Supplementary data

Enriching Amnestic Mild Cognitive Impairment Populations for Clinical Trials: Optimal Combination of Biomarkers to Predict Conversion to Dementia

Peng Yu[a,*], Robert A. Dean[a], Stephen D. Hall[a], Yuan Qi[b], Gopalan Sethuraman[a], Brian A. Willis[a], Eric R. Siemers[a], Ferenc Martenyi[a], Johannes T. Tauscher[a], Adam J. Schwarz[a]
and for the Alzheimer's Disease Neuroimaging Intiative[1]
[a] *Eli Lilly and Company, Indianapolis, IN, USA*
[b] *Department of Computer Science, Purdue University, West Lafayette, IN, USA*

Accepted 14 June 2012

CLASSIFICATION BASED ON PREDICTIVE AUTOMATED RELEVANCE DETERMINATION

In this work, we selected predictive variables in a classification framework, using the collected measurements $\mathbf{X} = \{\mathbf{x}^n\}_{n=1}^N$ of the aMCI subjects as input variables. A group label $\mathbf{t} = \{t_n\}_{n=1}^N$, $t = \pm 1$ was used to indicate whether a subject converted to AD within the 2 years follow-up period. At last, we employed the predicative automated relevance determination (pred-ARD) method to conduct variable selection and classification on the training data to obtain a classifier $\mathbf{w}$.

Pred-ARD is a hierarchical Bayesian approach that determines the relevance of input variables based on their prediction performance. It extends the classical Bayesian variable selection method, automatic relevance determination (ARD). Both ARD and pred-ARD model the prior distribution of the parameters in the classifier to explicitly represent the relevance of different input variables. It is usually accomplished by assigning hyperparameters to determine the range of variation for the parameters relating to a particular input variable. In particular, the ARD method models the width of a zero-mean Gaussian prior on those parameters:

$$p(\mathbf{w}|\alpha) = \prod_i N(w_i|0, \alpha_i^{-1}) \qquad (1)$$

where $i = 1, \dots p$, and p is the number of variables. In ARD, the hyperparameters are estimated to maximize the model evidence (marginal likelihood):

$$p(\mathbf{t}|\mathbf{X}, \alpha) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)d\mathbf{w} \qquad (2)$$

where $\mathbf{X} = \{\mathbf{x}^n\}_{n=1}^N$ denote data, and $\mathbf{t} = \{t^n\}_{n=1}^N$, $t = \pm 1$ denoted group labels, As described before.

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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. A complete list of ADNI investigators is available at http://adni.loni.ucla.edu/research/active-investigators/.

[*] Correspondence to: Peng Yu, Eli Lilly and Company, Indianapolis, IN 46285, USA. Tel.: +1 317 277 5528; Fax: +1 317 276 6545; E-mail: yu_peng_py@lilly.com.

ISSN 1387-2877/12/$27.50 © 2012 – IOS Press and the authors. All rights reserved

The Pred-ARD method, proposed by Qi et al. [1], assigns hyperparameters α in the same fashion, but estimates them to optimize the predictive performance

$$p(t^{new}|\mathbf{x}^{new}, \mathbf{t}) = \int p(t^{new}|\mathbf{x}^{new}, \mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w} \quad (3)$$

As a result, many elements of α go to infinity, which naturally prunes irrelevant variables in the data. Furthermore, this method uses Expectation Propagation (EP), a more accurate approximation method, for pred-ARD model estimation, and uses the LOO generalization error obtained directly from EP as estimates of predictive performance.

Using pred-ARD, we can not only select relevant variables, but also learn the posterior distribution of classifier $p(\mathbf{w}|\mathbf{D}, \alpha)$ from the training set $\mathbf{D} = \{(\mathbf{x}^1, t^1), \ldots, (\mathbf{x}^N, t^N)\}$. The posterior distribution can then be used to estimate the posterior predictive distribution for a new data point using Equation (3). In this two-class classification problem, we adopt the simple decision rule:

