Data mining algorithm predicts Alzheimer's Disease using known and novel biomarkers

Kevin Boehme[1], Ivan Arano[1,*], Ryan Hillary[1], Dan Haskin[2]

1 Deptment of Life Science, Brigham Young University, Provo, Utah, U.S.A

2 Deptment of Computer Science, Brigham Young University, Provo, Utah, U.S.A

∗ **E-mail: Corresponding brujo2204@gmail.com**

Abstract

Alzheimer's disease (AD) is the most common type of dementia that affects about 35 million people worldwide and this number is expected to increase exponentially in the upcoming years. Based on previous studies, we sought to build a robust CDR (clinical dementia rating) prediction model using the ADNI dataset with the advantage of a machine learning algorithm. We cleaned and ran quality control protocols on the ADNI data, and then we binned the dataset using the Ab42/ptau ratio that has consistently been shown to be a useful biomarker of disease progression. We used Weka, a data mining application, to find an optimized prediction model on a subset of the ADNI dataset (training set). We found a model which used as parameters known and novel biomarkers for AD. These markers include measures of cognition, brain measurements, genetic variants, as well as novel blood plasma biomarkers. Testing this simple logistic regression model on a test set we were able to achieve an AUC of .652 for prediction of CDR (0, 0.5, 1) and an AUC of .752 for prediction of the binary phenotype (case/control).

Introduction

Alzheimer's disease (AD) is the most common type of dementia that affects about 35 million people worldwide and this number is expected to increase exponentially in the upcoming years [1]. The most accepted theory for the neurodegeration in Alzheimer's disease is the widespread neuronal death triggered by high levels of a molecule called amyloid beta 42 (AB-42) [2, 3]. Another strong hypothesis for pathogenesis of AD is the mitochondrial cascade hypothesis [4]. This hypothesis claims that neurodegeneration is caused by deleterious changes in the structure and function of the mitochondria. As neuronal metabolism declines due to reduced mitochondrial function, other pathogenic changes are triggered such

as AB-42 and tau phosphorylation [5]. In addition, many other genetic variations in different genes have been associated with increased risk for AD [6, 7]. Of particular notoriety is the APOE e4 allele, which has been documented to provide increased risk by increasing AB-42 deposition [8]. However, mechanisms that provide protection or diagnosis for the disease are still not well understood.

Extensive research has been done in understanding the underlying biology of AD, Genetic factors, brain structural imaging, biomarker levels associated are well characterized to be associated with AD pathology [9, 10]. However, this understanding have provided limited insight in terms of a therapeutics and diagnosis. Identifying individuals with higher propensity of developing AD would be key to ensure better therapeutic strategies and open the door to a more targeted research [11,12]. Currently, there is no effective treatment for the disease once the first symptoms of onset appear. In addition, accurate diagnosis treatments that could track the progression of cognitive decline symptoms in elderly individuals are not available. Thus, it is imperative to identify a model that could effectively identify and track cognitive decline progression.

Current research efforts in AD focus on building tools that could track the AD pathogenesis and the progression of cognitive decline. Yang [13] proposed a simple but powerful model that synchronized sets of data from MCI and AD cases using the ADNI dataset. Delor [14] used the same approach as Yang but using clinical dementia rating scale (CDR-SOB) and identified scenarios of individuals with slow and fast rate of progression. Both models used a robust approach with time series measurements and features highly associated with AD [13,14]. Most recently, Mapstone (2014) develop a promising method by using sets of lipids from peripheral blood. This novel approach predicted conversion from MCI to AD in a period of 2-3 years with a 90% accuracy [15]. Although promising, this model needs to be further tested and confirmed since the phospholipids used have not been previously associated with AD.

