33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and *K*-Means

Javier Escudero [*], *Member, IEEE*, John P. Zajicek, Emmanuel Ifeachor, *Member, IEEE*, and the Alzheimer's Disease Neuroimaging Initiative [#]

***Abstract*—Alzheimer's Disease (AD) is the most common neurodegenerative disease in elderly people. There is a need for objective means to detect AD early to allow targeted interventions and to monitor response to treatment. To help clinicians in these tasks, we propose the creation of the Bioprofile of AD. A Bioprofile should reveal key patterns of a disease in the subject's biodata. We applied *k*-means clustering to data features taken from the ADNI database to divide the subjects into pathologic and non-pathologic groups in five clinical scenarios. The preliminary results confirm previous findings and show that there is an important AD pattern in the biodata of controls, AD, and Mild Cognitive Impairment (MCI) patients. Furthermore, the Bioprofile could help in the early detection of AD at the MCI stage since it divided the MCI subjects into groups with different rates of conversion to AD.**

I. Introduction

ALZHEIMER'S DISEASE (AD) is the most common neurodegenerative disease among elderly people [1]. In 2006, there were 26.6 million AD patients worldwide and its prevalence is expected to grow fourfold by 2050 due to the aging population [1]. The progression of AD starts several years before the first symptoms appear and it remains undetected in that time [2], [3]. Mild Cognitive Impairment (MCI) is often considered a transitional stage between normal aging and dementia [2], [4]. However, this condition is heterogeneous. Whereas some MCI patients develop AD or other dementias, others remain as MCI patients for many years [2], [4]. There is a need for objective means to help clinicians in the very early detection of AD to allow targeted therapeutic interventions and to monitor the response to treatment. Techniques are being developed to address these needs by using information derived from multimodal (i.e., clinical, neuroimaging, and biochemical) data [3], [4].

We conjecture that the healthcare of subjects with MCI and AD would greatly benefit from the creation of a "Bioprofile" able to identify the likelihood of progressing to AD at the MCI stage. The Bioprofile could also help to run Clinical Trials (CTs) in AD. Conceptually, a Bioprofile is a personal "fingerprint" that fuses together a person's bio-history. It contains the temporal evolution of his or her biomedical data and the results of automated decision-support tools (Machine Leaning) for diagnosis, prognosis, and monitoring of health [5]. The concept of Bioprofile may be useful to characterize diseases and detect them early. By applying suitable analysis algorithms to diverse biomedical variables, the Bioprofile may reveal the pattern of a specific disease in the subject's biodata. This idea is particularly appealing in complex conditions, such as AD [5], that evolve over a long period of time. With a Bioprofile of AD, it would be possible to scan the data of subjects in a patient (i.e., dementia) registry to look for those with the clearest signs of AD. By studying only this subset of patients, the effect of a new drug might be better shown in a CT.

Recently, MCI and AD subjects have been analyzed with Machine-Learning-based approaches that can be related to the idea of the Bioprofile [6], [7]. One study assessed whether Cerebrospinal Fluid (CSF) biomarkers reflect the AD pathology in Cognitive Normal (CN), MCI, and AD subjects without using the clinical diagnosis [6]. A CSF "signature" of AD was revealed in 90% of AD, 72% of MCI, and 36% of CN subjects with un-supervised Machine Learning [6]. In another study, the MCI subjects were considered unlabeled cases because it is difficult to ascertain who will progress from MCI to AD when few follow-ups are available [7]. A semi-supervised classifier was applied to Magnetic Resonance Imaging (MRI) scans to divide the MCI subjects into "AD-like" and "CN-like" and predict their progression [7]. These studies show that un-supervised and semi-supervised Machine Learning techniques might help to build a Bioprofile of AD from various biodata [6], [7].

Thus, our aim is to investigate whether a Bioprofile of AD emerges from clinical and/or biomarker data without using the diagnostic labels. This Bioprofile could help to

Manuscript received March 25th, 2011. This work was supported in part by the UK National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10124; Clinical Trial Methods in Neurodegenerative Diseases). The Foundation for the National Institutes of Health (www.fnih.org) coordinates the private sector participation of the $60 million ADNI public-private partnership that was begun by the National Institute on Aging (NIA) and supported by the National Institutes of Health. *Asterisk indicates corresponding author.*

J. Escudero [*] and E. Ifeachor are with the Signal Processing and Multimedia Communications Research Group, School of Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK (phone: +44 1752 586295; javier.escudero@ieee.org, e.ifeachor@plymouth.ac.uk).

