Supplementary Online Content

Doecke JD, Laws SM, Faux NG, et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. *Arch Neurol.* 2012. doi:10.1001/archneurol.2012.1282.

eAppendix. ADNI Cohort and Data Methodology

eTable. Top 21 Biomarkers (Variable Selection)

This supplementary material has been provided by the authors to give readers additional information about their work.

© 2012 American Medical Association. All rights reserved.

eAppendix. ADNI Cohort and Data Methodology

The Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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 magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (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 U.S. 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 www.adni-info.org.

© 2012 American Medical Association. All rights reserved.

Expanded Statistical Methodology

Data cleaning, transformation and imputation

For both cohorts, data cleaning consisted of removal of all data values, post log transformation, which were outside 10 standard deviations from the mean. All values that were below assay detectable limit were taken as missing data. Variables were removed if they were not 95% complete. From the AIBL cohort, a total of 177 biomarkers were tested; 111 RBM biomarkers, 52 clinical pathology biomarkers, 7 measures of circulating metals, 6 measures of Aβ, and total ApoE. From the ADNI cohort, a total of 174 biomarkers were tested, 136 RBM and 38 clinical pathology biomarkers. Individual samples with more than 50% missing data were deleted. Statistical imputation was conducted for both cohorts using the Multiple Imputation using Chained Equations (MICE) method.[1] Multiple imputation was conducted 100 times, with 100 separate iterated datasets created each time. The median imputed value was chosen to replace the missing value.

Biomarker selection

Variable selection was performed on both cohorts using a specific analysis pathway (Pathway 1, Figure 1) including including: Random Forest (RF), Boosted Trees (BT), Regression Trees (RT) and Linear Models for Micro Array (LIMMA) (Set 1, Figure 1).

Seventy percent of the data was randomly selected (training data) for biomarker selection (repeated 100 times). Biomarkers that were frequently selected in all four statistical methods in greater than 50% of repeats were chosen for class prediction analyses.

© 2012 American Medical Association. All rights reserved.

To validate the variable selections from the AIBL cohort, a separate set of variable selection techniques (Set 2, Figure 1) including: Best First (BF), Greedy Stepwise (GS), Forward Selection Regression, and Significance Analyses of Microarray (SAM) were used. Biomarkers that were chosen most frequently across all four methods were collated for further analyses. Furthermore, feature rankings from recursive feature elimination (RFE) coupled with a linear SVM were used to verify biomarker selection reproducibility. We show a Venn diagram (Figure 2) to demonstrate the reproducibility of the variable selections and confirm the choice of biomarkers for prediction, and a mind map (Figure 1) to illustrate the overall statistical design.

Disease predictions: AIBL cross validation

Once biomarkers were identified from the variable selection pathway (P1), a second statistical analysis pathway (Classification/Prediction, P2) was created, for the modelling of the chosen biomarkers together with demographic variables. Confounding variables such as education status, cardiovascular disease risk, physical activity and smoking history were assessed for contribution to disease prediction. Only those markers that added significant strength to predictions were included in the final models. Class predictions were performed using the classification methodologies Naiive Bayes (NB), RF and RFE-ordinal linear SVM. To assess the generalised performance of the classifiers, as measured by average sensitivity, specificity and AUC statistics, 3 fold cross-validation with 100 repeats were performed.

ADNI analyses and prediction validations

To identify biomarkers in ADNI that were significantly altered in the AD group, the GLM was used as described above. Classification performance and validation analyses were conducted

© 2012 American Medical Association. All rights reserved.

using two different approaches. Approach 1 consisted of using 3 fold cross validation (100 repeats) using the AIBL cohort with a) age, gender and *APOE* genotype only, b) age, gender, *APOE* genotype and a short list of biomarkers (Biomarker Set B), c) age, gender, *APOE* genotype and a full list of biomarkers (Biomarker Set A) and d) age, gender, *APOE* genotype and only those biomarkers from Biomarker Set A that were measured in both cohorts. Approach 2 randomly sampled (90%) across both AIBL and ADNI biomarker data sets; computed biomarker Generalized Linear Model using AIBL, and made class predictions on the ADNI data set. By taking a random sample across each AIBL and ADNI, we were able to iterate through a large proportion of times to increase accuracy.

© 2012 American Medical Association. All rights reserved.

eReference

1. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *J Stat Soft.* 2011;45(3):1-67.

© 2012 American Medical Association. All rights reserved.

Supplementary Material

AIBL	Fold^	P-value	ADNI	Fold^	P-value
Insulin like growth factor binding protein 2	1.61	<0.0001	**Pancreatic polypeptide**	1.75	0.0189
Pancreatic polypeptide	1.54	<0.0001	**Brain natriuretic peptide***	1.71	<0.0001
Erythrocyte sedimentation rate	1.46	0.002	**Tenascin C**	1.27	0.0017
B lymphocyte chemoattractant	1.45	0.002	Heparin-binding epidermal growth factor-like growth factor	1.26	0.0202
Carcinoembryonic antigen	1.4	0.001	Cancer antigen 19.9	1.24	0.0198
Cortisol	1.28	<0.0001	Alpha 1 microglobulin	1.20	0.0043
Tumor necrosis factor receptor like 2	1.27	0.0002	Adiponectin	1.18	0.0116
Homocysteine	1.23	0.002	Tissue inhibitor of metalloproteinases **1**	1.16	0.0018
Angiopoietin 2	1.23	0.003	**Eotaxin 3**	1.15	0.0008
Matrix metalloproteinase 9	1.19	0.001	Vascular cell adhesion molecule 1	1.13	0.0074
Vascular cell adhesion molecule 1	1.18	<0.0001	Matrix metalloproteinase 2	1.12	0.0009
Tissue inhibitor of metalloproteinases 1	1.18	0.0003	Alpha 2 macroglobulin	1.11	0.0024
Superoxide dismutase	1.16	<0.0001	Betacellulin	1.11	0.0151
Alpha 1 antitrypsin	1.11	0.0003	Mean cell hemoglobin	1.02	0.0120
Interleukin 10	1.1	<0.0001	Vitronectin	0.92	0.0048
Mean cell hemoglobin concentration	0.99	<0.0001	Transthyretin	0.88	0.0033
Albumin	0.95	<0.0001	Interleukin 16	0.85	0.0053
Haemoglobin	0.95	0.0003	Apolipoprotein AII*	0.83	0.0002
Calcium	0.94	<0.0001	Placenta growth factor	0.82	0.0218
Zinc (isotope 66)	0.91	<0.0001	Serum glutamic oxaloacetic transaminase	0.81	0.0007
Interleukin 17	0.87	<0.0001	**Tumor necrosis factor receptor like 2**	0.37	0.0089

Supplementary Table 1: Top 21 Biomarkers (Variable Selection)

RBM, clinical pathology and metal ions: Only those AIBL biomarkers with P-values less than 0.0003 are shown. For comparison, the same number of biomarkers from ADNI is shown. * Markers measured from the ADNI cohort significant p<0.0003