Bioprofile Analysis: A New Approach for the Analysis of Biomedical Data in Alzheimer's Disease

Javier Escudero[a,*], Emmanuel Ifeachor[a], John P. Zajicek[b] and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]*Signal Processing and Multimedia Communications Research Group, School of Computing and Mathematics, Plymouth University, Plymouth, UK*
[b]*Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Plymouth University, Derriford, UK*

Abstract. This article presents a new approach for the analysis of biomedical data to support the management and care of patients with Alzheimer's disease (AD). The increase in prevalence of neurodegenerative disorders such as AD has led to a need for objective means to assist clinicians with the analysis and interpretation of complex biomedical data. To this end, we propose a "Bioprofile" analysis to reveal the pattern of disease in the subject's biodata. From the Bioprofile, personal "Bioindices" that indicate how closely a subject's data resemble the pattern of AD can be derived. We used an unsupervised technique (k-means) to cluster variables of the ADNI database so that subjects are divisible into those with the Bioprofile of AD and those without it. Results revealed that there is an "AD pattern" in the biodata of most AD and mild cognitive impairment (MCI) patients and some controls. This pattern agrees with a recent hypothetical model of AD evolution. We also assessed how the Bioindices changed with time and we found that the Bioprofile was associated with the risk of progressing from MCI to AD. Hence, the Bioprofile analysis is a promising methodology that may potentially provide a complementary new way of interpreting biomedical data. Furthermore, the concept of the Bioprofile could be extended to other neurodegenerative diseases.

Keywords: Alzheimer's disease, Bioindex, biomarkers, Bioprofile, machine learning, mild cognitive impairment, unsupervised techniques

Supplementary data available online: http://dx.doi.org/10.3233/JAD-2012-121024

---

*Correspondence to: Javier Escudero, School of Computing and Mathematics, Plymouth University, Drake Circus, Plymouth, PL4 8AA, UK. Tel.: +44 1752 586295; E-mail: javier.escudero@ieee.org.

INTRODUCTION

As a result of the rise in life expectancy, the number of people suffering from long-term conditions is increasing and this places a huge burden on national healthcare systems. Thus, there is a need for healthcare provision to evolve in order to meet the challenges of improved patient management. Interpretation of the vast amounts of biomedical data that are now becoming

available to support early detection and monitoring of disease can be assisted by novel methods of analysis. In this context, we hypothesize that the healthcare of subjects with neurodegenerative disorders would benefit from the creation of a "Bioprofile" and "Bioindices" of disease from the patients' data.

We define a Bioprofile as an objective 'fingerprint' or 'pattern' in the data that provides information about a complex underlying disease that evolves with time. It represents the pattern of changes that the disease imposes on the variables. Moreover, an individual's data patterns can be compared against the Bioprofile of the disease. This is reflected by the Bioindex, a measure of how closely the subject's data patterns resemble such Bioprofile of disease. When it is feasible to acquire periodic measures of clinical variables for a subject, this information can be used to update the subject's Bioindices and, over time, this could provide an indication of disease progression and response to treatment. The long-term vision is to create a set of Bioprofiles for various conditions and different data modalities so that individual patient data can be compared against a set of Bioprofiles to provide a holistic view of his or her health status and aid in the early detection of disease. This idea is particularly appealing in complex conditions that evolve over a long period of time [1], such as neurodegenerative diseases.

Alzheimer's disease (AD) is the most common neurodegenerative disease [2]. In 2010, there were 35.6 million people with dementia in the world and the number of patients is expected to double every 20 years [2]. AD pathology starts years before the first symptoms appear, with quality of life already affected by the time any clinical diagnosis is made [3]. Thus, there is a need for new objective means to help clinicians in the analysis of AD-related data, particularly at the mild cognitive impairment (MCI) stage, to allow targeted interventions and to monitor disease progression and response to treatment [3].

Several biomarkers have been proposed in a number of studies [4–15] to overcome the difficulties of traditional neuropsychological scores in the characterization of AD [16]. In the context of Bioprofile analysis, biomarkers are relevant neuroimaging or biochemical features with potential information about the disease. A hypothetical model of biomarker dynamics in AD has recently been introduced [17–20]. This model may help to contextualize changes using some common biomarkers of AD as the disease progresses. For example, it suggests that the level of the 42-amino acid amyloid-β (Aβ_{42}) protein becomes abnormal first

[10] and that AD would later result in decreased brain glucose metabolism [4, 8, 9]. Afterwards, tau protein levels [10] would alter and AD would also cause brain atrophy [5–7]. Finally, the biological changes would lead to cognitive and other impairments which inform clinical diagnosis [17, 18].

Machine learning provides useful ways of analyzing data. These techniques can be categorized into supervised and unsupervised methodologies [21]. Whereas the supervised methods require ground-truth labels to learn which data patterns correspond to the predefined classes, the unsupervised techniques are able to reveal patterns in the data without the need for ground-truth references [21]. Some studies have used supervised methodologies to find 'signatures' or 'fingerprints' that represent AD features in a number of subjects diagnosed with the disease [10, 22–27], but unsupervised techniques can also be relevant in this setting. Unsupervised machine learning provides simple and easy-to-interpret algorithms for data analysis and it offers complementary perspectives to those of supervised classifiers [21, 28]. Unsupervised machine learning methods are data-driven and tend to be less subject to over-fitting [21, 28]. Thus, to develop the Bioprofile of AD, we selected unsupervised machine learning. We hypothesized that it may be well suited to reveal the pathological process of AD that takes place before clinical symptoms appear [29–31]. Cognitive normal (CN), MCI, and AD subjects have recently been analyzed with clustering-related techniques [29–32]. One study assessed if cerebrospinal fluid (CSF) biomarkers reflect the AD pathology in the three groups of subjects without using diagnosis labels [30]. Other authors clustered a mixed set of data from CN people to reveal that this is not a completely homogeneous subject group [29]. In another study, MCI subjects were considered unlabeled cases and a semi-supervised classifier was applied to the magnetic resonance imaging (MRI) data [32]. Hence, unsupervised data-driven processing based on machine learning may provide useful insights into AD.

In this article, our goal is to investigate if a Bioprofile of AD can be revealed from biomedical data and to present the Bioprofile methodology in tasks related to the assessment of progression from MCI to AD. This study, which extends previous preliminary results [31], makes the following contributions:

1) Introduction of the Bioprofile of AD as a new way to analyze biomedical data and to study the disease.

2) Provision of evidence of the link between the Bioprofile and a recent hypothetical model of AD evolution.

3) Illustration of the potential use of the Bioprofile and Bioindices to assess the risk of developing AD at the MCI stage.

4) Assessment of the change in the Bioindices with time for MCI subjects.

MATERIALS AND METHODS

ADNI database

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu/). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment 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 and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative 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 over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 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/.

