>
<th>1 y</th>
<th>2 y</th>
<th>3 y</th>
<th>4 y</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD n</td>
<td>42</td>
<td>42</td>
<td>37</td>
<td>38</td>
<td>14</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>76.5 ± 4.7</td>
<td>77.1 ± 4.7</td>
<td>77.7 ± 5.1</td>
<td>79.2 ± 4.4</td>
<td>81.8 ± 4.4</td>
</tr>
<tr>
<td>Gender (M/F)</td>
<td>18/24</td>
<td>18/24</td>
<td>16/21</td>
<td>17/21</td>
<td>8/6</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>29.2 ± 0.8</td>
<td>29.3 ± 1.0</td>
<td>29.4 ± 0.9</td>
<td>29.1 ± 1.3</td>
<td>29.5 ± 1.1</td>
</tr>
<tr>
<td>AD n</td>
<td>54</td>
<td>53</td>
<td>54</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>76.7 ± 7.0</td>
<td>77.1 ± 6.7</td>
<td>78.8 ± 7.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gender (M/F)</td>
<td>25/29</td>
<td>25/28</td>
<td>25/29</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>23.4 ± 2.1</td>
<td>22.4 ± 3.7</td>
<td>19.2 ± 5.8</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ncMCI n</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>44</td>
<td>—</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>75.1 ± 7.7</td>
<td>75.7 ± 7.7</td>
<td>76.8 ± 7.6</td>
<td>79.0 ± 7.4</td>
<td>—</td>
</tr>
<tr>
<td>Gender (M/F)</td>
<td>45/16</td>
<td>45/16</td>
<td>45/16</td>
<td>34/10</td>
<td>—</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>27.5 ± 1.9</td>
<td>27.4 ± 2.3</td>
<td>27.0 ± 3.1</td>
<td>27.3 ± 2.1</td>
<td>—</td>
</tr>
</tbody>
</table>

Key: AD, Alzheimer’s disease; cMCI, mild cognitive impairment (converters); F, female; M, male; MMSE, mini-mental state examination; ncMCI, mild cognitive impairment (nonconverters); SD, standard deviation.
is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from 50 sites across the United States and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55–90, to participate in the research, 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, see http://www.adni-info.org/.

2.4. SVM classification

For SVM classification, we used the freely available LIBSVM software (Chang and Lin, 2001). The data of AD patients and healthy controls were split randomly into a dataset used for training of the SVM classifier (n = 54/54, AD/healthy controls) and a separate dataset used for classification-based diagnosis (Tables 1 and 2). The training dataset was used to build a linear soft-margin SVM classifier using voxel-wise whole-brain GM information without any further parameter optimization and default cost value of 1 (Kloppel et al., 2008b). The independent diagnostic set included data of AD, cMCI, and ncMCI patients and healthy controls (Tables 2 and 3).

To obtain the precision of classification accuracy for the diagnostic cohorts, we apply bootstrapping with 100 permutations using whole-brain information from a randomly selected 2/3 of the training set without replacement and keeping the original proportion of AD and controls (36 AD/36 healthy subjects) from the whole training dataset to build a classifier to predict test data of healthy controls and cMCI, ncMCI, and AD patients at each data acquisition time point. Aiming at an unbiased and precise estimation of classification/prediction accuracy, we used bootstrapping, which was shown in the literature to provide a more accurate estimation of the classification accuracy compared with other approaches such as leave-one-out or other cross-validation modes (Dougherty et al., 2010; Jain et al., 1987).

The results are presented using 2 different timescales—for healthy controls, AD, and ncMCI, the zero point on time axis signifies the baseline scan, whereas for cMCI, it signifies the time point of conversion (time to conversion [TTC]). The cMCI data were ordered on the timescale relative to the TTC as this ordering indicates that all subjects reached a predefined level of cognitive decline. For ncMCI, classification performance was evaluated as assignment to the healthy control group, whereas for cMCI, correct classification was considered as assignment to the AD group. To enable longitudinal estimates of accuracy in the same group of cMCI patients, all classifications were repeated for a subset (n = 17) with imaging data for each of 3 years up to time point of conversion.