J.P. Zajicek is Head of the Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, University of Plymouth, Derriford, Plymouth, PL6 8BX, UK.

[#] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

characterize AD. We also aim at assessing the utility of the Bioprofile in the early detection of AD at the MCI stage. We present preliminary results that address these two questions. We build on previous studies [6], [7], the idea of Bioprofile [5], and an un-supervised technique (*k*-means) [8] and we consider five clinical scenarios. Each scenario represents a different clinical infrastructure and enables us to assess the presence of the Bioprofile in diverse multimodal settings.

II. Materials and Methods

A. ADNI Database

Data used in the preparation of this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). 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 CTs. The initial goal of ADNI was to recruit about 200 CN 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 www.adni-info.org.

B. Selection of Variables in Clinical Scenarios

The ADNI data used in this study are as of February 7th, 2011. The database was queried for basic demographic, clinical (number of ApoE ε4 alleles, ADAS-Cog, and MMSE scores), MRI [9], and CSF [10] data from CN, MCI, and AD subjects at baseline. This query reported 381 records. However, two AD and four MCI patients were excluded from the analysis because not all their CSF values were available at baseline. Thus, the analyses were carried out with 375 subjects, whose basic data are detailed in Table I.

MRI is included in the differential diagnosis of AD from vascular dementia [11]. The hippocampal atrophy and cortical thickness are correlated with the clinical decline in AD and MCI [4], [11]. In [9], features from T_1 MRI scans acquired at 1.5T were computed with FreeSurfer. The results were reviewed and minimally edited for accuracy. Additional details can be found in [9]. Here, we consider the average of the left and right hippocampal volumes normalized by the intra-cranial volume and entorhinal cortical thickness values.

On the other hand, the CSF $A\beta_{42}$ and p-Tau_{181P} protein levels might also have diagnostic information to predict the conversion from MCI to AD [3], [10]. As in [6], we consider the $A\beta_{42}$ and the log-transformed p-Tau_{181P} values. These protein concentrations were measured with a multiplex immunoassay platform from a CSF sample obtained with a lumbar puncture after overnight fast [10]. We also include the number of ApoE ε4 alleles, a risk factor for AD [3].

TABLE I
Basic Data of the ADNI Subjects Included in this Study

	CN (N=106)	MCI (N=178)	AD (N=91)
Gender (%)	51.9 / 48.1	65.2 / 34.8	57.1 / 42.9
Age	76.08 ± 5.34	74.71 ± 7.41	74.94 ± 7.99
Years of education	15.76 ± 2.81	15.78 ± 3.02	15.26 ± 3.27
ApoE ε4 (%)	74.5/23.6/1.9	45.5/43.3/11.2	30.8/46.1/23.1
ADAS-Cog	6.37 ± 2.93	11.67 ± 4.46	18.08 ± 6.03
MMSE	29.10 ± 1.04	26.92 ± 1.80	23.51 ± 1.95

Data are given as mean ± standard deviation (SD), except for the gender and ApoE ε4 distributions, where the relative frequencies of male / females and number of subjects with 0 / 1 / 2 ApoE ε4 alleles are given.

To inspect how the Bioprofile of AD may help to the early detection of AD at the MCI stage, the diagnoses of all 178 MCI subjects were retrieved at follow-ups of 6, 12, 18, 24, and 36 months. Not all subjects had the same number of follow-ups (mean: 4.12 visits). During this time, 72 MCI subjects progressed to AD. The follow-up diagnoses were not used to derive the Bioprofiles. They were used only as a validation of the predictive power of the Bioprofile of AD.

C. K-Means Clustering

The Bioprofile of AD should reveal itself from the subjects' biodata without relying on the diagnostic labels. In this study, we assume two underlying populations (AD-like and CN-like or, alternatively, pathologic and non-pathologic) [6]. Then, we apply clustering (i.e., unsupervised Machine Learning) to integrate the information from diverse variables into relevant disease patterns [8]. Given a set of instances, the clustering splits them into categories to explore their structure and provide insights for further analyses [8].