Selection of variables

The ADNI data used in this study are as of May 3, 2011. The database was queried for basic demographic, clinical, CSF, MRI, and 2-fluorodeoxy-D-glucose (FDG)-PET data of CN, MCI, and AD subjects. The retrieved biomarkers and clinical scores were used to construct six scenarios. The first five correspond to the modalities considered in the hypothetical model of AD evolution [17] (1: CSF $A\beta_{42}$, 2: FDG-PET, 3: MRI, 4: CSF tau, and 5: neuropsychological scores). These five scenarios are located at different times in the evolution of the disease [17]. We also considered a sixth scenario composed of the data from the subjects for whom all variables are simultaneously available at baseline. Further details about the variables included in each scenario are available in the Supplementary Data (available online: http://www.j-alz.com/ issues/32/vol32-4.html#supplementarydata04).

The subjects' age, gender, years of education, and the number of apolipoprotein E (ApoE) $\varepsilon4$ alleles were retrieved. The follow-up diagnoses of all MCI subjects were also obtained from ADNI. The diagnosis is only used for validation purposes and not to drive the clustering-based Bioprofile analysis.

Concept of cluster-based Bioprofile analysis for AD

The Bioprofile is a general concept that hypothesizes:

1) that long-term conditions impose a characteristic pattern on the biomarkers and clinical variables;

2) that such pattern can be revealed with machine learning methods (either supervised or unsupervised, depending on the nature of the condition) so that appropriate indices can be computed to reflect changes due to the disease.

In the case of AD, the abnormalities reflected in most biomarkers follow a sigmoid function over a relatively long period of time [19, 20]. Such changes start even before the manifestation of clinical symptoms [3, 16–18, 29, 30]. The gold standard for diagnosis of AD is only available postmortem, but we hypothesize that unsupervised techniques would allow the estimation of Bioprofiles of AD and "health" from a pool of data so that a disease signature can be revealed.

After clustering the data in each scenario, one cluster is assigned to the Bioprofile of AD, and the other to the Bioprofile of normality, on the basis of how this dementia is known to modify the variables (e.g., MRI atrophy for scenario 3). In this way, we account for two underlying populations: subjects with and without signs of pathology. Then, for each subject, we compute a Bioindex, a continuous variable ranging from 0 to 1. Conceptually, clusters of pathology and normality are depicted
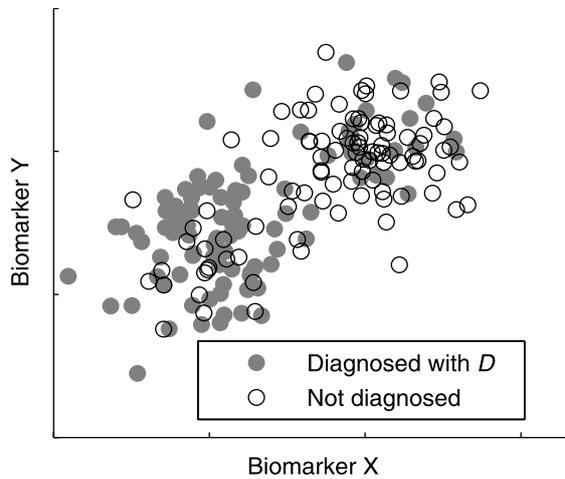
(a) Two hypothetical variables
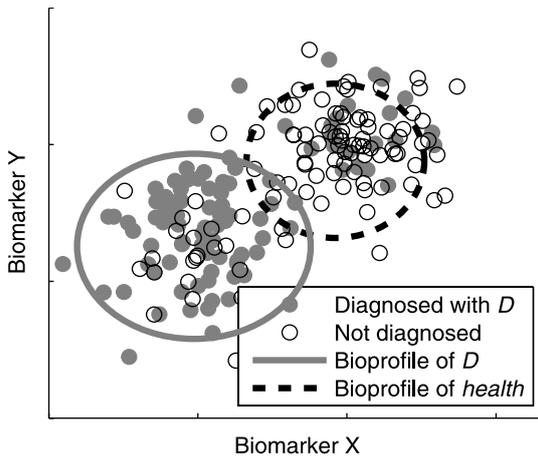
(b) Hypothetical Bioprofiles
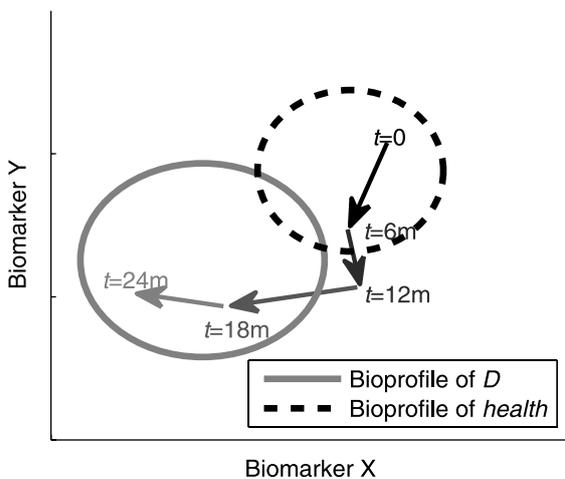
(c) Temporal evolution of a hypothetical subject

in Fig. 1, which also shows how the data of a hypothetical subject with a disease "*D*" may evolve over time.

Clustering with k-means

Clustering splits a set of instances into categories to explore their structure. It provides insights for data analysis and it has applications in diagnosis and knowledge extraction [21, 28, 33]. For a review, see [28].

We clustered the data with *k*-means [21, 28, 33]. This is a classic iterative distance-based method that works well in many practical problems [28]. A certain number of clusters (*k*) is specified in advance [21, 33]. Then, *k*-means partitions a number of data cases into *k* groups so that each case is assigned to the group (cluster) with the closest mean [21, 28, 33]. For the technical details of this procedure, the reader is referred to the Supplementary Data.

Computation of the Bioindices

A Bioindex is derived from the Bioprofile to provide information about a single subject. It is a continuous variable ranging from 0 (Bioprofile of normality) to +1 (Bioprofile of the disease). It can be seen as the likelihood that the subject's data belongs to the Bioprofile of disease and not to the Bioprofile of normality. Alternatively, the Bioindex shows the normalized distance from the subject's data to the Bioprofile of disease in comparison to the distance to the Bioprofile of normality.