In an additional step, we performed feature selection within the training dataset using weights provided by the classifier for each voxel as a measurement of its contribution to the separation between AD patients and healthy controls (weights close to zero indicate a low contribution). The absolute values of the weights (n = 41) randomly selected from the ADNI dataset and not used for further analyses. The obtained model parameters are then applied to the sMRI data accounting for any variance attributable to healthy aging in all subjects and patients.
were first sorted from lowest to highest. In a second step, we sequentially removed features with lowest weights as provided by the whole-brain classification. Similar to optimization of SVM parameters to a particular training dataset, iterative feature selection procedures often carry the risk of overfitting a classifier to the training dataset. This might be the case when the weights are redetermined after each selection step, and the newly obtained weights are used to guide further feature selection steps. This procedure leads to establishment of numerous feature maps each having different weights assigned to the same features. Selection of the best of them without accounting for the total number of calculated weight maps may therefore lead to overoptimistic accuracy estimations. To avoid overfitting the classifier by multiple testing, the distribution of weights was calculated only once, using the classifier trained on whole-brain information in the training cohort. The computed maximum weight was set as 100% (with a weight of zero corresponding to 0%). Subsequently, all features with weights <5% of the maximum were removed from classification. A classifier based on weights exceeding this threshold was used on all cMCI data at each time point. This procedure was repeated in 5% steps for the whole range of weights by consecutively removing features with lowest weights in the initial whole-brain classification, training a classifier on the remaining features, and reclassifying the cMCI data.

All classification accuracies with and without feature selection were compared with each other, between consecutive time points, and with chance-level distributions (obtained from the same data by randomly shuffling diagnostic labels within the training dataset) by using 2-sample t tests with a significance threshold of 0.05 (1 tailed, Bonferroni corrected for multiple comparisons).

2.5. Voxel-based morphometry

We performed analyses of covariance in the test cohort VBM data comparing correctly and incorrectly classified AD, cMCI, and ncMCI patients to each other and to healthy controls. For each patient, we used the first scan misclassified using whole-brain information in more than 50% of all permutations for VBM comparisons. As group sizes of correctly and incorrectly classified subjects differed within each clinical group, a random sample was drawn from the bigger group to match the smaller group in size. Age, gender, and total intracranial volume were included as covariates for all comparisons. A statistical threshold of p < 0.001 uncorrected at voxel level and an extent threshold of 200 voxels at cluster level were applied for all statistical analyses. We used the automated anatomic labeling atlas (Tzourio-Mazoyer et al., 2002) implemented in the Matlab-based WFU PickAtlas 2.3 toolbox.
(Maldjian et al., 2003, 2004) to assign imaging results to anatomic labels.

2.6. Statistical analysis of behavioral and demographic data

We performed analyses of variance (ANOVAs) to compare all testing and training groups at baseline in terms of age and mini-mental state examination (MMSE, Folstein et al., 1975). For all ANOVAs that revealed significant between-group differences, we conducted post hoc Bonferroni t tests with a significance threshold of $p < 0.05$ for pairwise group comparisons. Group differences regarding sex were evaluated using a chi-square test for independent samples. The statistical analyses were performed with the software package SPSS 17.0 (http://www.spss.com/statistics/).

To investigate potential differences in the development of cognitive profiles (as measured by MMSE) between misclassified and correctly classified subjects (as classified by SVM) within each clinical group, we calculated repeated-measures ANOVAs with factor time as a within-group factor and classification label (correct vs. misclassified) as a between-group factor. Only time points for which MMSE was available for a sufficient number of subjects in both correctly and misclassified groups (number of subjects per condition >5) were used for these analyses. If an ANOVA revealed significant group or group-by-time differences, post hoc t tests were performed to compare the groups at each time point. As we expected that subjects misclassified as AD in both the control and ncMCI groups would show greater cognitive deficits and vice versa subjects classified as controls would express fewer cognitive deficits, a 1-sided significance threshold of $p < 0.05$ was applied for these analyses.

