32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Discovery and use of efficient biomarkers for objective disease state assessment in Alzheimer's disease

Mark van Gils, *Member, IEEE*, Juha Koikkalainen, Jussi Mattila, Sanna-Kaisa Herukka, Jyrki Lötjönen, Hilkka Soininen and the Alzheimer's Disease Neuroimaging Initiative*

Abstract—**Objective and early detection of Alzheimer's disease (AD) is a demanding problem requiring consideration of many-modal observations. Potentially, many features could be used to discern between people without AD and those at different stages of the disease. Such features include results from cognitive and memory tests, imaging (MRI, PET) results, cerebral spine fluid data, blood markers etc. However, in order to define an efficient and limited set of features that can be employed in classifiers requires mining of data from many patient cases. In this study we used two databases, ADNI and Kuopio L-MCI, to investigate the relative importance of features and their combinations. Optimal feature combinations are to be used in a Clinical Decision Support System that is to be used in clinical AD diagnosis practice.**

I. Introduction

Dementia causes long and oppressive suffering to patients and their relatives and imposes enormous costs on society. About 25 million people suffered from dementia in 2000. As a 4-fold increase of this number is expected by 2050. Alzheimer's disease (AD) covers 60-70% of all dementia cases. No cure for AD exists, and effective and reliable early diagnostic techniques are lacking. Early diagnosis and progress monitoring of AD is a central part of treatment once future drugs and prevention strategies become available. There is a strong indication that several different biomarkers provide a reliable and early indication of AD prior to its major clinical signs. However, it is not known which combinations of biomarkers/features are optimal for detection at different disease stages. Many research initiatives to address this problem are ongoing world-wide. For example, many research efforts are done centered around the ADNI initiative [1,2].

The study is part of the EU co-funded project PredictAD (www.predictad.eu). The objective of this part of the study is 1) to find the best combination of biomarkers for AD diagnostics from heterogeneous data (imaging, electrophysiology, molecular level, clinical tests, demographics) and 2) to develop clinically useful tools integrating the optimal biomarker results. The basis for this is formed by the consortium's large databases.

This paper reports first results regarding identification of efficient biomarker sets containing information related to AD diagnosis. Measures of efficiency are accuracy for detecting AD and correspondence measures with severity of AD. Of special interest is detection of the early phase of the disease, manifested in mild-cognitive impairment (MCI).

The optimal combination of biomarkers may vary between populations and there is thus a need for personalization. Also, different biomarkers may have different roles during different stages of the disease, thereby making the composition of the optimal set variable with progression of the disease.

The efficient combinations of biomarkers are eventually used in a tool for clinical use by the healthcare professionals – it allows review of all different measurements performed for a specific patient and shows the classification result together with the possibility to explore the influence the different biomarkers have on the classification result.

This paper is structured as follows. The Methods section gives a short description of the databases and analysis methods are described that are used to select and assess the suitability of different variables to separate between disease states. In Section III, results of these methods as applied to respectively ADNI and L-MCI data are presented. Finally, the results and their implications are discussed and continuing steps are outlined.

II. Methods

A. Data sets

Data from two sources have been used: the publicly available ADNI database and data obtained from the Kuopio L-MCI study as provided by University of Eastern Finland.

ADNI data cohort

The ADNI database [1] contains many-modal patient data from AD diagnosed as well as healthy control cases. Data exist from eight studies (screening, baseline, and

Manuscript received April 23, 2010. This work was supported by the European Commission under the 7th Framework Programme (FP7-224328 – PredictAD). 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, a detailed list of sponsors can be found at www.loni.ucla.edu/ADNI .

Mark van Gils, Juha Koikkalainen, Jussi Mattila and Jyrki Lötjönen are with VTT Technical Research Centre of Finland, Tampere, Finland (e-mail: firstname.lastname@vtt.fi).

Sanna-Kaisa Herukka and Hilkka Soininen are with the Neurology Department, University of Eastern Finland, Kuopio, Finland (e-mail: firstname.lastname@uef.fi)

* 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 are listed at the abovementioned website.

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

follow-ups at months 6, 12, 18, 24, 30 and 36). In total 505 features divided over 17 categories were used (Table 1).