The computation of the Bioindex depends on the machine learning algorithm. In the case of clustering though, all procedures revolve around the ideas of internal homogeneity and external separation of clusters [28]. Thus, all these methods incorporate a formulation of similarity, distance, or degree of membership to each cluster [28]. Using these formulations, we define the Bioindex as the normalized difference

Fig. 1. Conceptual representation of the Bioprofile of a disease, "*D*". Panel (a) shows two variables collected from a pool of subjects. Some participants (grey filled circles) have been diagnosed with "*D*", while others are not (empty circles). Some asymptomatic subjects may actually have an on-going pathological process of "*D*" or some patients of "*D*" may have been misdiagnosed. Panel (b) shows hypothetical Bioprofiles derived with clustering. There is a Bioprofile of "*D*" and a Bioprofile for the absence of "*D*" (or "health"). Panel (c) illustrates the temporal evolution of the data of a new subject who, over 24 months, develops the disease "*D*". The subject's data progressively move from the Bioprofile of "health" to the Bioprofile of "*D*". A Bioindex can be computed to reflect this evolution and help in the monitoring and prediction of disease.

between a subject's degrees of membership to the Bioprofile of the pathology and normality. This definition is general and it can be used with different techniques and diseases.

In the case of k-means and AD, we compute the distances from the subject's data to the centroids of the clusters corresponding to the Bioprofile of AD and normality and the inter-cluster distance. Let these values be d_{AD}, d_{Nor}, and d_{Inter}, respectively, once all variables have been normalized to the [0,1] interval. The Bioindex is then defined as:

$$Bioindex = \frac{1}{2} \left(\frac{d_{Nor} - d_{AD}}{d_{Inter}} + 1 \right) \qquad (1)$$

Higher Bioindices suggest the presence of more AD-related abnormalities in the subject's data.

Experimental procedures

Four experiments were performed. In each of them, the variables were normalized to the [0,1] interval using only the corresponding training data. Further details about each experiment are provided in the Supplementary Data.

Experiment 1: Bioprofiles of AD

The first experiment aims at introducing the Bioprofile analysis and to evaluate whether a Bioprofile of AD can be found in different data modalities. k-means [28, 33] is used to group the subjects' data into two clusters, which are then allocated to the Bioprofiles of AD and normality considering how AD affects the variables. This procedure is repeated for each scenario and the proportion of Bioprofiles of AD in each subject group (CN, MCI, and AD) is measured. We compute the χ^2 statistic and p-value for the cross-tabulation of the results to clarify whether the Bioprofiles convey information about the disease.

Experiment 2: Link between the Bioprofile and a model of AD evolution

The second experiment investigates whether the Bioprofiles agree with the model of AD evolution [17, 18]. High Bioindex values are expected in scenarios that change early with the disease as the variables of such scenarios will have the most abnormal values.

The results of this experiment are based on the process of experiment 1 but only the first five scenarios are considered because the aim of this experiment is to inspect the relationships between individual modalities and the model of AD evolution.

We inspect the distributions of the resulting Bioindices visually with boxplots. Additionally, within each diagnostic group (CN, MCI, and AD), we apply a one-way ANalysis Of VAriance (ANOVA) with a Bonferroni correction to assess whether the Bioindices differ among scenarios, thus implying an ordering of the biomarkers in AD.

Experiment 3: Relationship between the Bioprofile of AD and the risk of developing AD at the MCI stage

The third experiment is concerned with the potential link between the risk of progression from MCI to AD and the Bioprofiles and Bioindices. The training set for this experiment consists only of CN and AD subjects. The Bioprofiles derived from them are then used to estimate the Bioindices and Bioprofiles for the MCI people.

Cox proportional hazard models are used to assess the risk of conversion from MCI to AD associated with the Bioprofiles and Bioindices. A feature selection process is used to decide which covariates are informative for this. The empirical cumulative survival distribution was estimated for the MCI subjects in the Bioprofile of AD and normality. To evaluate the relationship of the Bioprofiles and Bioindices with the progression from MCI to AD at the level of individual patients [34], two metrics were computed for each scenario:

1) The area under the ROC curve (AUC) [21] of the Bioindex for the separation of MCI converters (MCI subjects that progressed to clinical AD at some follow-up, cMCI) versus non-converters (those who remained as MCI, nMCI).
2) The accuracy [21] defined as the fraction of cMCI and nMCI subjects that were correctly assigned to the Bioprofile of AD and Bioprofile of normality, respectively.

For the sake of a fair comparison between supervised and unsupervised approaches, we also compute the classification performance of a state-of-the-art supervised classifier (support vector machine) in the classification of cMCI versus nMCI subjects. Further details are given in the Supplementary Data.

Experiment 4: Evolution of the Bioindices with time

The fourth experiment assesses how the values of the Bioindices change with time. Bioprofiles are calculated only with baseline data of CN and AD subjects. Then, they are used as a reference to monitor the change in the MCI patients over time by computing their baseline and

follow-up Bioindex values. High Bioindices are associated with data similar to the Bioprofile of AD. Thus, we hypothesize that subjects who progressed toward AD faster will have larger increases in their Bioindices. The MCI subjects were split into cMCI and nMCI for illustration purposes.

A sigmoid function is fit to the Bioindex values at the level of individual MCI subjects to visualize their evolution. The sigmoid function is selected because current evidence suggests that changes in the biomarkers and clinical variables follow this contour as the disease progresses [17–20]. Whereas previous studies fit a sigmoidal-shape function to data gathered from large groups of subjects, we fit it to data of individual subjects.

The sigmoidal function used to fit the baseline and follow-up Bioindices is defined in [19] as:

$$f(t) = \frac{a}{1 + \exp\left[(b - t)/c\right]} \quad (2)$$

where a, b, and c are the parameters of the function. t represents the time of follow-up. The outcome of the function is the expected value of the Bioindex according to Equation (2). For additional details about this function, the reader is referred to [19].

RESULTS

Experiment 1: Bioprofiles of AD

Two clusters were revealed (for the Bioprofile of AD and normality) from the baseline data in each of the six scenarios. Table 1 contains the fraction of CN, MCI, and AD subjects with the Bioprofile of AD in each case. Most AD and CN subjects had the Bioprofiles of AD and normality, respectively. A χ^2 test confirmed (p-value < 0.0001) that the Bioprofile had information about the disease.

We assessed whether age, gender, number of ApoE $\varepsilon4$ alleles, and number of years of education differed

Table 1
Percentage of control (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD) subjects in the Bioprofile of AD for the six scenarios. Most AD patients had the Bioprofile of AD while most CN subjects had the Bioprofile of normality

Scenario	% CN subjects	% MCI subjects	% AD subjects
1: CSF Aβ_{42}	37.7	72.2	91.2
2: FDG-PET	27.2	51.2	85.6
3: MRI	8.7	46.6	75.5
4: CSF Tau	9.6	31.8	43.0
5: Scores	0.4	38.3	96.9
6: All	0.0	46.7	95.7

between Bioprofiles. There were significant differences (t-test, p-value < 0.0001) between the mean age and education level of the subjects in each Bioprofile in scenarios 3 and 5, respectively. The number of ApoE $\varepsilon4$ alleles also varied significantly between the Bioprofiles in all scenarios (χ^2 test, p-value $= 0.0023$ for scenario 2; p-value $= 0.0004$ for scenario 6; otherwise: p-value < 0.0001).