We categorize the subjects' data with *k*-means, an iterative distance-based clustering method [8]. Briefly, a certain number of clusters (k) must be specified in advance (two in this case: CN-like and AD-like). Then, k points are randomly selected as cluster centers. All data instances are assigned to their closest center according to the Euclidean distance. The cluster centers are computed as the mean of all instances belonging to each cluster. This process is repeated until the same points are assigned to the same cluster in consecutive iterations [8]. We have applied *k*-means with the Weka software (version 3.6.3), which provides a comprehensive collection of Machine Learning techniques [12].

D. Bioprofile Analysis in Five Clinical Scenarios

Five different clinical scenarios are tested. The first one entails variables that can be obtained in a routine visit to the GP: two cognitive scales (ADAS-Cog and MMSE) and the ApoE genotype measured from a blood sample. The second and third scenarios include biomarkers: CSF and MRI, respectively. Several studies have shown the potential of MRI and CSF in AD diagnosis [2]–[4], [10], [11]. These biomarkers cover a wide range of the pathological changes in AD, which are hypothesized to start with abnormalities in $A\beta_{42}$, followed by p-Tau_{181P} and MRI atrophy [3]. The fourth and fifth scenarios are multimodal. Scenario four includes both CSF and MRI, while the fifth combines all cognitive scales and biomarkers. These scenarios reflect the different clinical set-ups needed to create the Bioprofile.

Two experiments are performed. The first involves clustering the subjects' data in each scenario and estimating the rate of appearance of the Bioprofile of AD in each group. This tests whether the Bioprofile of AD naturally emerges from the data and which variables may define it. The second experiment applies *k*-means only to CN and AD subjects. The resulting clusters are used to divide the MCIs into CN-like and AD-like. Then, the MCI subjects' rate of decline to AD is measured to assess whether the presence of the pathologic Bioprofile helps in the early detection of AD at the MCI stage. In the two experiments, *k*-means only considers baseline data and is blind to the diagnostic labels, which were only used for validation purposes.

III. Results

In Experiment 1, the rates of appearance of the Bioprofile of AD in the whole population of CN, MCI, and AD subjects were measured in five clinical scenarios. Table II gives the percentage of subjects that were assigned to the AD cluster in each case. More than 69% of the AD subjects and about half of the MCI individuals were always assigned to the pathological (i.e., AD) Bioprofile.

In Experiment 2, *k*-means was applied to baseline data from CNs and ADs to estimate clusters that split the MCI group into CN-like and AD-like. Table III shows the number of MCI subjects assigned to each cluster. The progression to AD of the CN-like and AD-like MCI subjects is shown in Fig. 1 at several follow-ups: 6, 12, 18, 24, and 36 months. Larger gaps between the AD-like (red) and CN-like (blue) lines suggest better ability of the baseline Bioprofile to predict future decline from MCI to AD. A two-sided Mann-Whitney U test was used to check the statistical significance ($p < 0.001$) of the differences in the rate of conversion.

IV. Discussion and Conclusions

We have carried out a preliminary assessment of the potential of five sets of biodata to create a Bioprofile of AD. This was motivated by the fact that AD is characterized by an unclear transient phase, which hinders the diagnosis and prognosis of the disease [4], [7]. Most AD patients were correctly assigned to the pathological cluster in all scenarios of Experiment 1. This confirms that un-supervised methods can characterize most AD subjects without being trained with the diagnosis labels [6]. On the other hand, the fraction of CN subjects assigned to the AD cluster ranged from 6.6% for MRI to the 33.0% for CSF. This might reflect that MRI is the latest biomarker to change in AD [3], [11]. The presence of the Bioprofile of AD in controls may indicate an undetected ongoing AD process [6]. The fraction of patients assigned to the Bioprofile of AD is smaller in Scenario 5 than in Scenarios 2 or 4. This could be due to the fact that all features are equally weighted in *k*-means [8], [12]. Hence, the inclusion of variables with different information for the Bioprofile of AD could have affected the assignment of CNs, MCIs, and ADs to the Bioprofile of AD in Scenario 5.