3. Results

3.1. Behavioral and demographic results

Groups used for testing and training did not differ in terms of age (Table 1). As expected, we observed significant differences in
MMSE scores between all groups. Gender differed significantly between AD and both MCI groups and between controls and both MCI groups but not between AD and controls and not between the MCI groups. Post hoc t-tests revealed significant differences in MMSE between healthy controls and other clinical groups for both the test (healthy controls vs. AD: t(106) = 19.51, p < 0.05; healthy controls vs. cMCI: t(182) = 9.41, p < 0.05; healthy controls vs. ncMCI: t(113) = 5.87, p < 0.05) and training cohorts (healthy controls vs. AD: t(106) = 2.81, p < 0.05) and from AD in the test cohort (cMCI vs. AD: t(194) = 11.40, p < 0.05; ncMCI vs. AD: t(113) = 10.97, p < 0.05). No significant differences in MMSE were observed between test and training sets for AD patients (AD training vs. AD testing: t(106) = 0.73, p > 0.99) and for healthy controls (t(94) = 1.41, p > 0.99). The overall comparison revealed a significant difference in gender between groups (χ²(t) = 16.80, p = 0.005). However, no significant differences were observed for gender between AD patients and healthy controls in the training (χ²(t) = 0.150, p = 0.699) and testing cohorts (χ²(t) = 0.113, p = 0.737). Further, no differences were observed between cMCI and ncMCI (χ²(t) = 1.343, p = 0.246). Both cMCI and ncMCI differed significantly in gender distribution from healthy controls (cMCI vs. healthy controls: χ²(t) = 6.94, p < 0.05; ncMCI vs. healthy controls: χ²(t) = 10.01, p < 0.05) and AD patients (cMCI vs. AD: χ²(t) = 6.02, p < 0.05; ncMCI vs. AD: χ²(t) = 9.08, p < 0.05).

The ANOVAs investigating differences in MMSE between correctly and incorrectly classified subjects revealed significant differences in all clinical groups. Thus, both in the control (F(1,140) = 6.41, p = 0.049) and ncMCI (F(1,59) = 4.7, p = 0.035) groups, subjects classified as controls had higher MMSE values compared with those who were classified as AD. The opposite pattern was observed in cMCI (F(1,158) = 6.0, p = 0.015) and AD (F(1,151) = 5.0, p = 0.03), with subjects correctly classified as AD showing significantly lower MMSE values. Group-by-time interactions failed to reach the significance threshold in any group though showed a trend (p < 0.1) in control and cMCI groups. However, in the post hoc t tests, differences between correctly and incorrectly classified subjects in control and ncMCI groups were not significant at baseline but became significant on follow-up examinations (Fig. 1). In the AD and cMCI groups, differences between correctly and incorrectly classified subjects were significant at all time points except for time points 2 and 5 for cMCI.

3.2. SVM classification

The results of the feature selection procedure are displayed in Fig. 2. All results described as with feature selection are referring to the feature set providing the maximum classification accuracy in this feature selection procedure.

Whole-brain SVM classification yielded average (over all time points) diagnostic accuracies of 80.3%, 73.5%, and 63.7% for healthy controls, AD, and cMCI, respectively. In the ncMCI group, 69.0% have been classified as control subjects. At a TTC of 4 and 3 years, cMCI patients were classified at chance level (Fig. 3). At a TTC of 2 years, the accuracy increased above the chance level and kept increasing annually. In all other groups, accuracy was significantly above the chance level at all time points (Figs. 3 and 4).