TABLE I
CATEGORIES AND NUMBER OF FEATURES IN ADNI DATA

Demographic Information	16
Medical History	19
Diagnostic Summary	28
Diagnosis and Symptoms Checklist	28
Baseline Changes	16
Alzheimer Disease Assessment Scale(ADAS)	23
Apolipoprotein E risk gene (ApoE)	3
Clinical Dementia Rating (CDR)	7
Cerebrospinal Fluid (CSF)	4
Function Assessment Questionnaire (FAQ)	11
Geriatric Depression Scale (GDS)	17
Modified Hachinski (MH)	17
Mini-Mental State Exam (MMSE)	31
Magnetic Resonance Imaging (MRI)	136
Molecular Tests	93
Neuropsychological Battery (NB)	46
Neuropsychiatric Inventory Questionnaire (NIQ)	1
Positron Emission Tomography (PET)	9

At baseline the number of cases was 821: 343 (42%) female and 478 (58%) male. This includes 229 healthy control cases, 402 with MCI and 190 cases with AD. The MCI cases can be subdivided further into 281 stable MCI cases (S-MCI, those with mild-cognitive impairment that do not further develop into AD) and 121 progressive MCI (P-MCI) cases that did develop into AD. Mean age at baseline was 74.9 years (standard deviation 6.9 year.)

For follow-up data, the number of available cases decreases as data is still being collected and obvious drop-out exists in longitudinal studies. The numbers of cases are: 818 (after 6 months), 811 (12 months), 725 (18 months), 446 (24 months), 121 (30 months), and 48(36 months).

Kuopio L-MCI data cohort

This data is collected from a prospective longitudinal study where healthy-to-AD converters are known. The data include MRI data, blood tests and clinical data. The follow-up time has been up to 5 years. Table 2 gives an overview.

At baseline the number of cases was 977 (of which 589 (60%) female and 388 (40%) male). This is distributed as 687 healthy control cases, 249 with MCI and 77 cases with AD. Mean age at baseline was 68.9 years with a standard deviation of 5.0 year. The MCI cases were subdivided further into 176 stable S-MCI cases and 73 P-MCI. Again, in follow-ups this number decreases (861 cases in1st follow-up, 136 in 2nd, 109 in 3rd and 38 cases in 4th follow-up).

TABLE II
CATEGORIES AND NUMBER OF FEATURES IN L-MCI DATA

Basic datarecord information	15
Diagnostic Summary	17
Demographic Information	9
Medical History	68
Concurrent Medications	36
Apolipoprotein E risk gene (ApoE)	8
Cerebrospinal Fluid (CSF)	4
Vital Signs	9
Magnetic Resonance Imaging (MRI)	59
Geriatric Depression Scale (GDS)	1
Heaton Visual Retention Test (HVRT)	13
Wechsler Memory Scale (WMS)	5
Buschke Memory Test (BMT)	19
Activities of daily living inventory (ADCS-ADL)	1
Mini-Mental State Exam (MMSE)	19
CERAD neuropsychological battery (CERAD)	67
Trail Making Test (TMT)	11
Clock Drawing Test (CDT)	9
Wechsler Adult Intelligence Scale (WAIS-R)	3
Blessed Test (BT)	27
Short Portable Mental Status Questionnaire (SPMSQ)	11
Clinical Dementia Rating (CDR)	8
Scent Recognition (SR)	33
Beck Depression Inventory (BDI)	22
Stroop Test (ST)	6

B. *Analysis methods for defining sets of features*

The above described data have been compiled into one common SQL database that can be queried using a dedicated software tool that allows exploration as well as exporting of parts of interest for direct use in data processing software. For data processing, the latest versions of Matlab (The Mathworks Inc., Natick MA), SPSS (SPSS Inc, Chicago IL), Microsoft Excel (Microsoft, Redmond WA) and in-house developed software utilities were used.

Initial statistical analysis included pairwise comparison tests of feature values in different subject groups to assess potential usefulness of features. Different types of data (scalar, ordinal, nominal) require different statistical tests. For scalar and ordinal features, first the Kolmogorov-Smirnov test was used to study if the features are normally distributed. Based on this, either as a t-test or Wilcoxon test was used to compare between different subject groups. The Chi-square goodness-of-fit test was used for comparisons of the nominal features. A significance level of 0.01 was considered as statistically significant. As in this phase the purpose is to explore potentially useful features and not do formal hypothesis testing, no corrections for multiple-testing (such as Bonferroni) were done.

