Journal of Alzheimer's Disease 32 (2012) 1–7
IOS Press

Supplementary Data

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*

Accepted 3 July 2012

For the sake of clarity, this document is organized following the same structure as the main text. Only sections for which there are relevant supplementary details are included here.

MATERIALS AND METHODS

Selection of variables

The variables included in the scenarios are described in the following paragraphs:

- Scenario 1: Cerebrospinal fluid (CSF) amyloid-β $(A\beta)_{42}$ protein concentration [1, 2] (one variable). In the Alzheimer's Disease Neuroimaging Initiative (ADNI), the $A\beta_{42}$ level—together with those of total tau (tTau) and phosphorylated tau ($pTau_{181p}$) proteins—were measured with a multiplex immunoassay platform from a CSF sample obtained with a lumbar puncture after overnight fast. Additional details appear elsewhere [1].
- Scenario 2: Mean 2-fluorodeoxy-D-glucose (FDG)-positron emission tomography (PET) glucose uptake level in two regions: left middle/inferior temporal and bilateral posterior cingulate gyri [2–5]. These two regions were generated from a literature review [4]. They were selected for being the only areas where mild cognitive impairment (MCI) patients showed a tendency toward greater annual decline than cognitively normal (CN) subjects in [4].
- Scenario 3: Atrophy of the magnetic resonance imaging (MRI) hippocampal volume and entorhinal cortical thickness [2, 6–9]. We consider the average of the left and right hippocampal volumes, normalized by the intra-cranial volume, and left and right entorhinal cortical thickness. The results were reviewed and minimally edited for accuracy in [8], where further details can be found.

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu/). 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. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

*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.

ISSN 1387-2877/12/$27.50

- Scenario 4: CSF tTau and $pTau_{181p}$ protein concentrations [1, 2, 10] (two variables), which were measured from the CSF acquired as described in scenario 1.
- Scenario 5: Two neuropsychological scores: the Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog). These scores are two of the most widespread clinical tests for Alzheimer's disease (AD). They account for the last period of AD evolution, when the symptoms of dementia become apparent [11, 12].
- Scenario 6: A combination of baseline data from all previous five scenarios (all nine variables) was used to create a sixth scenario which contains all data modalities.

An ADNI subject is included in as many scenarios as relevant. For instance, a person for whom FDG-PET, MRI, and neuropsychological scores were measured would be included in scenarios 2, 3 and 5, but not in 1, 4, and 6. The total numbers of subjects included in scenarios 1 to 6 are 414, 403, 737, 409, 817, and 186, respectively. Additional information about the ADNI sample and its procedures is detailed elsewhere [11]. The supplementary Table 1 contains the main characteristics of the subjects included in each scenario.

Clustering with k-means

This section presents a more detailed description of the algorithmic procedures of *k*-means clustering.

Formally, *k*-means minimizes the sum-of-squared-error criterion [13]:

$$J(\mathbf{\Gamma}, \mathbf{M}) = \sum_{i=1}^{k} \sum_{j=1}^{N} \gamma_{ij} \left\| \mathbf{x}_j - \mathbf{m}_i \right\|^2 \quad \text{(S1)}$$

where $\mathbf{\Gamma} = \{\gamma_{ij}\}$ is a partition matrix that indicates to which cluster each instance belongs:

$$\gamma_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in \text{cluster } i \\ 0 & \text{otherwise} \end{cases}, \quad \text{(S2)}$$

$$\sum\nolimits_{i=1}^{K} \gamma_{ij} = 1 \forall j. \quad \text{(S3)}$$

The matrix $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_K]$ contains the cluster centroids computed as the mean of all objects assigned to the cluster:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ij} \ \mathbf{x}_j. \quad \text{(S4)}$$

In Equations (S1) and (S4), N and N_i denote the total number of subjects in the dataset and in the cluster i, respectively [13].

k-means may be trapped in local minima if the initial partition is not appropriate [13–15]. To avoid this problem, we have applied *k*-means to the data ten times with different random initial centroids. Afterwards, we selected the solution with the lowest value of $J(\mathbf{\Gamma}, \mathbf{M})$.

Experiment 1: Bioprofiles of AD

To clarify the steps followed in experiment 1, supplementary Figure 1 shows a block diagram. In addition to the analyses detailed in the article, in experiment 1 we also carry out additional procedures to: 1) minimize the risk of over-fitting and 2) assess the consistency of the assignments of subjects to the Bioprofile of AD or the Bioprofile of normality. These are described in the following paragraphs.