Feature selection resulted in a feature set consisting of 814 voxels, corresponding to 0.21% of the initial brain volume, in bilateral hippocampus, amygdala, precentral gyrus, middle frontal gyrus, left inferior temporal gyrus, crus I, supramarginal gyrus, right calcarine sulcus, caudate nucleus, and cerebellum VI (Fig. 2C). Average classification accuracy over all time points using this feature set was significantly increased compared with classification without feature selection—82.0%, 76.8%, and 67.2% for healthy controls (t(344) = 3.21, p < 0.05), AD (t(320) = 7.23, p < 0.05), and cMCI (t(1068) = 7.36, p < 0.05), respectively. Mean assignment for ncMCI patients as control subjects significantly decreased to 60.6% (t(452) = 23.64, p < 0.05). With feature selection, the accuracy for cMCI at 62% was higher than chance level at a TTC of 4 years and was mostly stable at consecutive time points (Fig. 3). The classification accuracy for the 17 cMCI patients in the feature selection set was similar to that in the whole cMCI group.
After feature selection, diagnostic accuracy for AD significantly increased at baseline \((t(106) = 8.41, p < 0.05)\) and 1-year follow-up \((t(104) = 5.61, p < 0.05)\) compared with classification without feature selection [Fig. 4]. For ncMCI, the percentage of assignment to the control group significantly decreased at all time points—baseline \((t(120) = 14.96, p < 0.05)\), 1-year follow-up \((t(120) = 12.86, p < 0.05)\), 2-year follow-up \((t(120) = 12.00, p < 0.05)\), and 3-year follow-up \((t(86) = 10.20, p < 0.05)\). For healthy controls, accuracy significantly increased only at 3-year follow-up \((t(74) = 3.26, p < 0.05)\). For CMCI, significant increases in accuracy were observed at TTC of 4 years \((t(32) = 4.39, p < 0.05)\), 3 years \((t(74) = 8.43, p < 0.05)\), 2 years \((t(164) = 9.99, p < 0.05)\), 1 year \((t(254) = 8.96, p < 0.05)\), and at time point of conversion \((t(256) = 3.24, p < 0.05)\). For the CMCI subgroup consisting of the same subjects at all time points, accuracy increased at all time points (TTC of 3 years: \(t(32) = 4.92, p < 0.05\); TTC of 2 years: \(t(32) = 6.60, p < 0.05\); TTC of 1 year: \(t(32) = 4.16, p < 0.05\); time point of conversion: \(t(32) = 3.22, p < 0.05\)).

3.3. Voxel-based morphometry

The comparison of incorrectly classified CMCI to healthy controls revealed atrophy restricted to bilateral hippocampus, amygdala, ventral putamen, left thalamus, and right parahippocampal gyrus (Fig. 5A, Table 4). Correctly classified cMCI showed extensive atrophy extending to bilateral hippocampus, fusiform gyrus, superior, middle, and inferior temporal gyri, inferior parietal lobule, supramarginal gyrus, parahippocampal gyrus, insula, angular gyrus, amygdala, and putamen. Left-sided atrophy was observed in the postcentral gyrus. A direct comparison of correctly and incorrectly classified CMCI revealed significant atrophy in correctly classified CMCI in a temporoparietal-hippocampal network (Fig. 5B, Table 5). The opposite contrast revealed atrophy in misclassified cmCI in the cerebellum bilaterally.

In cmCI classified as AD, we observed significant atrophy compared with healthy controls in bilateral hippocampus, amygdala, and parahippocampal gyrus and also atrophy restricted to the left hemisphere in fusiform gyrus and middle and inferior temporal gyri. cMCI classified as controls showed no differences relative to healthy controls. A direct comparison in the ncMCI cohort revealed atrophy in ncMCI classified as AD in temporal and hippocampal regions. In the opposite contrast, we found calcarine sulcal atrophy in ncMCI classified as controls.