The aim was to find a limited-sized subset of features that can be used in an accurate and generally useful classifier. Three feature selection methods were employed: forward selection, backward selection, and stepwise selection. Forward selection starts with a one-('best')-feature classifier and subsequently adds features; the backward selection approach initially uses all features and then removes features which affect classification performance the least; and stepwise selection is a combination of these two in turns: after removing *n* features, *m* features are added to the feature set. This helps to avoid the so-called 'nesting' issue that is possible in the forward and backward approaches in which features, once added or removed, can not be changed anymore – possibly leading to sub-optimal feature sets. The feature selection procedures produce a set of feature combinations, which are then tested using a separate test set, and the combination producing the best performance is selected.

Prior to the analysis, a validation set was selected by randomly selecting 20% of the subjects. This set was used only at the end in evaluating the final performance of the feature selection and classification.

Two criteria were tested to select the features added/removed. The first criterion was the classification error on a test set (separate from the training set). The feature set producing the lowest classification error was elected for further consideration. The second criterion was so-called minimal-redundancy-maximal-relevance criterion (mRmR) [3]. Its objective is to maximize features dependencies on the class and, at the same time, to minimize the redundancy of the features in the analysis (to not include the same information multiple times in the analysis).

As classification method to assess the performances of different feature sets the support vector machine classifier (SVM) paradigm was used (using libSVM Matlab toolbox, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, [4]). Radial basis functions were used and the parameters were optimized by iterating different values.

For performance assessment, a multi-classifier approach was used. In total 50 optimal feature sets were searched and 50 classifiers were trained using different training and test sets. The 50 training and test sets were obtained by randomly dividing the dataset 50 times so that 75% of the subjects were in the training set and 25% in the test set. The training set was used to compute the mRmR criterion and to train the classifiers. The test set was used to compute the classification error that was then used to select the optimal number of features. After the optimal feature sets were obtained for each data set, the classifiers were trained and then applied to the validation set. Finally, a voting rule was applied to the classifications of all the classifiers, giving the final classification result.

Since a large portion of the data is not yet available in its final form for the L-MCI data (not all MRI features, metabolomics/protoeomics etc are available yet), a similar detailed classification/performance analysis is feasible only when these data become available. To get an idea of the potential performance of the set though, linear models were developed using the features available and using a stepwise feature selection algorithm. An assessment of the performance was then made using a separate validation set (containing 20% of the original data). The history of the performance increase during the stepwise selection process was examined and it was explored which features are deemed the most important in a linear model – which, although not giving optimal classification performance, gives a good indication of which features are important.

III. Results

A. Feature set analysis

The separate testing results of individual features showed that approximately 30% of both the ADNI and the L-MCI features are potentially useful for this task as they have very significantly ($p<0.001$) different values for different disease groups. Some overall indications for the usefulness of different features:

- Age is significantly different between the different groups, but not between P-MCI and S-MCI
- Apolipoprotein E risk gene (ApoE) variables are significantly different for all comparisons
- CSF variables are different between the groups, but not between controls and S-MCI
- Level and duration of education seem especially useful to separate between control and MCI (but not between S-MCI and P-MCI)
- Estrogen usage duration differs for the different groups, but not between P-MCI and S-MCI
- Many MRI features are significantly different for all the tests, especially hippocampus and enthorinal-cortex volume related features
- Geriatric dementia scale did not show differences between groups
- Cognitive and memory tests, such as Wechsler memory scale, Heaton visual retention, Buschke memory, MMSE, CERAD, Clock drawing, Trail making and Blessed and Short Portable tests results differ between different groups, but not between P-MCI and S-MCI

As example, classification results for a SVM classifier using classification error as criterion in stepwise selection for different feature groups in ADNI data are shown in Table III.

TABLE III
CLASSIFICATION PERFORMANCE FOR DIFFERENT FEATURES

Classification task	Features from one single category	Combined features	Most important feature category
AD vs. P-MCI	84 %	94 %	CDR
AD vs. S-MCI	93 %	99 %	MMSE
AD vs. Control	100 %	100 %	CDR, NB
P-MCI vs. S-MCI	81 %	87 %	NB
P-MCIvs. Control	100 %	100 %	CDR
S-MCI vs Control	100 %	100 %	CDR

For abbreviations of feature category names, see Table I

Using a stepwise feature selection approach in which features from any test may be combined gives a model in which the classification performance does not increase after approximately 10 features. These features include, CSF values, number of ApoE-4 alleles, MRI features, age and length of education. Using a linear model including those features gives an impression of the classification potential of them, as depicted in Fig.1.