Based on these previous studies, we seeked to build a model using a more robust ADNI dataset [14] with the advantage of a machine learning algorithm. This dataset of 800 cases and controls contains several genotypic and phenotypic features that have been associated with Alzheimer's disease. Briefly, our features contain brain measurements from MRI, up to 80 protein biomarkers from CSF and blood plasma, genetic data from associated SNPs, cognitive decline scores, and demographic variables such as age, gender and education level. Our main hypothesis involved building a robust model that could predict cognitive decline progression with a confidence of 85% specificity and sensitivity or predict current diagnosis at a 90% specificity and sensitivity rates.

Materials and Methods

Data Preparation

The data for this manuscript was gathered from current datasets availalble on the Alzheimers Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/). The data available on ADNI has been gathered as a collaborative effort to validate the usefulness of biomarker information, blood test results, and brain measurements in studying the presence, development and physiology of Alzheimer's Disease. With AD being a very complex disease, it was vital to use many different measurements to try to capture as much of the perceived complexity as possible in our model and to observe their effects in detecting AD.

In creating and testing our model we used a dataset that originally consisted of 1117 individuals. These data were very diverse and consisted of measurements of individual portions of the brain, tau proteins, AB-42, and various biomarkers along with age, gender, years of education, APOE genotype and carrier status, and ADAS cognitive total score baseline. We decided to also include SNP data for each of the individuals which were previously found to be significant by Lambert et al. [16] . We included the SNP to observe Among this data we also decided to include the ratio of of tau versus AB-42. This method was discussed by Yang et al. and examined by Delor et al. when they undertook to study and compensate for the pathophysiological timeline of Alzheimers and the differences that exist [13, 14]. This method of compensation for the timeline of AD takes into account the fact that the individuals are possibly in different states in their progression and development of AD. To assess how to group the individuals we simply took the available measurements of AB-42 and divided by the available measurement for tau. (The datasets that did not contain data for tau and AB42 were excluded from this study). These newly calculated ratios were included in the dataset, and from these ratios we were able to group the individuals into three groups. Individuals whose AB-42/tau ratio was less than -.80 were assigned to group 1, individuals whose ratio was between -.79 and -.1 were assigned to group 2, and individuals whose ratio was greater than -.09 were assigned to group 3. Another column that contained the individuals group number was then also added to the dataset.

The dataset was then trimmed as to remove any missing data. Trimming was done using Microsoft Excel. Sadly of the data available, a small proportion of the dataset had individuals whose data contained all the needed fields. We were able to use 96 of the samples from our original dataset after trimming

for missing data. These missing data included, no data for biomarkers, tau, AB42, age, gender, APOE genotype and carrier status and CDR. 459 of the individuals data did not contained missing data for APOE, tau, and AB42. 369 of the individuals did not contain the needed information for Age, years of education, and gender. 191 of the individuals had missing data for one or more biomarkers and were excluded.

Model Selection

Using WEKA, we tried many different data mining algorithms on our data. We decided on using a simple logistic data mining algorithm for two reasons. First, this data mining algorithm yielded a model for prediction of CDR which was interpretable (some yield a model which might accurately predicts CDR model). Second, this algorithm was seemingly resistant to unknown values. Rather than choosing to get rid of unknown values, then, we chose to use an algorithm that could look at all the available information, rather than removing instances with lots of unknown values.

We needed to stratify the data, but with only 30 examples in the full CDR class, it would mean throwing out all but around 30 examples of the other two classes as well. To mitigate this, we first combined the partial and full CDR cases, considering that having full CDR or partial CDR as a patient may well be caused by the same risk factors. After combining the classes, we stratified the data by throwing out instances of the newly formed nontrivial CDR group until the no CDR and nontrivial CDR groups had roughly equal membership within the training set. We first shuffled the data to make sure we threw out extraneous examples of the nontrivial CDR class in a random way.

Results

Model Prediction Results

Our initial three class model seemed to be confused (Table 1). There are many factors that are only considered in a single column which is counter-intuitive. It also appears that this model is skewed. For example, observe the large constant factor in the simple logistic model associated with class 1, partial CDR. This is due to the fact that 70% of our data is on partial CDR patients, and so this constant factor says that a new instance in novel data is likely to be partial CDR a priori since 70% of previously seen examples in the training set were partial CDR as well.