In the comparison of correctly classified AD to healthy controls, we observed significant atrophy in AD in bilateral hippocampus, parahippocampal gyrus, fusiform gyrus, superior and middle temporal gyri, supramarginal gyrus, superior and inferior parietal lobules, angular gyrus, and amygdala. Only left-hemispheric decreases were observed in the postcentral gyrus and ventral putamen. Right-hemispheric changes were restricted to the inferior temporal gyrus. Correctly classified AD showed greater atrophy compared with misclassified AD in the right inferior temporal and left lingual gyrus. Misclassified AD showed atrophy in the right gyrus rectus only.

The comparison of correctly and incorrectly classified control subjects revealed an AD-typical pattern of atrophy in misclassified control subjects in a temporoparietal-hippocampal network. No other significant differences were observed.

4. Discussion

Here, we investigate the longitudinal changes in predictive value of computer-based diagnosis of AD. We detect systematic differences in classification accuracy of CMCI that were dependent to closeness to conversion time. Whole-brain SVM classification 3–4 years before conversion provides chance-level accuracy for AD detection. After feature selection restricting the analysis to structures of the medial temporal lobe and parietal cortex, we achieve accuracies significantly higher than chance level already 4 years before conversion. Our findings not only provide a novel perspective on the time- and state-dependent changes of diagnostic accuracy within the preconversion period from MCI to AD but also bring new evidence for potentially differential pathophysiological processes underlying neurodegeneration in clinically defined AD.

Our findings bring strong evidence that sMRI-based information can be used to predict conversion from CMCI to AD at disease stages earlier than previously suggested [Jack et al., 2010]. On the one side, this demonstrates that already at early disease stages, sMRI data contain information on AD-related pathologic changes, which can be used for prediction of conversion from MCI to AD. On the other side, the relatively low accuracy of 62% obtained using an unbiased classifier, which was not optimized for MCI subjects or cross-validation, also suggests that interindividual variability in the sMRI measurements might be the impeding factor in our study preventing high diagnostic accuracies at this early disease stage. Using optimized preprocessing and classification methods and integration of longitudinal trajectories into the algorithms might provide remedy.

Additionally, we find strong evidence that AD-related structural changes can be observed before the occurrence of any significant...
cognitive deficit. This finding is in line with recent studies demonstrating an increased atrophy rate in cognitively normal subjects with high beta-amyloid deposition and greater atrophy in the right medial temporal lobe in cognitively normal subjects with future cognitive impairment (Chételat et al., 2012; Tondelli et al., 2012). Control subjects and ncMCI classified as AD demonstrate a pattern of structural GM abnormalities in a temporoparietal-hippocampal network that is very similar to the one previously reported in AD patients (Devanand et al., 2007; Frisoni et al., 2002; Schroeter et al., 2009). Corroborating these structural observations, we find statistically significant yet subclinical cognitive deficits in control and ncMCI subjects classified as AD compared with those classified as controls. These differences increase over time without reaching the significance threshold of a formal interaction analysis. However, the observation of AD-typical structural changes combined with a decline in cognitive performance in control subjects and in ncMCI patients strongly supports the conjecture that AD-related pathology can be detectable at much earlier time points than proposed in current models of AD progression (Jack et al., 2010). The finding of AD-like atrophy patterns linked with faster cognitive decline in control subjects and ncMCI patients classified as AD is in line with previous studies performed by Davatzikos et al. (2009, 2011), who demonstrated faster cognitive decline in control and ncMCI subjects with an AD-like atrophy pattern.