The risk of over-fitting is low because the diagnosis is not used to find the clusters. However, 100 complete runs of a stratified ten-fold cross-validation [14] are applied in this experiment. This means that, in each of the runs, the dataset is randomly divided into ten folds. Nine of these folds are used to find the clusters (i.e., Bioprofiles). Then, the subjects in the left-out fold are assigned to the Bioprofile of AD or normality accordingly to the clusters derived in the training step. The results are aggregated over all folds.

The previously described process also enables us to check the consistency of the assignment of a subject to the Bioprofile. This is necessary to ensure that the Bioprofiles do not depend on the initial conditions of *k*-means. If that were the case, those 100 runs of the cross-validation would result in very different assignments to the Bioprofiles: sometimes a subject would be assigned to the Bioprofile of AD and sometimes he or she would be assigned to the Bioprofile of normality. To check that the Bioprofile does not suffer from this problem, we consider all 4,950 possible pairs out of the 100 realizations and, for each pair, we compute the proportion of subjects that are assigned to the same Bioprofile in both executions of the pair.

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

To clarify the steps followed in experiment 2, supplementary Figure 2 shows a block diagram.

Supplementary Table 1

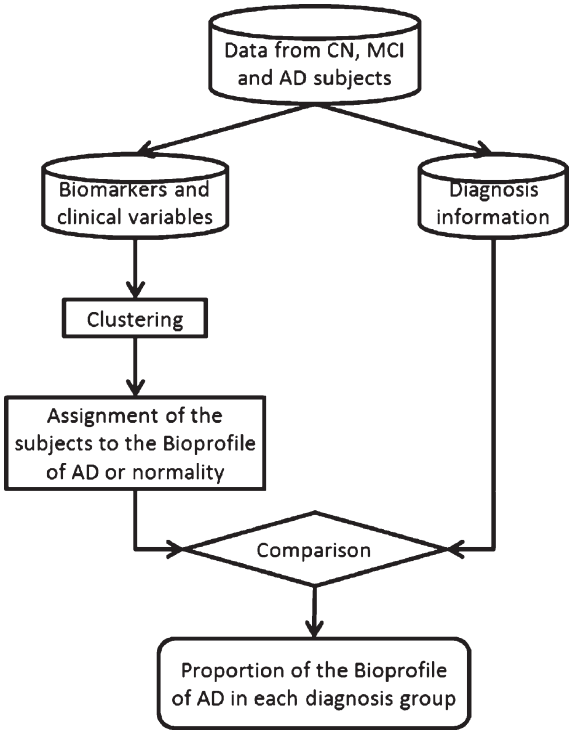
Basic baseline data for the six scenarios given as mean ± standard deviation (SD) in all cases apart from the gender distribution (number of males (M) and females (F)) and ApoE $\varepsilon 4$ alleles (number of subjects with 0, 1, and 2 alleles)

