Electronic Health Records Based Prediction of Future Incidence of Alzheimer’s Disease Using Machine Learning

Ji Hwan Park\textsuperscript{a}, Han Eol Cho\textsuperscript{b}, Jong Hun Kim\textsuperscript{g}, Melanie Wall\textsuperscript{d}, Yaakov Stern\textsuperscript{c,d}, Hyunsun Lim\textsuperscript{f}, Shinjae Yoo\textsuperscript{a}, Hyoung-Seop Kim\textsuperscript{g}, Jiook Cha\textsuperscript{dh}

\textsuperscript{a} Computational Science Initiative, Brookhaven National Laboratory, Upton, New York, USA; 
\textsuperscript{b} Department of Rehabilitation Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea; 
\textsuperscript{c} Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea; 
\textsuperscript{d} Department of Psychiatry, Columbia University, New York, USA; 
\textsuperscript{e} Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University, New York, USA; 
\textsuperscript{f} Research and Analysis Team, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea; 
\textsuperscript{g} Department of Physical Medicine and Rehabilitation, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea; 
\textsuperscript{h} Data Science Institute, Columbia University, New York, USA.
Affordable EHR for Screening Alzheimer’s disease (AD)

- Biomarkers - the collection of bio-specimen (e.g., serum or fluid) or imaging data
  - Time consuming

- Electronic health records (EHR)
  - not require additional time or effort for data collection
  - Increase the size of EHR data due to digitalization
Overview of EHR
A few predefined features

• In prior work, predefined features
  • sociodemographic (age, sex, education)
  • lifestyle (physical activity)
  • midlife health risk factors (systolic blood pressure, BMI and total cholesterol level)
  • cognitive profiles

• Multi-factor models best predict risk for dementia

⇒ Machine learning
Machine learning on high-dimensional EHR

- Use a large nationally representative (South Korea) sample cohort

- Construct and validate data-driven machine learning models to predict future incidence of AD using the extensive measures collected within high-dimensional EHR

- Demonstrate the feasibility of developing accurate prediction models for AD
Korean EHR data

• Korean National Health Insurance Service - National Elderly cohort Database

• 6,435 features

• 430,133 individuals (> 65 yrs, 10% sample of randomly selected elderly individuals)

• 2002 – 2010, South Korea
# High-dimensional Features

**National Elderly cohort Database (DB)**

<table>
<thead>
<tr>
<th>Health Screening (HS) DB</th>
<th>Participant Insurance Eligibility (PIE) DB</th>
<th>Healthcare Utilization (HU) DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>21 Features: laboratory values, health profiles, history of family illness</td>
<td>2 Features: sex, age</td>
<td>6,412 features including ICD-10 codes and medication codes</td>
</tr>
</tbody>
</table>
Machine learning analysis

• Input: High-dimensional EHR data

• Methods
  • Random forest, support vector machine (SVM), logistic regression

• Task: Can machine learning be used to predict future incidence of Alzheimer’s disease using electronic health records?
Definition of data

• Two criteria
  • (Korean) ICD-10 code:
    • Dementia in AD - F00, F00.0, F00.1, F00.2, F00.9
    • AD - G30, G30.0, G30.1, G30.8, G30.9
  • Dementia medication: e.g., donepezil, rivastigmine, galantamine, and memantine

• **Definite AD**: ICD-10 code + medication

• **Probable AD**: only ICD-10
Data range for n-year prediction

• AD group: between 2002 and the year of incident AD – n
• Non-AD group: 2002 to 2010 – n

Example:

Non-AD: 2002 2003 2004 2005 2006 2007 2008 2009 2010
AD (1yr): 2002 2003 2004 2005 2006 2007 2008 2009 2010
AD (2yr): 2002 2003 2004 2005 2006 2007 2008 2009 2010
AD (3yr): 2002 2003 2004 2005 2006 2007 2008 2009 2010
AD (4yr): 2002 2003 2004 2005 2006 2007 2008 2009 2010

AD

Non-AD:
Data Preprocessing

• EHR alignment

• ICD-10 and medication coding
  • the first disease category codes: e.g., F00.0
  • the first 4 characters for the medication codes representing main ingredients: e.g., 149801ATB

• Rare disease exclusion (\(\leq 5\))

• Records exist in all the three databases (HS, PIE,HU)
# of data samples

- **430,133** elderly individual
- **389,349** excluded
  - Not in HS DB
- **40,736** individual
  - In all DBs
- **Probable AD: 2,026** individual
  - Definite AD: **614** individual
- **Non-AD: 38,710** individual
# Sample characteristics

<table>
<thead>
<tr>
<th></th>
<th>Definite AD</th>
<th>Probable AD</th>
<th>Non-AD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td>614</td>
<td>2,026</td>
<td>38,710</td>
</tr>
<tr>
<td><strong>Income</strong></td>
<td>$ 60k ($57.3k-$62.7k)</td>
<td>$59k ($58.7k-$59.3k)</td>
<td>$60.2k ($58.7k-$61.7k)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>80.67 (80.2-81.1)</td>
<td>79.2 (79.0-79.5)</td>
<td>74.5 (74.4-74.5)</td>
</tr>
<tr>
<td><strong>sex</strong></td>
<td>Male:229 (37%)</td>
<td>Male:733 (36%)</td>
<td>Male:18,200 (47%)</td>
</tr>
<tr>
<td></td>
<td>Female:285 (63%)</td>
<td>Female:1,293 (64%)</td>
<td>Female:20,510 (53%)</td>
</tr>
</tbody>
</table>