We further demonstrate that detection accuracy of conversion from MCI to AD significantly increases with closeness to the time point of conversion. Interestingly, classification after feature selection provides accuracies higher than chance level 4 years before conversion opposed to classification of whole-brain data reaching similar level of accuracy at earliest at the time point of clinical conversion. The localization and spatial extent of the anatomic feature set are in line with previous research showing the precedence of pathologic changes in hippocampus and parietal cortex and relative sparing of cerebellum (Braak and Braak, 1991; Jack et al., 2010; Schroeter et al., 2009). These studies have shown that beta-amyloid and tau pathologies start focally in AD and spread with time to other areas of GM whereas the further accumulation of histopathologic changes in already affected regions slows with time. Atrophy is assumed to follow this spreading pattern of beta-amyloid and tau depositions with a delay of several years. Based on these findings, one would expect that atrophy is restricted to a few brain regions in earlier AD stages. Other parts of the brain are not expected to contain relevant information for disease detection at this early stage. Given that the standard SVM assigns a weight to all—including noninformative—features, the inclusion of a large number of features (as performed in a whole-brain approach) may decrease the signal-to-noise ratio and therewith lower the possible classification accuracy. Feature selection at this early AD stage will remove noisy features and so increase classification accuracy. In contrast, at later AD stages, atrophy involves large parts of cortex and so all brain regions carry some information that are relevant for discrimination between AD patients and healthy control subjects. Performing feature selection at this stage will lead to removal of informative features and may decrease classification accuracy. Both effects are observed in our study: the gain achieved in cMCI using feature selection rapidly decreases with closeness to the conversion time point. Correspondingly, feature selection in manifest AD and also in healthy control subjects adds little, giving a similar accuracy to that with a whole-brain approach. This observation is consistent with the literature evaluating the effect of feature selection on
classification accuracy (Chu et al., 2012) and studies reporting similar accuracies for differentiation between AD patients and healthy controls using sMRI with and without feature selection (Dukart et al., 2011a; Gerardin et al., 2009; Hinrichs et al., 2009; Koppel et al., 2008a, 2008b). Our result therefore emphasizes the importance of feature selection procedures at the MCI stage of early incipient AD using sMRI.

Another finding is that there is a significant reduction in assignment of ncMCI to healthy controls using feature selection. In general, all MCI patients from the ADNI cohort used in our study correspond to the amnestic MCI type. Previous studies with a follow-up of up to 10 years have shown that at longer follow-up, most of the amnestic MCI convert to AD (Fischer et al., 2007; Visser et al., 2006). It is therefore likely that a 4-year follow-up is insufficient to detect all cases of premanifest AD in the group. Given the strong boost in accuracy we observe after feature selection at the MCI stage of early incipient AD, we would expect the classifier trained on selected features to be much more sensitive to potential AD also in the ncMCI cohort. Supportive of this reasoning is our VBM result comparing ncMCI classified as AD with healthy controls. This reveals a GM atrophy pattern that is very similar to that observed in correctly classified AD and cMCI. An evaluation of assignment of ncMCI patients with a short follow-up as controls is therefore highly problematic in terms of interpretation.

We further detect substantial differences between correctly and incorrectly classified control subjects and MCI and AD patients using the VBM approach. Both correctly classified AD and cMCI show extensive atrophy in a temporoparietal-hippocampal network, whereas only minor changes restricted to the hippocampus are detected in misclassified patients. The finding of a temporoparietal-hippocampal atrophic network is consistent with a similar set of regions reported for cMCI and AD in the literature (Karow et al., 2010; Schroeter et al., 2009). However, cumulative evidence shows that current diagnostic selection criteria do not discriminate accurately between AD and other types of dementia (Dubois et al., 2007; Varma et al., 1999) and that hippocampal atrophy is also found in other dementia syndromes such as frontotemporal lobar degeneration (Barnes et al., 2006, 2007; van de Pol et al., 2006). Taking both findings into account questions the reliability of selection criteria used in the ADNI cohort and of suggested diagnostic algorithms that use only hippocampal atrophy as a biomarker for AD (Devand et al., 2007; Laakso et al., 1995; Morra et al., 2009a, 2009b). Indeed, the diagnostic accuracy we achieve for AD of 80.3% is compatible with results published in pathologic studies of patients dying with a clinical diagnosis of AD and differ from the SVM-based classification accuracy of ~95% when scans from pathologically verified cases of AD were used (Koppel et al., 2008b). Given very recent pathologic-clinical diagnostic results, this is clearly an important area for further research (Beach et al., 2012).