$$t^{new} = \arg\max p(t^{new}|\mathbf{x}^{new}, \mathbf{t}) \quad (4)$$

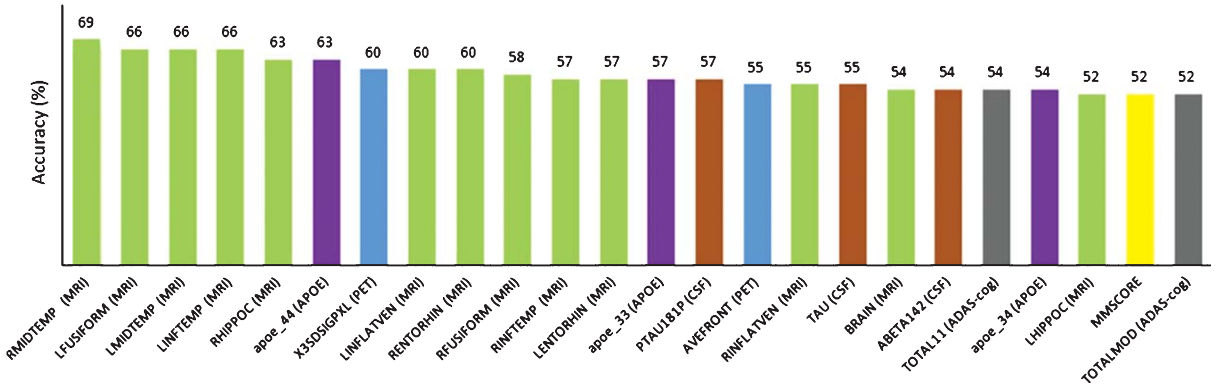
Using this method, we can estimate the posterior distribution $p(\mathbf{w}|\mathbf{D}, \alpha)$ of the classifier, where only variables relevant to separating converters from non-converters have nonzero weights. Moreover, we can rank the importance of these variables to the classification by using their corresponding weights. The larger the magnitude of the weight, the more significant the variable is for distinguishing the two groups. Finally, we can use $p(\mathbf{w}|\mathbf{D}, \alpha)$ to calculate the prediction probability on testing subject.

DISCUSSION ON CLASSIFICATION METHODS

The classification method we employed in this paper is a wrapper method that jointly select features while building the classification model. This kind of wrapper method can help to more accurately find only relevant biomarkers for predicting MCI to AD conversion, compared with a filtering approach. Secondly, the Bayesian

Supplementary Table 1
Summary of numeric biomarker and clinical variables used in this study

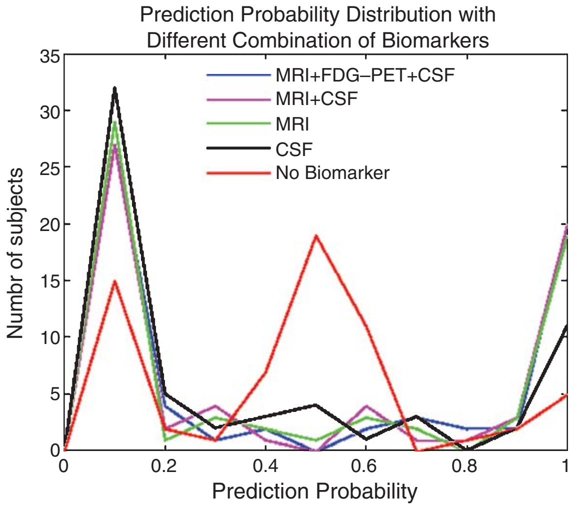
Variable	Annotation	Laboratory	Modality
L/R MIDTEMP	Left/right middle temporal cortex	University of California, San Diego	MRI
L/R INFTEMP	Left/ right inferior temporal cortex	University of California, San Diego	MRI
L/R FUSIFORM	Left/ right fusiform cortex	University of California, San Diego	MRI
L/R ENTORHIN	Left/ right entorhinal cortex	University of California, San Diego	MRI
BRAIN	Whole brain	University of California, San Diego	MRI
VENTRICLES	Ventricle	University of California, San Diego	MRI
L/R HIPPOC	Left/right hippocampus	University of California, San Diego	MRI
L/R INFLATVEN	Left/right inferior lateral ventricles	University of California, San Diego	MRI
$A\beta_{42}$	Amyloid beta 1-42	University of Pennsylvania	CSF
P-Tau_{181}	Phosphorylated Tau	University of Pennsylvania	CSF
tTau	Total Tau	University of Pennsylvania	CSF
AVEASSOC	Average cerebral metabolic rate of glucose (CMRglc) in frontal parietal and temporal cortices	University of Utah	FDG-PET
AVEFRONT	Average CMRglc in frontal cortex	University of Utah	FDG-PET
X2SDSIGPXL	Number of pixels with hypometabolic activity two standard deviations below normal mean	University of Utah	FDG-PET
X3SDSIGPXL	Number of pixels with hypometabolic activity three standard deviations below normal mean	University of Utah	FDG-PET
ROI-avg	The average signal from the right/left angular, right/left temporal, and bilateral posterior cingulate	University of California, Berkeley	FDG-PET
CV-fROI	Cross-validated region	Banner Alzheimer's Institute	FDG-PET
ApoE 23	ApoE Genotype 23	UPENN	ApoE
ApoE 24	ApoE Genotype 24	UPENN	ApoE
ApoE 33	ApoE Genotype 33	UPENN	ApoE
ApoE 34	ApoE Genotype 34	UPENN	ApoE
ApoE 44	ApoE Genotype 44	UPENN	ApoE
TOTALMOD	85 point total including Delayed Word Recall and Q14		ADAS-Cog
TOTAL11	Classic 70 point total excluding Delayed Word Recall and Number Cancellation		ADAS-Cog
MMSESCORE	Total Score		MMSE