TABLE II
Fraction of CN, MCI, and AD Subjects Assigned to the Pathological Cluster (Bioprofile of AD) in Experiment 1

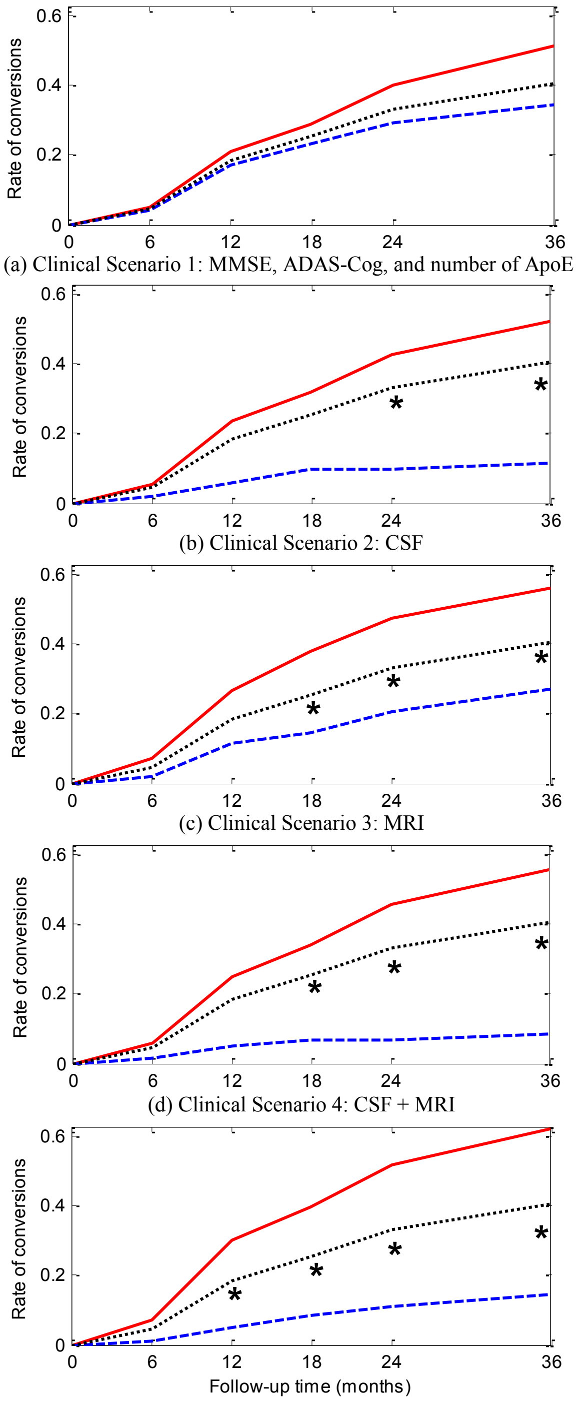
Clinical scenario	% CN	% MCI	% AD
1: MMSE+ADAS+ApoE	25.5	54.5	69.2
2: CSF	33.0	69.7	89.0
3: MRI	6.6	46.6	73.6
4: CSF+MRI	18.9	68.5	90.1
5: All	15.1	48.3	79.1

TABLE III
Number of MCI subjects in each Bioprofile in Experiment 2

Clinical scenario	CN-like Bioprofile	AD-like Bioprofile
1: MMSE+ADAS+ApoE	116	62
2: CSF	52	126
3: MRI	96	82
4: CSF+MRI	58	120
5: All	82	96

In Experiment 2, the follow-up information was only used to assess the conversion from MCI to AD diagnosis. It was not used to guide the assignment of MCI subjects to either Bioprofile. Furthermore, the Bioprofiles were based only on baseline data. Even so, the clustering of MCI people into AD-like and CN-like groups led to significant differences in their rates of conversion to AD in all Scenarios but the first one. All variables included in Scenario 1 can be obtained in a GP practice. Despite being closer to the clinical practice [4], this scenario did not perform better than the biomarkers in the early detection of AD at the MCI stage. On the other hand, MRI equipment is costly and the patients consider the lumbar puncture as an invasive procedure. These scenarios may help to assess the usefulness of various biodata in the creation of the Bioprofile. Every scenario entails a different infrastructure in clinical practice or in CTs. In this sense, future CTs may benefit from the creation of a Bioprofile of AD. Nowadays, patient registries with information from volunteers to participate in CTs are becoming widespread. The Bioprofile of AD could be compared against the data of the subjects in the registry to pinpoint those with higher likelihood of progressing to AD in the near future. The treatment of the CT could then be tested only on this more relevant subpopulation of AD-like subjects to evaluate if it helps in their clinical management.