Running our model generator in WEKA again on this new, binary CDR dataset, we obtained the following model (Table 2). In this model we see some striking details:

- Each term listed is used both in determining if the patient has CDR, and also are used to determine if they do not have CDR, which makes more sense.
- Each term's factor listed under "No CDR" is precisely negated in the "Non-trivial CDR" column
- There are a lot of attributes listed that are known correlatives to AD
- There are a lot of attributes listed that are not known correlatives to AD

Discussion

A striking feature of our model is how much improvement in predictive ability is gained when considering a binary (case/control) disease phenotype instead of a trinary CDR (0,0.5,1). Much of the confusion that Weka reported in the trinary model was due to incorrect classifications of CDR 0 and CDR 1 patients to CDR 0.5. Although it was much more apt at correctly reporting a CDR 0 individual than a CDR 1. This was to be expected as many features found in a mild cognitive decline patient (CDR 0.5) would also be seen in a AD patient (CDR 1).

Many of the markers found to be informative in our final model were indeed, known Alzheimer disease markers. These include lsubptau, gender, ADAS cog score, change in ADAS cog score, Entorhinal cortex thickness at baseline, rs2718058, rs10792832. lsubptau and gender are established correlates to AD risk and were expected to be found in the model. It is interesting to note however that AB42 levels were not found in any of the models. Perhaps this is because AB42 is a very early correlate of disease whereas ptau increases are found closer to the point of disease onset. This means that ptau levels could be a more important factor in disease conversion.

The snps rs2718058 and rs10792832 are found in the genes NME8 and PICALM respectively . Previous studies have shown that the recessive A allele of rs10792832 is protective against AD. It is interesting to note that this allele was a significant factor in our model in homozygous alternate allele (AA) carriers (Table 2). rs2718058 was also found to have a protective effect when found in its minor allele form (G). It was a significant factor when found in heterozygotes(AG) (Table 2). Novel biomarkers that were found in the optimized model include: Hepatocyte growth factor, Interleukin 6 receptor, Monokine Induced by

Gamma Interferon, Serum Glutamic Oxaloacetic Transaminase, Thrombomodulin. Interleukin 6 receptor is a protein that plays an important role in the immmune response.

Conclusion

In this study we attempted to leverage current knowledge on AD pathophysiology, the unique ADNI data-set, and past work on AD prediction models to create an accurate model of AD disease progression. After cleaning and binning the dataset by AB42/ptau ratio we used WEKA to compute an optimized logistic regression model. The model we found was somewhat effective at predicting a trinary CDR trait and even more effective at predicting case control status. This model combined known AD biomarkers and novel biomarkers to make its prediction. Testing this simple logistic regression model on a test set we were able to achieve an AUC of .652 for prediction of CDR (0, 0.5, 1) and an AUC of .752 for prediction of the binary phenotype (case/control).

Further investigation is merited for the novel biomarkers that were found to be significant in this study. Future work should incorporate a time syncronization variable due to the long duration of disease progression. In addition, better leveraging of the longitudinal data found in the ADNI dataset could help refine the predictive power of this model.

Acknowledgments

References

1. Geneva (2012:112) WHO: Dementia: A Public Health Priority.

2. Reitz C, Brayne C, Mayeux R (2011) Epidemiology of Alzheimer disease. Nature reviews Neurology 7.

3. Glenner GG, Wong CW (1984) Alzheimer's disease: initial report of the purificationand characterization of a novel cerebrovascular amyloid protein. Biochem Biophys Res Commun 120.

4. Swerdlow R, Khan S (2004) A mitochondrial cascade hypothesis for sporadic Alzheimer's disease. Med Hypotheses 63: 8–20.

5. Ridge P, Maxwell T, Corcoran C, Norton M, Tschanz J, et al. (2012) Mitochondrial genomic analysis of late onset alzheimer's disease reveals protective haplogroups h6a1a/h6a1b: The cache county study on memory in aging. PLoS ONE .