Scenario 1: CSF $A\beta_{42}$			
($n = 414$)	CN ($n = 114$)	MCI ($n = 198$)	AD ($n = 102$)
Gender distribution	58 M/56 F	132 M/66 F	59 M/43 F
ApoE $\varepsilon 4$ alleles	87/25/2	92/85/21	31/48/23
Age	76.14 ± 5.21	75.08 ± 7.50	75.53 ± 7.88
Years of education	15.72 ± 2.83	15.80 ± 2.99	15.16 ± 3.30
$A\beta_{42}$	205.59 ± 55.09	163.66 ± 54.89	142.98 ± 40.79
Scenario 2: FDG-PET			
($n = 403$)	CN ($n = 103$)	MCI ($n = 203$)	AD ($n = 97$)
Gender distribution	62 M/41F	137 M/66 F	58 M/39 F
ApoE $\varepsilon 4$ alleles	77/24/2	95/82/26	33/49/15
Age	76.39 ± 4.80	75.52 ± 7.21	76.23 ± 7.36
Years of education	15.82 ± 3.10	15.79 ± 2.88	14.65 ± 3.21
Left middle/inferior temporal	1.227 ± 0.132	1.161 ± 0.144	1.041 ± 0.159
Bilateral posterior cingulate	1.379 ± 0.165	1.285 ± 0.173	1.132 ± 0.146
Scenario 3: MRI			
($n = 737$)	CN ($n = 218$)	MCI ($n = 356$)	AD ($n = 163$)
Gender distribution	115 M/103 F	226 M/130 F	82 M/81 F
ApoE $\varepsilon 4$ alleles	158/55/5	162/151/43	54/78/31
Age	76.47 ± 5.09	75.09 ± 7.43	75.34 ± 7.63
Years of education	16.05 ± 2.86	15.66 ± 3.04	14.78 ± 3.19
Hippocampal volume $\times 10^4$	24.91 ± 2.69	22.02 ± 3.29	20.38 ± 3.30
Entorhinal cortical thickness	3.251 ± 0.302	2.926 ± 0.468	2.592 ± 0.443
Scenario 4: CSF tau			
($n = 409$)	CN ($n = 114$)	MCI ($n = 195$)	AD ($n = 100$)
Gender distribution	58 M/ 56 F	130 M/65 F	58 M/42 F
ApoE $\varepsilon 4$ alleles	87/25/2	90/84/21	21/46/23
Age	76.14 ± 5.21	75.07 ± 7.49	75.43 ± 7.90
Years of education	15.72 ± 2.83	15.82 ± 3.00	15.11 ± 3.30
tTau	69.68 ± 30.37	103.56 ± 60.91	121.61 ± 57.57
$pTau_{181p}$	24.86 ± 14.58	35.67 ± 18.13	41.70 ± 19.98
Scenario 5: Neuropsychological scores			
($n = 817$)	CN ($n = 229$)	MCI ($n = 397$)	AD ($n = 191$)
Gender distribution	119 M/110 F	256 M/141 F	101 M/90 F
ApoE $\varepsilon 4$ alleles	168/56/5	185/165/47	65/90/36
Age	76.51 ± 5.03	75.38 ± 7.44	75.93 ± 7.49
Years of education	16.03 ± 2.85	15.67 ± 3.04	14.71 ± 3.15
ADAS-Cog	6.20 ± 2.91	11.50 ± 4.42	18.60 ± 6.31
MMSE	29.11 ± 1.00	27.03 ± 1.78	23.39 ± 2.02
Scenario 6: All variables			
($n = 186$)	CN ($n = 49$)	MCI ($n = 90$)	AD ($n = 47$)
Gender distribution	32 M/17 F	59 M/31 F	30 M/17 F
ApoE $\varepsilon 4$ alleles	35/14/0	41/39/10	11/24/12
Age	75.69 ± 5.21	75.51 ± 7.01	75.80 ± 7.66
Years of education	15.69 ± 3.14	15.80 ± 2.90	14.83 ± 3.58

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

To clarify the steps followed in experiment 3, supplementary Figure 3 shows a block diagram.

For the sake of a fair comparison between the unsupervised Bioprofile approach and supervised techniques in the prediction of the progression from MCI to AD, we used a state-of-the-art supervised classifier (support vector machine, SVM) [6, 11, 16, 17] to separate MCI converters (cMCI, MCI patients



Supplementary Figure 1. Block diagram of the steps carried out to compute the Bioprofile results in experiment 1.

who declined from MCI to AD at a later follow-up) from MCI non-converters (nMCI, MCI subjects who remained as MCI in the future). Following the same approach as with the Bioprofile-based approach of experiment 3, the supervised classifiers were trained with CN and AD subjects' data and they were evaluated in terms of Area Under the ROC Curve (AUC) and accuracy for the separation of cMCI versus nMCI. The SVMs were optimized using a cross-validated grid-search approach on the training set varying the value of C, the kernel (polynomial or Gaussian), and a kernel parameter (degree of the polynomial kernel or value of gamma for the Gaussian one) [14, 17].

Experiment 4: Evolution of the Bioindices with time

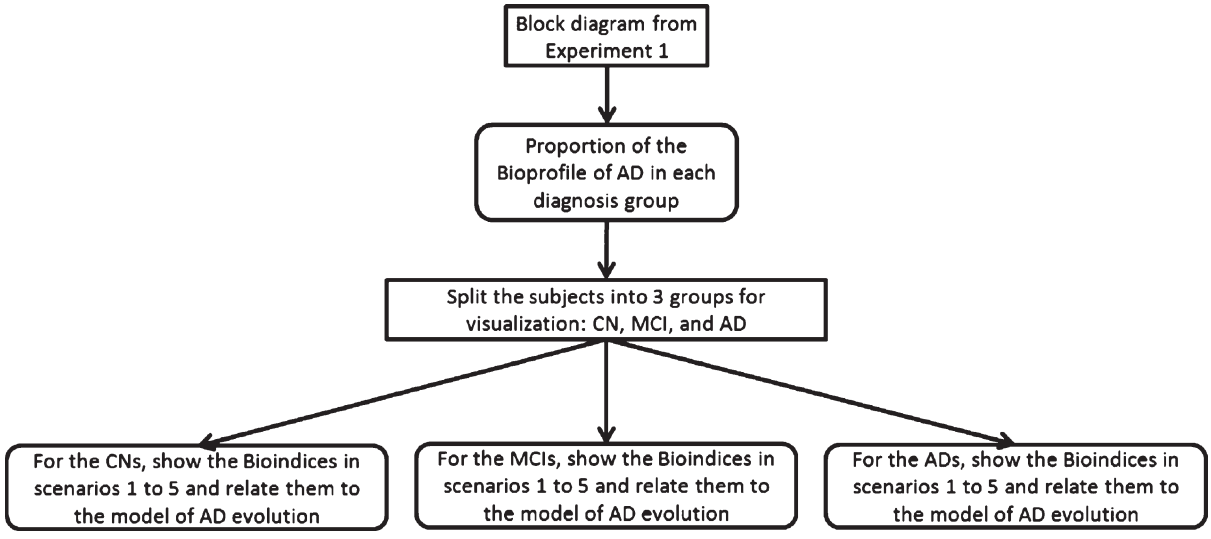
To clarify the steps followed in experiment 4, supplementary Figure 4 shows a block diagram.