Our preliminary results indicate that MRI and CSF may reveal the pathological pattern of AD more clearly than the clinical scales and ApoE. Only the combination of all variables in Scenario 5 provided significant differences between the evolution of CN-like and AD-like subjects at the 12-month follow-up, suggesting that a multimodal combination of clinical tests and biomarkers outperforms each modality alone. Additional analyses are needed to corroborate these results. There is no gold-standard in this setting as the diagnosis of AD can only be confirmed with an autopsy [2] but our results agree with previous studies [6].



(a) Clinical Scenario 1: MMSE, ADAS-Cog, and number of ApoE ε4

(b) Clinical Scenario 2: CSF

(c) Clinical Scenario 3: MRI

(d) Clinical Scenario 4: CSF + MRI

(e) Clinical Scenario 5: All previous variables

Fig. 1. Rates of conversion from MCI to AD. The black dotted line indicates the conversion in the whole sample of MCI subjects. The red full and blue dashed lines show the rate of conversion for the subsets of AD-like and CN-like MCI subjects according to the Bioprofile of each scenario: (a) MMSE, ADAS-Cog, and ApoE, (b) CSF, (c) MRI, (d) CSF and MRI, and (e) all previous variables. '*' marks follow-up points when the difference in the number of conversions between AD-like and CN-like subjects was significant ($p < 0.001$, two-sided Mann-Whitney U test).

In summary, we have used *k*-means [8] as a first step towards the creation of a Bioprofile of AD [5] in five scenarios. The Bioprofile emerged from the subjects' biodata without considering the diagnosis [6]. Moreover, the presence of the Bioprofile of AD led to significantly higher conversion rates from MCI to AD. This could help in the early detection of the disease. Yet, further analyses are needed to fully develop this concept. These include the inspection of other cluster tools (e.g., Expectation-Maximization algorithm) [8] and biomarkers (e.g., electrophysiological recordings and PET) [3], [5], [13]. Additionally, the potential of the Bioprofile in other neurodegenerative (e.g., Parkinson's Disease) or long-term (e.g., cancer) conditions should be investigated.

Acknowledgment

This article presents independent research commissioned by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10124). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimers Dement.*, vol. 3, no. 3, pp. 186–191, Jul. 2007.

[2] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's Disease," *Lancet*, vol. 368, no. 9533, pp. 387–403, Jul. 2006.

[3] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *Lancet Neurol.*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[4] J. L. Cummings, "Integrating ADNI results into Alzheimer's disease drug development programs," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1481–1492, Aug. 2010.

[5] P. Hu, L. Sun, and E. Ifeachor, "A Framework for Bioprofile Analysis Over Grid," *IEEE Syst. J.*, vol. 3, no. 4, pp. 520–535, Dec. 2009.

[6] G. De Meyer, F. Shapiro, H. Vanderstichele, E. Vanmechelen, S. Engelborghs, P. P. De Deyn, E. Coart, et al., "Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people," *Arch. Neurol.*, vol. 67, no. 8, pp. 949–956, Aug. 2010.

[7] R. Filipovych, C. Davatzikos, and the Alzheimer's Disease Neuroimaging Initiative, "Semi-supervised Pattern Classification of Medical Images: Application to Mild Cognitive Impairment (MCI)," *NeuroImage*, vol. 55, no. 3, pp. 1109–1119, Apr. 2011.

[8] R. Xu and D. C. Wunsch II, "Clustering Algorithms in Biomedical Research: A Review," *IEEE Rev. Biomed. Eng.*, vol. 3, no. 1, pp. 120–154, Dec. 2010.

[9] D. Holland, J. B. Brewer, D. J. Hagler, C. Fennema-Notestine, A. M. Dale, and the Alzheimer´s Disease Neuroimaging Initiative, "Subregional neuroanatomical change as a biomarker for Alzheimer's disease," *Proc. Nat. Acad. Sci.*, vol. 106, no. 49, pp. 20954–20959, Dec. 2009.

[10] L. M. Shaw, H. Vanderstichele, M. Knapik-Czajka, C. M. Clark, P. S. Aisen, R. C. Petersen, K. Blennow, et al., "Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects," *Ann. Neurol.*, vol. 65, no. 4, pp. 403–413, Apr. 2009.

[11] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, Jun. 2009.

[13] R. Hornero, D. Abásolo, J. Escudero, and C. Gómez, "Nonlinear analysis of electroencephalogram and magnetoencephalogram recordings in patients with Alzheimer's disease," *Philos. Trans. R. Soc. A-Math. Phys. Eng. Sci.*, vol. 367, no. 1887, pp. 317–336, Jan. 2009.