Predicting Future Incidence of Alzheimer's Disease Using Electronic Health Records

Ji Hwan Park[1], Jiook Cha[2], Shinjae Yoo[1]

[1]Computational Science Initiative
Brookhaven National Laboratory

[2]Department of Psychiatry
Vagelos College of Physicians and Surgeons;
Center for Computing Systems for Data Driven Science
Data Science Institute
Columbia University

Data Science Institute COLUMBIA UNIVERSITY

COLUMBIA | PSYCHIATRY

Computing Systems for Data-Driven Science

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Motivation

Recently, the amount of electronic health system (EHR) data has surged. EHR data contains complementary information associated with patients such as demographics, diagnosis codes, medication codes, and lab results, which is sparse, high-dimensional, and temporal information [1]. This rich, abundant EHR may give us an opportunity for predictive modeling in medicine that could lead to early detection and intervention of debilitating diseases, such as Alzheimer's disease (AD). In the AD literature, such a predictive model has yet to be reported. Here we applied data-driven machine learning techniques to EHR data for predicting Alzheimer's disease.

EHR data: Korean NHIS-National Sample Cohort

We used Korea health insurance data from Korean Health Insurance Review and Assessment Service (KHIRAS), which contains demographics, disease and medication codes, and physical exam records from 1,086,516 people. The data was collected from 2002 and 2010.

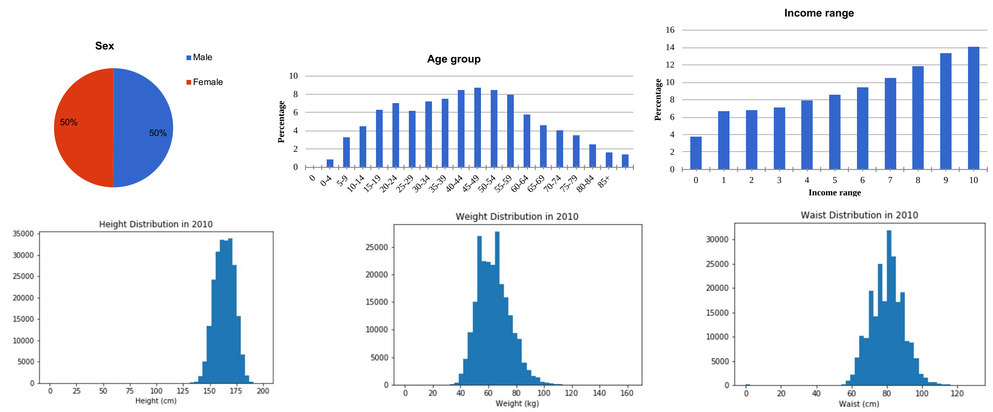

Figure 1. Distribution of some features from Korean NHIS-National Sample Cohort.

Experiment

The goal of our experiment is to predict Alzheimer's disease (AD) 1 year prior to the initial AD diagnosis. To achieve this goal, first we identified 2,049 unique individuals diagnosed with AD with age>=65yr. Since our data is imbalanced, we performed a random sampling to have balanced data, totaling 4,098 individuals (mean age: AD=78±6.1; non-AD=74±5.3). The total number of unique features was 4,938, and the average number of disease and medication code is 128.3±127.7 for AD and in 191.5±68 for non-AD. We applied three popular machine learning techniques (logistic regression, support vector machine, random forest) to our EHR data. For model validation, we used 5×5 nested cross-validation.

Experimental Result

(): standard deviation

	Avg. Accuracy	Avg. Sensitivity	Avg. Specificity	AUC
Logistic Regression	0.883 (0.009)	0.856 (0.023)	0.911 (0.012)	0.884
Support Vector Machine	0.880 (0.013)	0.827 (0.024)	0.932 (0.006)	0.880
Random Forest	0.905 (0.011)	0.858 (0.014)	0.953 (0.010)	0.906

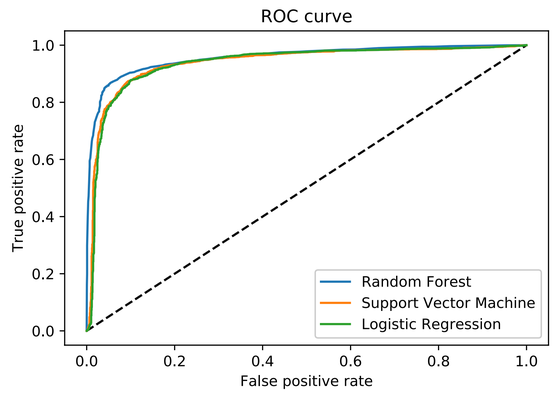

Figure 2. Our experimental results: averaged (avg.) accuracy, sensitivity and specificity and AUC (top), and ROC curves (bottom).

Important Features

Coef.: coefficient
Db: database

feature	coef.	Db
hemoglobin	-0.26	screen
LDL cholesterol	-0.25	screen
waist	-0.21	screen
triglyceride	-0.20	screen
HDL cholesterol	-0.19	screen
age	0.16	eligibility
amoxicillin	-0.08	utilization
smoking	0.10	screen
Essential hypertension	-0.08	utilization
urine glucose	0.06	screen

Table 1. Top features contributing to AD prediction from logistic regression.

Conclusion

We explored the possibility of machine learning techniques using EHR data to predict Alzheimer's disease (AD). Our results show accurate prediction of one-year subsequent incidence of AD. This suggests the utility of both Korean EHR, as well as data-driven machine learning frameworks for predictive modeling in AD. FuIn future, we will test model generalizability on independent data.

Acknowledgments

National Institute of Mental Health (K01-MH109836, PI-Cha), Brain and Behavior Research Foundation (Young Investigator Award, PI-Cha), Korean American Scientists and Engineers Association (Young Investigator Grant, PI-Cha).

References

1. Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu, "Risk Prediction with Electronic Health Records: A Deep Learning Approach." SIAM International Conference on Data Mining, 432–440, 2016