Experiment 2: Link between the Bioprofile and a model of AD evolution

We investigated the relationship between the Bioindices derived from the Bioprofiles and the model of AD evolution [17]. In Table 1, the percentage of subjects in the Bioprofile of AD tended to decrease from scenarios 1 to 4, something that suggests a temporal order in the biomarkers.

For the sake of a fair comparison among scenarios, we only considered the 186 subjects with all variables available. Figure 2 contains the boxplots of the Bioindices for CN, MCI, and AD subjects. In all three diagnosis groups, the overall level of the Bioindices tended to decrease from scenario 1 to 4. A one-way ANOVA with Bonferroni correction was carried out and it confirmed the significance of the differences in the Bioindices among scenarios and subject groups (detailed in the Supplementary Data).

Experiment 3: Relationship between the Bioprofile of AD and the risk of developing AD at the MCI Stage

In each scenario, we clustered only the CN and AD subjects' baseline data. Then, we derived from them the baseline Bioindex values and Bioprofiles for the MCIs.

The subjects' age, gender, number of ApoE $\varepsilon4$ alleles, and years of education were introduced in Cox regression models with a forward feature selection applied to all MCI subjects with available follow-up data. Only the number of ApoE $\varepsilon4$ alleles was selected for inclusion in scenarios 1 to 5 but such variable was not selected in scenario 6. This could be due to the smaller number of subjects in this scenario. Then, the presence of the Bioprofile of AD in the subject's data (a binary variable: yes/no) was entered into the Cox regression. Alternatively, the Bioindex value was used instead of the Bioprofile assignment.

The Bioprofile was significantly associated with the progression from MCI to AD in all scenarios but 4. Table 2 shows the corresponding p-value and the risk

Scenarios for CN subjects

(a) CN subjects

Scenarios for MCI subjects

(b) MCI subjects

Scenarios for AD subjects

(c) AD subjects

Fig. 2. Boxplots showing the distribution of the baseline Bioindex values for (a) CN, (b) MCI, and (c) AD subjects for inspection of the relationship between Bioprofile and the model of AD evolution in the five scenarios composed of individual data modalities. There is a trend to have lower Bioindices from scenario 1 to 4. This suggests that some biomarkers might become abnormal earlier than others in the disease.

factors associated with the Bioprofile of AD (with 95%CIs). The survival analysis was replicated with the Bioindices instead of the Bioprofiles and similar results were found (also in Table 2). The empirical survival functions for the Bioprofiles of AD and normality appear in Fig. 3, showing that the MCI subjects in the Bioprofile of AD are more likely to convert to AD than those with the Bioprofile of normality in all six scenarios.

We also computed results at the level of individual subjects for the ability of the baseline Bioprofile analysis to predict the progression from MCI to AD in terms of AUC and accuracy values. These results are contained in Table 3.

Experiment 4: Evolution of the Bioindices with time

The fourth experiment assessed whether the Bioindices reflected changes over time in each of the six scenarios. The CN and AD subjects' baseline data were used to derive the Bioprofiles. From these clusters, we computed the Bioindices for baseline and follow-up data of nMCI and cMCI subjects. For the subjects with three or more follow-ups (the process needs at least as many points as parameters), we fit a sigmoidal function to their Bioindex values. These results are plotted in Fig. 4. Grey dots and thin lines represent individual measures and sigmoidal fits for individual subjects, respectively, while the thick black line depicts the median of the sigmoidal regressions for the groups of cMCI and nMCI subjects. Steeper increments indicate that the subjects' data become more similar to the Bioprofile of AD. Of note is that, in scenarios 1, 4, and 6, there were a much smaller number of subjects with data available because these scenarios require the acquisition of CSF not only at baseline, but also at two or more follow-ups.

DISCUSSION

We have presented the Bioprofile as a surrogate method of reflecting the underlying pathological processes of AD using clinical variables and biomarkers collected in different clinical scenarios. The Bioprofile may also be useful in other neurodegenerative conditions such as Parkinson's disease. Our main objectives were to show that a Bioprofile of AD can be found in the data with unsupervised machine learning and to illustrate the potential utility of this approach in specific tasks. In order to do so, we ran four experiments.

Table 2
p-values and hazard ratios [with their 95%CI] for the progression from MCI to AD associated with the Bioprofiles and Bioindices in each scenario. The Bioprofile and Bioindex were associated with the risk of decline from MCI to AD

Scenario	Bioprofile as predictor		Bioindex as predictor	
	p-value	Risk for unit of change in the predictor	*p*-value	Risk for unit of change in the predictor
1: CSF Aβ$_{42}$	0.0027	2.778 [1.424, 5.421]	0.0014	3.509 [1.627, 7.566]
2: FDG-PET	<0.0001	3.147 [1.934, 5.121]	<0.0001	5.526 [2.570, 11.879]
3: MRI	<0.0001	2.053 [1.471, 2.866]	<0.0001	3.505 [2.189, 5.612]
4: CSF Tau	0.3162	1.258 [0.803, 1.972]	0.0503	1.879 [0.999, 3.535]
5: Scores	<0.0001	2.133 [1.557, 2.922]	<0.0001	6.539 [3.418, 12.509]
6: All	0.0482	1.917 [1.005, 3.655]	0.0010	23.011 [3.561, 148.702]

The first experiment assessed the presence of the Bioprofile of AD in six scenarios. Except for scenario 4, the Bioprofile of AD emerged from the data in over 75% of the AD patients, even though the diagnosis was not considered in the clustering procedure that led to the Bioprofiles. Yet, this study included only AD-related data. The methodology assumes that one Bioprofile is associated with each pathological process so that other diseases would have their own Bioprofiles. This is, though, a hypothesis that requires future work with datasets containing more than one disease.

The results from scenario 1 agreed with those reported with a Gaussian mixture model of CSF Aβ$_{42}$ [30]. Furthermore, the results suggest that unsupervised machine learning can consistently reveal a Bioprofile of AD in other data modalities, either when considered on their own (scenarios 2 to 5) or combined (scenarios 6) [31]. Moreover, the significant differences in mean age and education level of subjects in each Bioprofile in scenarios 3 and 5 support that these covariates are relevant factors in brain atrophy and cognitive performance, respectively [3]. In addition, the number of ApoE ε4 alleles varied significantly between the subjects in both Bioprofiles in all six scenarios. This agrees with the fact that ApoE ε4 is a major risk factor for AD [3, 16, 35]. Current guidelines for the clinical diagnosis of AD do not consider the carriage of ApoE ε4 allele specific enough to increase the certainty that the dementia suffered by people who meet the core clinical criteria for probable AD is indeed caused by AD [36]. However, ApoE ε4 increases the risk for progressing from MCI to AD within a few years [37] and a recent study [35] advocated raising the importance of the role of ApoE ε4 in AD. In our study, we consider ApoE as a covariate not to downplay its relevance in AD, but because the genotype does not change during the patient's lifetime. On the other hand, the values of all other variables in our scenarios change with the disease [14, 18, 38].