This experiment requires the acquisition of the biomarkers at two or more follow-ups so that the sigmoidal function can be fit to the data. This reduces the number of cases available for analysis, as not all subjects have enough validated follow-up acquisitions of the biomarkers, especially for CSF measurements.

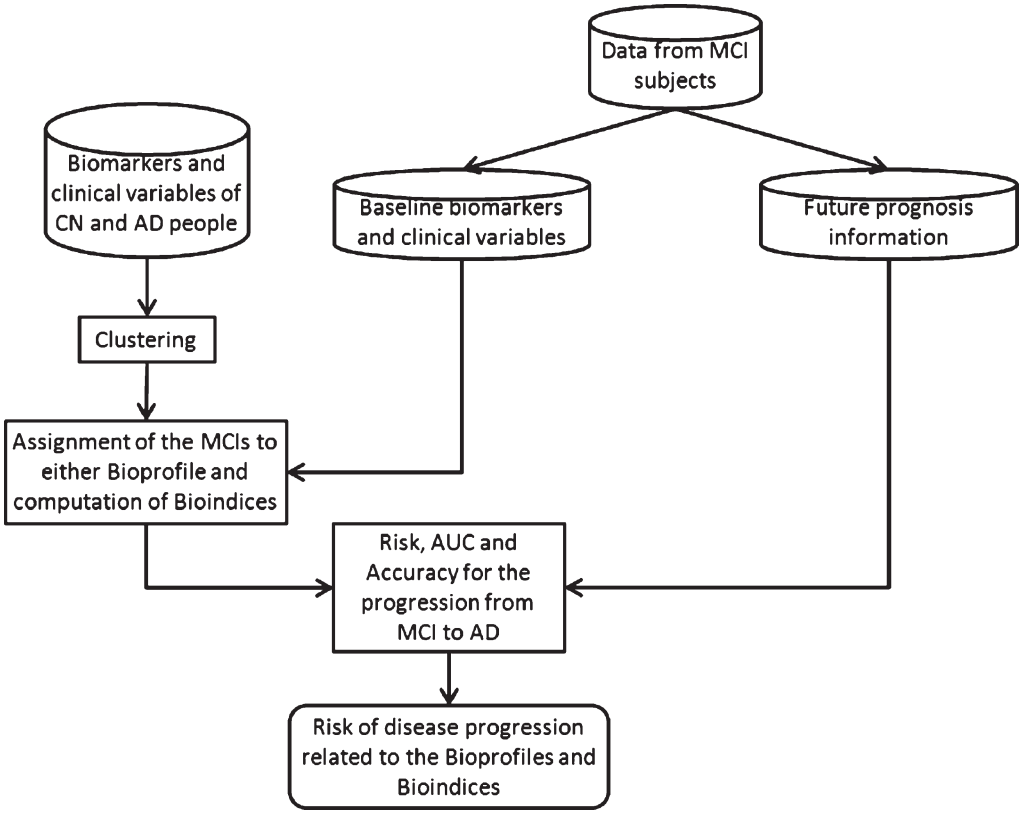
RESULTS

Experiment 1: Bioprofiles of AD

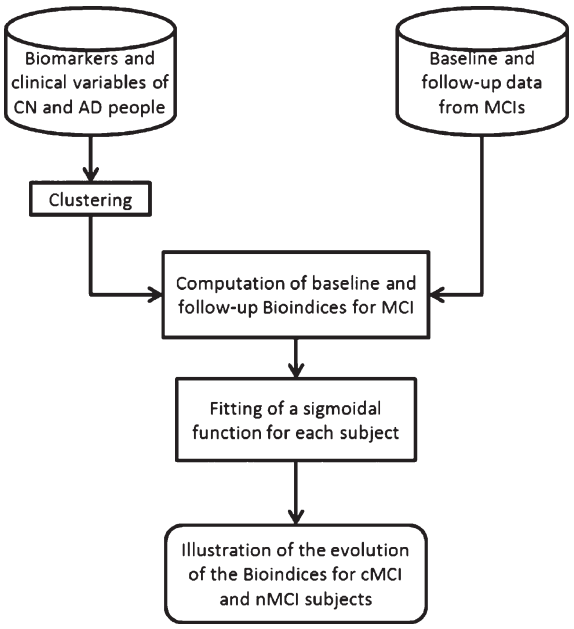
In order to verify that the Bioprofiles are consistent and that they do not depend on the random initialization of k-means, we ran 100 different realizations of the ten-fold cross-validation. For all possible pairs out of the 100 realizations, we computed the proportion of subjects that were found in the same Bioprofile. These results are given in supplementary Table 2 as



Supplementary Figure 2. Block diagram of the steps carried out to compute the Bioprofile results in experiment 2.



Supplementary Figure 3. Block diagram of the steps carried out to compute the Bioprofile results in experiment 3.



Supplementary Figure 4. Block diagram of the steps carried out to compute the Bioprofile results in experiment 4.

mean $\pm$ standard deviation (SD) and [minimum, maximum] range of those proportions for each scenario. The results indicate that the average level of agreement was about or over 0.974 in all scenarios, suggesting that the assignment to the Bioprofiles of AD or normality is consistent and it has very little dependence on the initial conditions of *k*-means.

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

Here, we detail which pairs of comparisons were significantly different (significance level: $\alpha = 0.05$) in experiment 2 (see Fig. 2 in the main text).

In the case of CN, the Bioindices computed for CSF $A\beta_{42}$ were significantly higher than those obtained for MRI, CSF tau, and the neuropsychological scores. The Bioindices for FDG-PET were significantly higher than those for CSF tau and the scores.

As for the Bioindices for the MCIs, the values for CSF $A\beta_{42}$ were significantly higher than for any other scenario. The Bioindices for FDG-PET were

Supplementary Table 2
Consistency of k-means in the assignment of subjects to the Bioprofiles over 100 runs of a ten-fold cross-validation. Results are given as mean ± SD and [minimum, maximum]

Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
0.998 ± 0.002	0.980 ± 0.007	0.993 ± 0.003	0.981 ± 0.007	0.993 ± 0.002	0.974 ± 0.010
[0.990, 1.000]	[0.948, 1.000]	[0.985, 1.000]	[0.954, 1.000]	[0.984, 1.000]	[0.930, 1.000]