Supplementary Figure 1. Leave one out classification accuracy by individual variables. Only variables with an accuracy >50% are shown in this figure. The bars are color-coded by modality.

Supplementary Table 2
RID (subject ID) of ADNI aMCI subjects used in this study. Converters and non-converters are listed in separate columns

Converters	Non-converters
101	51
204	150
214	291
222	292
231	293
240	314
256	361
258	378
294	424
344	443
511	552
723	566
904	626
906	634
930	673
941	718
978	722
997	746
1010	748
1033	783
1077	925
1130	932
1217	950
1398	973
1423	994
	1030
	1034
	1043
	1073
	1120
	1224
	1260
	1265
	1315
	1351
	1380
	1414
	1419



Supplementary Figure 2. Histograms of predicted conversion probability for 63 subjects using different combination of biomarker modalities in additional to ApoE, ADAS-Cog, and MMSE.

method we used is designed to build models that can be generalized well to other datasets, especially when the available training set is relatively small (63 subjects), and has been shown to provide better prediction performance compared with benchmark methods such as support vector machines [1]. This Bayesian method is also computationally efficient with the advanced approximation method for inference; in the present study, it took about 16 seconds to build the model with 63 subjects and 22 variables on a standard PC.

Furthermore, as a Bayesian method, it estimates the probability distribution of the classifier, instead of making a point estimation. Using these classifiers built with this Bayesian method, we can readily estimate the probability of conversion (0–100%) for any MCI subject using their corresponding biomarker measurements. Prediction accuracies shown in this paper were calculated with a probability threshold of 50%. Using this threshold, test patients with predicted conversion probability >50% were diagnosed as MCI to AD converters and patients with predicted conversion probability ≤50% were diagnosed as non-converters. When applying these classification models, we can change the probability threshold to increase either sensitivity or specificity. The change of classification threshold will subsequently affect the cost saving and patient-screening time when using this method for patient enrollment. For example, if we change the classification threshold to 85% (probability >85% are converters and probability ≤85% are non-converters), we would potentially further enrich our population with patients more likely to convert to AD. Accordingly, we would need to recruit fewer patients and reduce the cost associated with the clinical trial. However, since we are using more rigorous inclusion criteria (higher threshold), we would need to screen more patients and increase the cost and time associated with screening. These scenarios can be quantitatively simulated with these classification models to assess the logistical benefit and select the best threshold to use for patient selection in MCI clinical trials. As demonstrated in the results, the use of biomarkers (including CSF and imaging modalities) can generate strong predictions with >80% of subjects assigned to the first and fourth quartiles of prediction probability p (i.e., $p < 25\%$ or $p > 75\%$). In contrast, screening based on genotype and cognitive tests alone generates less informative predictions with borderline conversion probabilities, with 75% of subjects assigned a prediction probability between 25 and 75%.

In theory, for prospective application of our model in a clinical trial, it may not be necessary to use the same data acquisition or analysis methods as long as equivalent measurements, expressed in the same physical units, as those used to build the classification model are used. In practice, however, care must be taken for vMRI measures in particular; at the present time different structural MRI segmentation software packages and methods are likely to generate slightly different values for nominally the same brain structures. This is exemplified by the current ongoing effort to harmonize the manual delineation of the hippocampus [2]. Similarly, for FDG-PET, a composite SUVR measure should use consistent mask and reference regions. Thus, a classifier of the type presented here should be trained using summary measures generated using the same analysis method that will be applied in its prospective application to clinical trial data.

REFERENCES

[1] Qi Y, Minka TP, Picard RW, Ghahramani Z (2004) Predictive automatic relevance determination by expectation propagation. *Proceedings of Twenty-first International Conference on Machine Learning*, p 85, doi: 10.1145/1015330.1015418.

[2] Frisoni GB, Jack CR Jr (2011) Harmonization of magnetic resonance-based manual hippocampal segmentation: A mandatory step for wide clinical use. *Alzheimers Dement* **7**, 171-174.