Experiment 2 was carried out because experiment 1 showed that the number of subjects in the Bioprofile of AD decreased from scenario 1 to 4. Differences in the average level of the Bioindices in those scenarios were inspected both visually (with boxplots) and statistically (with a Bonferroni-corrected ANOVA). The results confirmed that the Bioindices tended to decrease from scenario 1 to 4. This suggests the following sequence of biological changes from healthy aging to AD: Aβ$_{42}$ ⟶ FDG-PET ⟶ MRI ⟶ tau. This ordering mostly agrees with [17, 18] and complements the first validations of the model of AD evolution, where the values of the biomarkers were regressed against Alzheimer's Disease Assessment Scale-cognitive subscale scores [19, 20]. However, this neuropsychological test is not an optimal instrument to measure disease progression [16, 26] and nMCI subjects were left out of the analysis in [19]. In our results, the average level of the Bioindices for CSF tau seems lower than that of MRI. This may indicate that changes in tau occur slightly after, not before, the atrophy measured in the MRI, something that has been found in other studies [11, 20]. However, this may be because CSF tau might stabilize at later stages of the disease [16].

For a single patient, the presence of the Bioprofile of AD in a number of scenarios might serve as an indicator of a potentially abnormal pathological process requiring closer inspection by a clinician. There is a number of healthy elderly people (about 30%) who are "amyloid positive" but do not suffer from AD. This percentage roughly corresponds with the prevalence of AD dementia one decade later [18, 38] and it is also similar to the number of CN subjects assigned to the Bioprofile of AD in the first two scenarios of our study. However, it is unclear whether abnormal levels of such biomarkers are a risk factor for developing the clinical syndrome of AD or whether they constitute an early stage of AD itself [18]. Hence, the presence of the Bioprofile of AD in a number of scenarios cannot
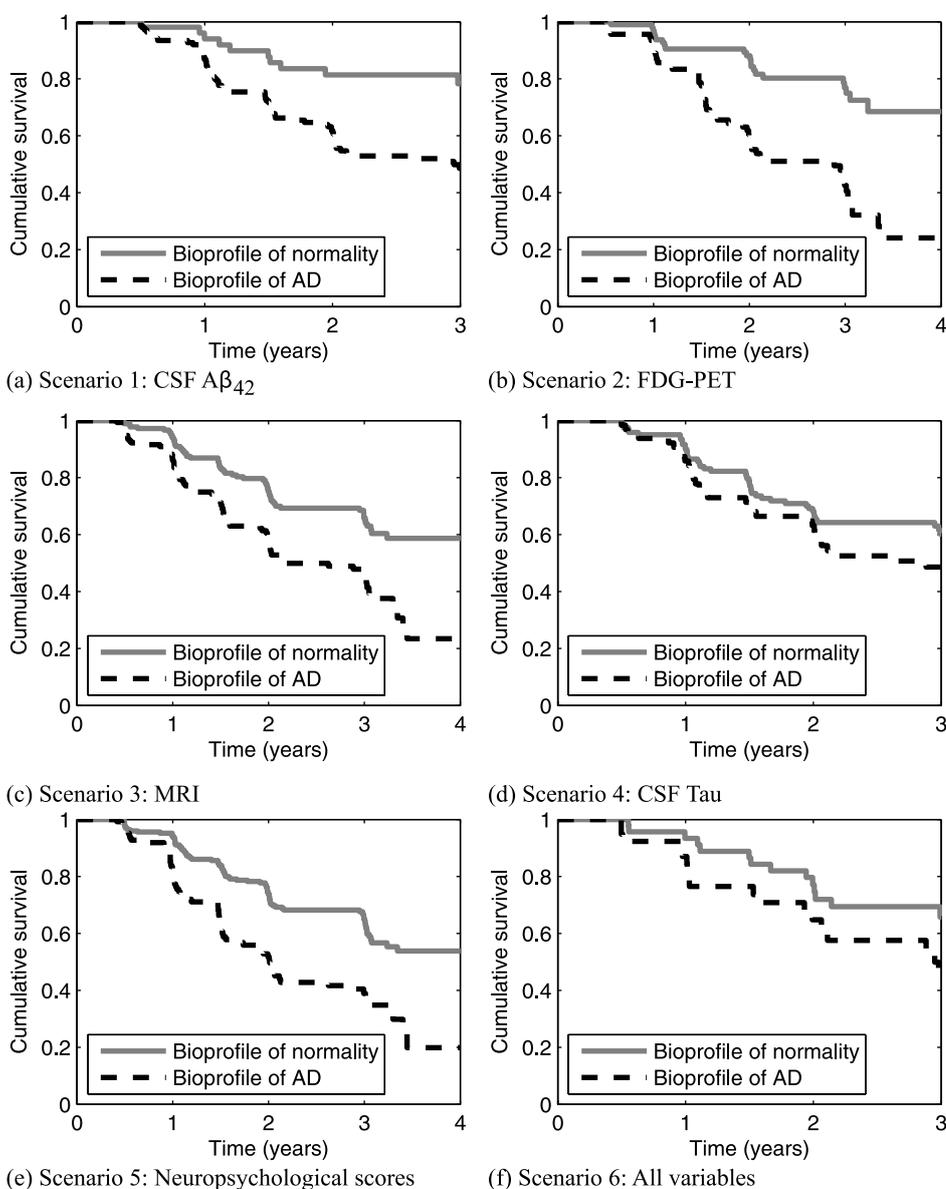
Fig. 3. Empirical survival functions for the conversion from MCI to AD for both Bioprofiles in scenarios 1 (a) to 6 (f). The lines represent the cumulative ratio of nMCI subjects over time. The presence of the Bioprofile of AD in the subject's data (black dashed line) implies a higher risk of progression to AD than the Bioprofile of normality (grey full line). Temporal axis shows the number of years for which follow-up data is available for both Bioprofiles in each scenario.