Supplementary Table 3
AUC and accuracy values and optimized set-up for the supervised SVM-based classification of cMCI versus nMCI subjects in experiment 3

Optimal	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
set-up for the SVM	Polynomial kernel, degree = 1, C = $10^{0.5}$	Polynomial kernel, degree = 2, C = 10^{1}	Polynomial kernel, degree = 2, C = $10^{1.5}$	Gaussian kernel, gamma = 1, C = 10^{1}	Polynomial kernel, degree = 2, C = 10^{2}	Gaussian kernel, gamma = 1, C = $10^{0.5}$
AUC	0.659	0.651	0.619	0.649	0.646	0.591
Accuracy	0.631	0.645	0.615	0.646	0.642	0.586

significantly higher than for CSF tau and the scores, and even the Bioindices of MRI data were significantly different from those of CSF tau.

For the ADs, CSF tau had significantly lower Bioindices than any other scenario. CSF $A\beta_{42}$ and the neuropsychological scores provided significantly higher Bioindices than MRI.

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

Here, we report the results obtained with an optimized SVM [6, 11, 16, 17] in the separation of cMCI from nMCI in each scenario. The objective is to compare the performance of the unsupervised Bioprofile approach with that of an optimized supervised classifier.

Supplementary Table 3 contains the corresponding results for the supervised methodology, together with the optimal set-up of each SVM computed with a grid-search optimization. The corresponding results for the unsupervised Bioprofile approach are detailed in Table 3 in the main text.

The results showed that the unsupervised approach provided similar performances to those of the supervised SVM, even though clustering is a simpler algorithm easier to interpret by the medical community. However, we acknowledge that the Bioprofile methodology should be further developed to be used in clinical practice and that accuracies of about 65% are not high enough for clinical diagnosis. In any case, it is important to note that the estimation of progression from MCI to AD is a far more difficult problem that the separation of CN against AD subjects [6].

ACKNOWLEDGMENTS

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

This article was funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0707-10124.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical

sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

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

[16] 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.

[17] 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.