6. Jun G, Naj A, Beecham G, Wang L, Buros J, et al. (2010) Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. Arch Neurol 67: 1473–1484.

7. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J, et al. (2011) Common variants at ABCA7, MS4a6a/ms4a4e, EPHA1, CD33 and CD2ap are associated with Alzheimer's disease. Nat Genet 43: 429–435.

8. Mahley R, et al. (2009) Apolipoprotein E: structure determines function, from atherosclerosis to Alzheimer's disease to AIDS. Journal of lipid research .

9. Hulstaert F, et al. (1999) Improved discrimination of AD patients using beta-amyloid 42 and tau levels in CSF. Neurology 52.

10. Klunk WE, et al. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. Ann Neurol 55.

11. Swainson R, et al. (2001) Early detection and differential diagnosis of alzheimer's disease and depression with neuropsychological tasks. Dement Geriatr Cogn Disord 12:265-280 (DOI:10.1159/000051269).

12. Bennett P (2003) Early diagnosis of alzheimer's disease: Clinical and economic benefits. Journal of the American Geriatrics Society 51.

13. Yang E, et al. (2011) Quantifying the pathophysiological timeline of alzheimer's disease. Journal of Alzheimer's Disease 26:745-753 745.

14. Delor I, et al. Modeling Alzheimer's Disease Progression Using Disease Onset Time and Disease Trajectory Concepts Applied to CDR-SOB Scores From ADNI. CPT Pharmacometrics Syst Pharmacol 2, e78; doi:10 2.

15. Mapstone M, et al. Plasma phospholipids identify antecedent memory impairment in older adults. Nature Medicine.

16. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics .

Figure Legends

Tables

Table 1. Logistic Regression Three Class Model

	No CDR	Partial CDR	Full CDR
Constant Term	-0.37	1.84	0.9
% Change ADAS Cognative Score	-0.26	0	0.09
Baseline Entorhinal Cortex Thickness	0.42	0	0
Baseline Left Hippocampal Volume	0	-5.56	0
Hepatocyte Growth Factor (HGF) (ng/mL)	0	2.19	0
1 if rs2718058 is GG, 0 otherwise	0	-0.33	0
1 if rs4147929 is AG, 0 otherwise	0	-0.26	0
Baseline Parahippocampal Cortex Thickness	0	0	-0.76
1 if rs7274581 is TT, 0 otherwise	0	0	-0.42

Table 2. Logistic Regression Two Class Model

	No CDR	CDR
Constant Term (a priori bias)	-2.81	2.81
lsubptau	1.48	-1.48
1 if gender is female, 0 otherwise	0.3	-0.3
Baseline ADAS Total Cognative Score	-0.02	0.02
% Change in ADAS Total Cognative Score	-0.14	0.14
Baseline Entorhinal Cortex Thickness	0.46	-0.46
Hepatocyte Growth Factor (HGF) (ng/mL)	-1.69	1.69
Interleukin 6 receptor (IL 6r) (ng/mL)	2.59	-2.59
Monokine Induced by Gamma Interferon (MI) (pg/mL)	0.76	-0.76
Serum Glutamic Oxaloacetic Transaminase (ug/mL)	-1.51	1.51
Thrombomodulin (TM) (ng/mL)	1.32	-1.32
1 if rs2718058 is GA, 0 otherwise	-0.23	0.23
1 if rs10792832 is AA, 0 otherwise	0.31	-0.31

Table 3. Training Set: Model Predicition Results

	Correctly Classified Instance	Incorrectly Classified Instance
Trinary	67.031%	32.969%
Binary	71.3287%	28.6713%

Table 4. Test Set: Model Predicition Results

	ROC Area	Correctly Classified Instance	Incorrectly Classified Instance
Trinary	0.652	57.7465%	42.2535%
Binary	0.752	69.0141%	30.9859%