Table 3

AUC and Accuracy values for the unsupervised Bioprofile-based separation of cMCI versus nMCI subjects in experiment 3. The baseline Bioprofile analysis showed some ability to predict the future progression from MCI to AD at the level of individual subjects

Scenario	1: CSF Aβ_{42}	2: FDG-PET	3: MRI	4: CSF Tau	5: Scores	6: All
AUC	0.642	0.682	0.649	0.654	0.650	0.642
Accuracy	0.601	0.675	0.601	0.579	0.620	0.578
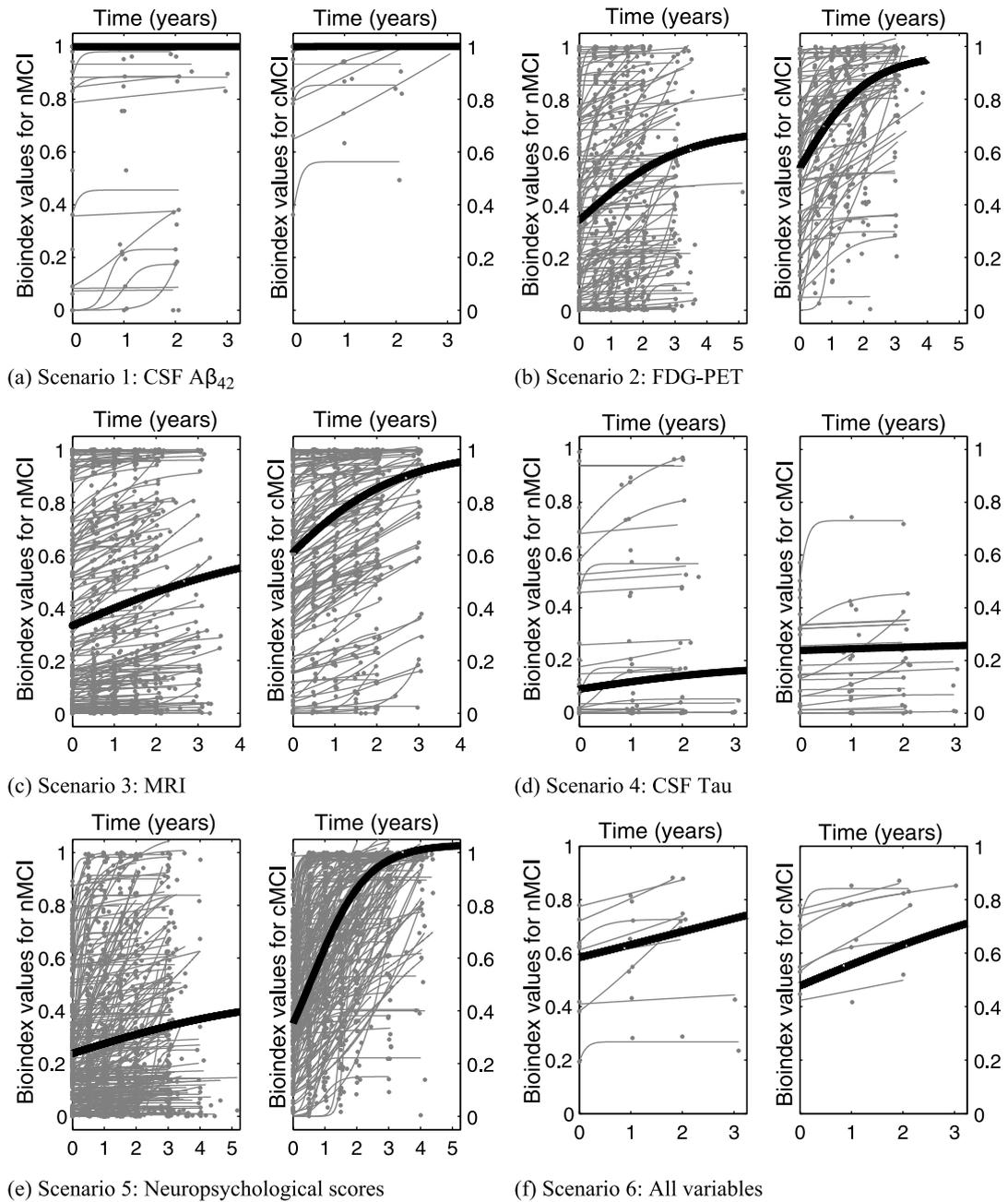
Fig. 4. Temporal evolution of the Bioindex values for the nMCI (left) and cMCI (right) subjects in scenarios 1(a) to 6 (f). Grey dots and lines represent individual measures and fittings of a sigmoidal function, respectively. The full black line shows the median trajectory of each group. The Bioindices of the cMCI subjects tended to increase faster than those of the nMCI, suggesting that the Bioindices may help to illustrate disease progression.

be considered yet as an early diagnosis of AD in an otherwise healthy person.

Experiment 3 elucidated whether the Bioprofiles and Bioindices had information about the risk of developing AD at the MCI stage. The Bioprofiles were computed only from baseline data; follow-up information was not used in their calculation. Even so, the clustering of MCI people into Bioprofiles of AD and normality was significantly associated with their future rates of conversion to AD in all scenarios but

4. Moreover, the Bioindex had a closer association with the risk of progression than the crisp rule of the assignment to either Bioprofile, including the case of scenario 6 (combination of all variables). This could be due to the fact that the Bioindices, being a continuous variable, are better able to provide a faithful representation of the continuum of AD pathological changes than the binary assignment to either Bioprofile. The results show that the Bioprofile analysis is linked with decline from MCI to AD. The AUC and hazard ratio values for the Bioprofiles and Bioindices are comparable to those reported in the literature. Semi-supervised and supervised MRI-based machine learning classifications provided an AUC of 0.69 [32] and accuracies of about 70% [5] for the separation of nMCI versus cMCI. A number of biomarkers and clinical variables have been associated with hazard ratios for this task ranging from 2 to 6.3 [4]. These values are comparable to those found in this study.

Finally, experiment 4 illustrated the evolution of the Bioindices with time. cMCI subjects showed steeper progressions over time toward the Bioprofile of AD than nMCI people. We fit the sigmoidal functions at the level of individual subjects rather than grouping together the data of all subjects as in previous articles [19, 20]. It is also worth noting that the results obtained in scenarios 1, 4, and 6 in experiment 4 are limited by the fact that fewer participants had follow-up CSF measures in the ADNI. For instance, scenario 6 in experiment 4 included very few subjects (10 CN, 8 nMCI, 7 cMCI, and 6 AD).

These experiments illustrate the process of deriving Bioprofiles and Bioindices from a biomedical dataset in AD. This dementia affects the brain years before the clinical symptoms appear [17, 18]. Bearing this in mind, we introduced the Bioprofile as a tool to aid in the identification of a potentially ongoing process of AD. We suggest that, if routinely collected data are available in the future, the Bioprofile approach, together with other disease models, could be used to extract knowledge from such data for clinical and research purposes.