As previously shown, sMRI can be used to predict conversion from MCI to AD (Devand et al., 2007; McEvoy et al., 2009; Plant et al., 2010) and even to distinguish between cMCI and ncMCI with reasonable accuracy (Ewers et al., 2010; Hinrichs et al., 2011). However, in these studies, cMCI patients were pooled over all time points and so obscuring the influence of TTC on diagnostic accuracy. Additionally, disease-related processes have been shown to influence the estimation of healthy aging and vice versa—healthy aging has been shown to interact with disease detection (Dukart et al., 2011b, Franke et al., 2010). For these reasons, we removed the voxel-wise variance explained by healthy aging from all imaging data based on estimates from an independent healthy cohort (Dukart et al., 2011b). All observed changes in accuracy estimates over time can therefore be attributed to disease-related pathologic processes.

The average accuracies obtained in our study for detection of AD and healthy controls and of conversion from MCI to AD are comparable with those reported in the literature. However, most previous studies applied feature selection algorithms optimizing the cross-validation performance within their own datasets. These types of algorithms carry the risk of overestimating the conversion from MCI to AD detection accuracy because of multiple testing and

<table>
<thead>
<tr>
<th>Region</th>
<th>Side</th>
<th>MNI coordinates</th>
<th>Cluster size</th>
<th>Peak T value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amygdala, fusiform gyrus, hippocampus, parahippocampal gyrus, putamen, middle and superior temporal poles</td>
<td>L</td>
<td>−24 −5 −16</td>
<td>2099</td>
<td>5.65</td>
</tr>
<tr>
<td>Amygdala, fusiform gyrus, hippocampus, parahippocampal gyrus, middle and superior temporal poles</td>
<td>R</td>
<td>28 −8 −18</td>
<td>1611</td>
<td>5.30</td>
</tr>
<tr>
<td>Angular gyrus, supramarginal gyrus, middle temporal gyrus, superior temporal gyrus</td>
<td>L</td>
<td>−58 −54 9</td>
<td>3437</td>
<td>5.01</td>
</tr>
<tr>
<td>Middle occipital and middle temporal gyrus</td>
<td>R</td>
<td>40 −74 16</td>
<td>715</td>
<td>4.61</td>
</tr>
<tr>
<td>AD misclassified &lt; controls</td>
<td>Putamen</td>
<td>R</td>
<td>31 3 −6</td>
<td>270</td>
</tr>
<tr>
<td>Amygdala, fusiform gyrus, hippocampus, insula, middle occipital, olfactory cortex, parahippocampal gyrus</td>
<td>L</td>
<td>−26 −8 18</td>
<td>4486</td>
<td>6.52</td>
</tr>
<tr>
<td>Amygdala, hippocampus, parahippocampal gyrus</td>
<td>R</td>
<td>27 −8 18</td>
<td>2793</td>
<td>6.12</td>
</tr>
<tr>
<td>Angular gyrus, supramarginal gyrus</td>
<td>L</td>
<td>−52 −59 29</td>
<td>3537</td>
<td>5.00</td>
</tr>
<tr>
<td>Middle temporal gyrus</td>
<td>L</td>
<td>−54 −12 −21</td>
<td>940</td>
<td>4.21</td>
</tr>
<tr>
<td>Middle occipital gyrus</td>
<td>R</td>
<td>44 −77 18</td>
<td>109</td>
<td>3.92</td>
</tr>
<tr>
<td>Fusiform gyrus, inferior temporal gyrus</td>
<td>R</td>
<td>42 −17 −31</td>
<td>364</td>
<td>3.87</td>
</tr>
<tr>
<td>Superior temporal pole</td>
<td>R</td>
<td>54 −9 15</td>
<td>383</td>
<td>3.60</td>
</tr>
<tr>
<td>cMCI misclassified &lt; controls</td>
<td>Amygdala, hippocampus, parahippocampal gyrus</td>
<td>R</td>
<td>28 −11 −18</td>
<td>701</td>
</tr>
<tr>
<td>Thalamus</td>
<td>L</td>
<td>−12 −30 0</td>
<td>332</td>
<td>4.00</td>
</tr>
<tr>
<td>Amygdala, hippocampus</td>
<td>L</td>
<td>−27 −11 −10</td>
<td>293</td>
<td>3.86</td>
</tr>
<tr>
<td>ncMCI misclassified &lt; controls</td>
<td>Inferior temporal gyrus, middle temporal gyrus</td>
<td>R</td>
<td>39 −9 −19</td>
<td>602</td>
</tr>
<tr>
<td>Amygdala, hippocampus, parahippocampal gyrus</td>
<td>L</td>
<td>−44 1 −45</td>
<td>1140</td>
<td>4.28</td>
</tr>
<tr>
<td>Amygdala, fusiform gyrus, hippocampus, parahippocampal gyrus</td>
<td>R</td>
<td>28 −12 −18</td>
<td>1016</td>
<td>4.09</td>
</tr>
</tbody>
</table>

