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

Ji Hwan Park¹, Jiook Cha², Shinjae Yoo²

¹Computational Science Initiative Brookhaven National Laboratory
²Department of Psychiatry Vagelos College of Physicians and Surgeons; Center for Computing Systems for Data Driven Science Data Science Institute Columbia University

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.

Experimental Result

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg. Accuracy</th>
<th>Avg. Sensitivity</th>
<th>Avg. Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.883 (0.009)</td>
<td>0.856 (0.023)</td>
<td>0.911 (0.012)</td>
<td>0.884</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.880 (0.013)</td>
<td>0.827 (0.024)</td>
<td>0.932 (0.006)</td>
<td>0.880</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.905 (0.011)</td>
<td>0.858 (0.014)</td>
<td>0.953 (0.010)</td>
<td>0.906</td>
</tr>
</tbody>
</table>

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

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

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. Fuln 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