The Bioprofile analysis was carried out using several scenarios, which represent different data modalities located at different times in the progression of AD pathology [17, 18]. We included variables frequently reported in AD literature, which are also discussed in the diagnostic criteria for this disease [5–11, 14, 16–18, 36–38]. The scenarios enabled us to assess the usefulness of different multimodal data in this context. The neuropsychological scores of scenario 5 can be obtained in a General Practice without the need

for additional biomarker analysis facilities. However, many current clinical scales have intrinsic limitations in their measurements [16]. On the other hand, the lumbar puncture to collect CSF provides information about both $A\beta_{42}$ and tau proteins but patients usually consider it a very invasive procedure. MRI is not invasive and can be performed several times without the learning effects of the clinical scores but it is not suitable for some patients (e.g., people with pacemakers). PET can also provide *in vivo* information about the brain, but it is expensive. Thus, every scenario entails its own advantages and limitations and requires different specific facilities. In practice, not all facilities may be available. Our analysis accounts for this because, by working with different scenarios, the results can be directly extended to cases where only some of the modalities are available. Additionally, we also considered scenario 6 as a combination of variables because, in clinical practice, all available evidence should be considered. However, the results seem to suggest that just grouping the information from different modalities may not outperform the results obtained from the most relevant modality on its own. For instance, FDG-PET seems to be better suited to track the clinical decline from MCI to AD than just considering all variables equally informative like in scenario 6. We hypothesize that this is because different modalities evolve at different rates with the disease [17] and some variables are better suited to monitor different disease stages than others.

There is no universally agreed-upon definition of cluster [28]. Considering our aims, we can define a cluster as a set of data that conveys the pattern of a disease in contrast with the pattern corresponding to health. The usefulness of the clustering approach in this analysis relies on the absence of gold standard evidence for the presence of AD while the patient is still alive, as this can only be obtained in an autopsy [3]. Results from experiment 3 showed that the Bioprofile provided similar performance in the separation of cMCI from nMCI patients to that of a state-of-the-art optimized supervised classifier (see the Supplementary Data). We used diagnostic information only to validate the results in experiments 1 and 2 and to split the subjects into training and testing sets in experiments 3 and 4. Our analysis does not use that information in the computations of the clusters. Although it might be possible that we introduced a small bias in the results by considering the diagnostic labels to create the training and testing sets, such bias, if existing at all, would be smaller than that of studies relying on the diagnosis of patients with-

out autopsy confirmation of AD to train supervised classifiers.

It is important to relate our results to the current clinical and research guidelines for AD diagnosis [18, 36–38]. Such criteria advocate the distinction between the AD pathological process and the clinically observable symptoms caused by that process [38]. The clinical diagnosis of AD and MCI must only be made on the basis of the patient's cognitive and behavioral symptoms [18, 36–38]. Further research about biomarkers is needed because the biomarkers cannot be included in the core criteria for clinical diagnosis yet [18, 36–38]. However, biomarkers can increase or decrease the certainty that clinical symptoms are due to AD pathology [18, 36, 37]. In the light of these recommendations, we suggest that high AD Bioindex values may support AD pathology as the cause of cognitive deficit. It is also possible to establish a hypothetical link between Bioindices and the spectrum of AD pathology; the Bioindices, being a continuous variable ranging from 0 to 1, reflects the suggestion that AD pathology is a continuum [18, 36–38]. Another potential clinical application of the Bioprofile concept lies on the fact that the assignment to the Bioprofile of AD or the Bioprofile of normality represents a clear distinction of normality from disease. This distinction might be useful in the future to decide up to which level biomarkers can be considered normal. This matter is still open to question because, nowadays, there are no established cut-off points that define when a biomarker can be considered normal or abnormal [36–38]. Finally, if the Bioprofile approach is confirmed by further results, the concept could contribute to defining stages in the disease process, i.e., the presence of abnormally high Bioindices in some early scenarios but not in other later ones could reflect a specific stage. Obviously, we acknowledge that the ultimate diagnostic decision must rely on the clinician's judgment [36, 37].

Some limitations merit consideration. Further analyses are needed with other clustering methods to corroborate our results and extend them to other modalities, paying special attention to genotype information. Secondly, autopsy-confirmed diagnoses are not available for all ADNI subjects [16]. Hence, it is impossible to ensure that the Bioprofile of AD appears in, and only in, the subjects with autopsy-confirmed pathology. Finally, we also acknowledge that additional research on biomarkers and on the Bioprofile methodology is needed before these can be incorporated into clinical practice [18].

To sum up, it is essential to develop techniques to assist clinicians in the interpretation of data related to AD [16]. To this end, we introduced the Bioprofile and Bioindices of AD to quantify abnormal disease patterns in the subjects' data. We used *k*-means [21, 28, 33] to derive Bioprofiles of AD in six scenarios. The Bioprofile emerged from the subjects' biodata without considering the diagnosis. Moreover, the Bioprofiles and Bioindices mostly agreed with a recent hypothetical model of AD evolution and they contained information related to the evolution of the disease. Therefore, they are a promising methodology that may help in the understanding of AD.

Yet, additional analyses are needed to extend the Bioprofile concept to more variables (including electromagnetic brain activity [12, 13] and functional connectivity assessments [14, 15]) and diseases. The collection of Bioprofiles across populations would constitute a valuable resource for research and personalized healthcare. An individual's relative characteristics with respect to a set of Bioprofiles could be studied to help in the interpretation of his or her data. In this sense, it is expected that the pattern of biological changes in other conditions, such as Parkinson's disease, will also be revealed with Bioprofile techniques.

ACKNOWLEDGMENTS

REFERENCES

[1] Hu P, Sun L, Ifeachor E (2009) A Framework for Bioprofile Analysis Over Grid. *IEEE Sys J* **3**, 520-535.

[2] Wimo A, Prince M (2010) *World Alzheimer Report 2010: The global economic impact of dementia.* Alzheimer's Disease International.

[3] Blennow K, de Leon MJ, Zetterberg H (2006) Alzheimer's disease. *Lancet* **368**, 387-403.

[4] Chen K, Ayutyanont N, Langbaum JBS, Fleisher AS, Reschke C, Lee W, Liu X, Bandy D, Alexander GE, Thompson PM, Shaw L, Trojanowski JQ, Jack CR Jr, Landau SM, Foster NL, Harvey DJ, Weiner MW, Koeppe RA, Jagust WJ, Reiman EM (2011) Characterizing Alzheimer's disease using a hypometabolic convergence index. *NeuroImage* **56**, 52-60.

[5] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage* **56**, 766-781.

[6] Frisoni GB, Fox NC, Jack CR, Scheltens P, Thompson PM (2010) The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol* **6**, 67-77.

[7] Holland D, Brewer JB, Hagler DJ, Fennema-Notestine C, Dale AM, the Alzheimer's Disease Neuroimaging Initiative (2009) Subregional neuroanatomical change as a biomarker for Alzheimer's disease. *Proc Natl Acad Sci U S A* **106**, 20954-20959.