Key: AD, Alzheimer’s disease; cMCI, mild cognitive impairment (converters); L, left; MNI, Montreal Neurological Institute; ncMCI, mild cognitive impairment (nonconverters); R, right.
potential overfitting of selected features to a specific dataset. The main goal of our study was not the further development and optimization of current classification algorithms but rather to provide evidence that sMRI-based classification is able to detect AD substantially before conversion and to characterize the evolution of detection accuracy. We therefore avoided in our selection and classification procedures most potential sources of overfitting by using independent training and test datasets and by avoiding parameter optimization to improve the classification accuracy of test data. Similarly, previous studies indicated that training cohort is an important factor influencing achievable accuracy (Abdulkadir et al., 2011; Casanova and Hsu, 2012; Chu et al., 2012). We would therefore expect that the accuracy we obtained for classification of all clinical groups could be further increased by using larger training datasets.

Another related issue is the use of only AD patients in the training dataset. It is likely that higher accuracies could have been achieved for early AD detection if only cMCI subjects had been used as a training dataset. However, from our point of view, it is very important for clinical practice to establish a single classifier that will be capable to detect AD irrespective of stage. We therefore avoided training a separate classifier for application only on MCI subjects. Nonetheless, a classifier built on already manifested AD is also likely not to be the best for AD detection in all its stages. The issue of establishing an optimum classifier with maximum accuracy for early AD detection should be addressed in future research, for example, by integrating both cMCI and AD patients in the same training dataset to allow SVM or other techniques to detect a stage-independent AD pattern.

A potential source of limitation in our study that also needs consideration is that all the data used for training and testing, although treated independently, were extracted from the same cohort (ADNI). Any potential selection bias inherent to the ADNI cohort could therefore limit the generalizability of our results to other sources of patients.

In summary, our study provides first evidence that sMRI might contain information that can be used for detection of AD already at early disease stages—4 years before conversion from MCI. We further demonstrate that control subjects and nMCI that are classified as AD using automated computer-based diagnostics show preclinical cognitive deficits align with an atrophy pattern that is characteristic for AD. Both findings are supportive for the capability of automated machine learning algorithms based on sMRI data to detect a possible incipient AD.

Disclosure statement
The authors have no conflicts of interest relevant to the subject of the manuscript, including no institutional contracts relating to this research or any other agreements of the authors or their institutions that could be seen as involving a financial interest in this work.

Acknowledgements
S.A. is supported by the Scientific Exchange Program Scien-X-NMS-ch (Scientific Exchange Programme between the New Member States and Switzerland). J.D. is supported by the Swiss National Science Foundation (National Centres of Competence in Research [NCCR] Synapsy). B.D. is supported by the Swiss National Science Foundation (Project Grant Nr 320030_135679, NCCR Synapsy, and SPUM 33CM30_140332/1). Foundation Parkinson Switzerland, Foundation


S. Adaszewski et al. / Neurobiology of Aging 34 (2013) 2815–2826