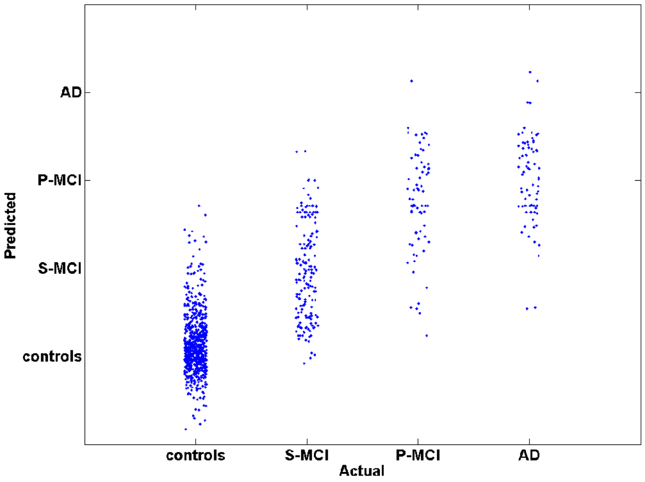

Fig 1. Predicted diagnosis of disease stage vs. actual using a linear prediction model (L-MCI data).

IV. DISCUSSION

Rresults indicate that it is relatively easy to separate controls from healthy cases. However, the clinically more interesting question, to detect differences in people with stable MCI compared to those with progressive MCI is considerably more challenging. This can be seen both in the relatively low accuracies obtained with SVM's on the ADNI data (Table III) as well as the overlap in outputs of a linear model for the L-MCI data (Fig. 1). The initial analyses give an indication that a subset of variables is responsible for the largest part of the classifier performance.

The data sets contain many missing values, especially for later follow-ups. The missing values make the use of dimension reduction techniques (such as principal component analysis) problematic. Therefore, such methods were not applied in this study. Also, the current set of features to be used by a classifier includes CDR – this is a feature used by clinicians to base diagnoses upon – to develop a classifier that is completely 'independent' next feature selections are done without considering CDR.

V. FUTURE STEPS

The data available consists of separate groups of features. For example, features from a memory test or features from MRI establish clear groups. The objective was to study which categories work best in the classification. After this, different categories can be combined to study how much complementary information they include. When costs are associated to performing different tests (like MRI, PET, CSF) it can be estimated how much a certain improvement in accuracy costs, and whether such investment is cost-effective. This is subject of continuing research.

The feature combinations are to be employed in methods implemented in a Clinical Decision Support System (CDSS) which organizes patient data, extracts features and biomarkers, analyzes them statistically against previously diagnosed cases [5]. It is used to characterize newly incoming patient data and can be used to explore the databases. The tool includes, amongst others, a timeline view of a patient's history allowing easy inspection of changes in condition over time. A statistical engine compares patient data to prior cases of corresponding age group and demographics. Combinations of features are then analyzed to identify those which, alone or concurrently, have statistical significance in predicting AD.

REFERENCES

[1] Alzheimer's Disease Neuroimaging Initiative: http://www.loni.ucla.edu/ADNI/

[2] G.B. Frisoni, and M.W. Weiner, "Alzheimer's Disease Neuroimaging Initiative special issue," *Neurobiology of Aging*, Article in press, 2010.

[3] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. PAMI*, vol. 27, no.4, pp. 1226-1238, Aug. 2005.

[4] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using the second order information for training SVM," *Journal of Machine Learning Research*, vol. 6, pp. 1889-1918, Dec. 2005.

[5] J. Mattila, J. Koikkalainen, D. Ververidis, M. van Gils, J. Lötjönen, G. Waldemar, A. Simonsen, D. Rueckert, L. Thurfjell, and H. Soininen, "Clinical decision support system based on statistical analysis of heterogeneous clinical data and Alzheimer's disease biomarkers," Accepted for publication in *Proc.s of the of the ICAD-Alzheimer's Association International Conference on Alzheimer's Disease 2010*, Jul. 2010.