[8] Landau SM, Harvey D, Madison CM, Koeppe RA, Reiman EM, Foster NL, Weiner MW, Jagust WJ (2011) Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiol Aging* **32**, 1207-1218.

[9] Nordberg A, Rinne JO, Kadir A, Langstrom B (2010) The use of PET in Alzheimer disease. *Nat Rev Neurol* **6**, 78-87.

[10] Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, Petersen RC, Blennow K, Soares H, Simon A, Lewczuk P, Dean R, Siemers E, Potter W, Lee VM-Y, Trojanowski JQ (2009) Cerebrospinal fluid biomarker signature in Alzheimer's Disease Neuroimaging Initiative subjects. *Ann Neurol* **65**, 403-413.

[11] Fagan AM, Head D, Shah AR, Marcus D, Mintun M, Morris JC, Holtzman DM (2009) Decreased CSF Aβ42 correlates with brain atrophy in cognitively normal elderly. *Ann Neurol* **65**, 176-183.

[12] Hornero R, Abásolo D, Escudero J, Gómez C (2009) Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with Alzheimer's disease. *Philos Trans R Soc A-Math Phys Eng Sci* **367**, 317-336.

[13] Babiloni C, Vecchio F, Lizio R, Ferri R, Rodriguez G, Marzano N, Frisoni GB, Rossini PM (2011) Resting state cortical rhythms in mild cognitive impairment and Alzheimer's disease: Electroencephalographic evidence. *J Alzheimers Dis* **26**, 201-214.

[14] Drago V, Babiloni C, Bartrés-Faz D, Caroli A, Bosch B, Hensch T, Didic M, Klafki H-W, Pievani M, Jovicich J, Venturi L, Spitzer P, Vecchio F, Schoenknecht P, Wiltfang J, Redolfi A, Forloni G, Blin O, Irving E, Davis C, Hårdemark H, Frisoni GB (2011) Disease tracking markers for Alzheimer's disease at the prodromal (MCI) stage. *J Alzheimers Dis* **26**(Suppl 3), 159-199.

[15] Escudero J, Sanei S, Jarchi D, Abásolo D, Hornero R (2011) Regional coherence evaluation in mild cognitive impairment and Alzheimer's disease based on adaptively extracted magnetoencephalogram rhythms. *Physiol Meas* **32**, 1163-1180.

[16] Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ (2012) The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimers Dement* **8**, S1-S68.

[17] Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* **9**, 119-128.

[18] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280-292.

[19] Caroli A, Frisoni GB (2010) The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. *Neurobiol Aging* **31**, 1263-1274.

[20] Yang E, Farnum M, Lobanov V, Schultz T, Raghavan N, Samtani MN, Novak G, Narayan V, Dibernardo A (2011) Quantifying the pathophysiological timeline of Alzheimer's disease. *J Alzheimers Dis* **26**, 745-753.

[21] Witten IH, Frank E, Hall MA (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier.

[22] Dickerson BC, Stoub TR, Shah RC, Sperling RA, Killiany RJ, Albert MS, Hyman BT, Blacker D, deToledo-Morrell L (2011) Alzheimer-signature MRI biomarker predicts AD dementia in cognitively normal adults. *Neurology* **76**, 1395-1402.

[23] Dickerson BC, Bakkour A, Salat DH, Feczko E, Pacheco J, Greve DN, Grodstein F, Wright CI, Blacker D, Rosas

HD, Sperling RA, Atri A, Growdon JH, Hyman BT, Morris JC, Fischl B, Buckner RL (2009) The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cereb Cortex* **19**, 497-510.

[24] Bakkour A, Morris JC, Dickerson BC (2009) The cortical signature of prodromal AD. *Neurology* **72**, 1048-1055.

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

[26] Llano DA, Laforet G, Devanarayan V (2011) Derivation of a new ADAS-cog composite using tree-based multivariate analysis. *Alzheimer Dis Assoc Dis* **25**, 73-84.

[27] Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, Waldemar G, Soininen H, Lötjönen J (2011) A disease state fingerprint for evaluation of Alzheimer's disease. *J Alzheimers Dis* **27**, 163-176.

[28] Xu R, Wunsch DC (2010) Clustering algorithms in biomedical research: A review. *IEEE Rev Biomed Eng* **3**, 1-35.

[29] Nettiksimmons J, Harvey D, Brewer J, Carmichael O, DeCarli C, Jack CR Jr, Petersen R, Shaw LM, Trojanowski JQ, Weiner MW, Beckett L (2010) Subtypes based on cerebrospinal fluid and magnetic resonance imaging markers in normal elderly predict cognitive decline. *Neurobiol Aging* **31**, 1419-1428.

[30] De Meyer G, Shapiro F, Vanderstichele H, Vanmechelen E, Engelborghs S, De Deyn PP, Coart E, Hansson O, Minthon L, Zetterberg H, Blennow K, Shaw L, Trojanowski JQ, for the Alzheimer's Disease Neuroimaging Initiative (2010) Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol* **67**, 949-956.

[31] Escudero J, Zajicek JP, Ifeachor E, The Alzheimer's Disease Neuroimaging Initiative (2011) Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and k-means. In *Proc 33rd Ann Int Conf IEEE Eng Med Biol Soc* IEEE EMBC2011, Boston, MA, USA, pp. 6470-6473.

[32] Filipovych R, Davatzikos C (2011) Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *NeuroImage* **55**, 1109-1119.

[33] Hartigan JA, Wong MA (1979) A k-means clustering algorithm. *J R Stat Soc Ser C* **28**, 100-108.

[34] Haller S, Lovblad KO, Giannakopoulos P (2011) Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease. *J Alzheimers Dis* **26**(Suppl 3), 389-394.

[35] Genin E, Hannequin D, Wallon D, Sleegers K, Hiltunen M, Combarros O, Bullido MJ, Engelborghs S, Deyn PD, Berr C, Pasquier F, Dubois B, Tognoni G, Fiévet N, Brouwers N, Bettens K, Arosio B, Coto E, Zompo MD, Mateo I, Epelbaum J, Frank-Garcia A, Helisalmi S, Porcellini E, Pilotto A, Forti P, Ferri R, Scarpini E, Siciliano G, Solfrizzi V, Sorbi S, Spalletta G, Valdivieso F, Vepsalainen S, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Bossu P, Hanon O, Piccardi P, Annoni G, Seripa D, Galimberti D, Licastro F, Soininen H, Dartigues J-F, Kamboh MI, Broeckhoven CV, Lambert JC, Amouyel P, Campion D (2011) APOE and Alzheimer disease: A major gene with semi-dominant inheritance. *Mol Psychiatr* **16**, 903-907.

[36] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 263-269.

[37] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.

[38] Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 257-262.