*Based on the 0-year prediction model.*
N-year prediction for definite AD
N-year prediction for probable AD

![Graph showing the AUC over different year predictions for probable AD.](image)
Model prediction result - ROC

Receiver-Operating Characteristics

Definite AD

Probable AD

Sensitivity

Specificity

0 yr
1 yr
2 yr
3 yr
4 yr

0 yr
1 yr
2 yr
3 yr
4 yr
## Important features

<table>
<thead>
<tr>
<th>Name</th>
<th>b value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemoglobin (H)</td>
<td>-0.902</td>
</tr>
<tr>
<td>Age (Demo)</td>
<td>0.689</td>
</tr>
<tr>
<td>Urine protein (H)</td>
<td>0.303</td>
</tr>
<tr>
<td>Zotepine (antipsychotic drug) (M)</td>
<td>0.303</td>
</tr>
<tr>
<td>Nicametate Citrate (vasodilator) (M)</td>
<td>-0.297</td>
</tr>
<tr>
<td>Other degenerative disorders of nervous system in diseases</td>
<td>-0.292</td>
</tr>
<tr>
<td>classified elsewhere (D)</td>
<td></td>
</tr>
<tr>
<td>Disorders of external ear in diseases classified elsewhere (D)</td>
<td>0.274</td>
</tr>
<tr>
<td>Tolfenamic acid 200mg (pain killer) (M)</td>
<td>0.266</td>
</tr>
<tr>
<td>Adult respiratory distress syndrome (D)</td>
<td>-0.259</td>
</tr>
<tr>
<td>Eperisone Hydrochloride (antispasmodic drug) (M)</td>
<td>0.255</td>
</tr>
</tbody>
</table>

(H): Health checkup  
(M): Medication  
(Demo): Demographics  
(D): Disease
Summary (1)

- Our model AUC: 0.887 (0yr), 0.781 (1yr), 0.662 (4yr)
- Prior models AUC: 0.5 ~ 0.78
- Detected interesting EHR-based features associated with incident AD
Summary (2)

- Presents the first data in predicting future incident AD using data-driven machine learning based on large-scale EHR

- Support to the development of EHR-based AD risk prediction that may enable better selection of individuals at risk for AD in clinical trials or early detection in clinical settings
Future work

• Generalize our findings to ethnicities other than Korean or to different healthcare systems

• Apply deep neural networks such as a recurrent neural network (RNN)
# Model prediction results (1)

<table>
<thead>
<tr>
<th></th>
<th>Classifier*</th>
<th>AD/non-AD</th>
<th>AUC</th>
<th>Sensitivity** (when 90% specificity)</th>
<th>Specificity** (when 90% Sensitivity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 yr</td>
<td>RF</td>
<td>614/38,710</td>
<td>0.887</td>
<td>0.687</td>
<td>0.737</td>
</tr>
<tr>
<td>1 yr</td>
<td>SVM</td>
<td>672/38,967</td>
<td>0.781</td>
<td>0.380</td>
<td>0.475</td>
</tr>
<tr>
<td>2 yr</td>
<td>SVM</td>
<td>640/38,605</td>
<td>0.739</td>
<td>0.281</td>
<td>0.400</td>
</tr>
<tr>
<td>3 yr</td>
<td>SVM</td>
<td>605/29,983</td>
<td>0.686</td>
<td>0.227</td>
<td>0.291</td>
</tr>
<tr>
<td>4 yr</td>
<td>RF</td>
<td>491/14,196</td>
<td>0.662</td>
<td>0.000</td>
<td>0.151</td>
</tr>
</tbody>
</table>

*best classifiers based on AUC. **closest values with sensitivity or specificity set to 90%. LR, logistic regression; RF, random forest; SVM, support vector machine
## Model prediction results (2)

<table>
<thead>
<tr>
<th>Probable AD (AD codes)</th>
<th>Classifier*</th>
<th>AD/non-AD</th>
<th>AUC</th>
<th>Sensitivity** (when 90% specificity)</th>
<th>Specificity** (when 90% Sensitivity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 yr</td>
<td>RF</td>
<td>2,026/38,710</td>
<td>0.805</td>
<td>0.240</td>
<td>0.456</td>
</tr>
<tr>
<td>1 yr</td>
<td>RF</td>
<td>2,049/38,967</td>
<td>0.730</td>
<td>0.170</td>
<td>0.338</td>
</tr>
<tr>
<td>2 yr</td>
<td>LR</td>
<td>1,892/38,605</td>
<td>0.645</td>
<td>0.136</td>
<td>0.301</td>
</tr>
<tr>
<td>3 yr</td>
<td>LR</td>
<td>1,697/29,983</td>
<td>0.575</td>
<td>0.085</td>
<td>0.253</td>
</tr>
<tr>
<td>4 yr</td>
<td>RF</td>
<td>1,412/14,196</td>
<td>0.602</td>
<td>0.020</td>
<td>0.018</td>
</tr>
</tbody>
</table>

*best classifiers based on AUC. **closest values with sensitivity or specificity set to 90%. LR, logistic regression; RF, random forest; SVM, support vector machine