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

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

a. Computational Science Initiative, Brookhaven National Laboratory, Upton, New York, USA;
b. Department of Rehabilitation Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea;
c. Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
d. Department of Psychiatry, Columbia University, New York, USA;
e. Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University, New York, USA;
f. Research and Analysis Team, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
g. Department of Physical Medicine and Rehabilitation, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
h. Data Science Institute, Columbia University, New York, USA.

Affordable EHR for Screening Alzheimer's disease (AD)

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

  ➔ Time consuming

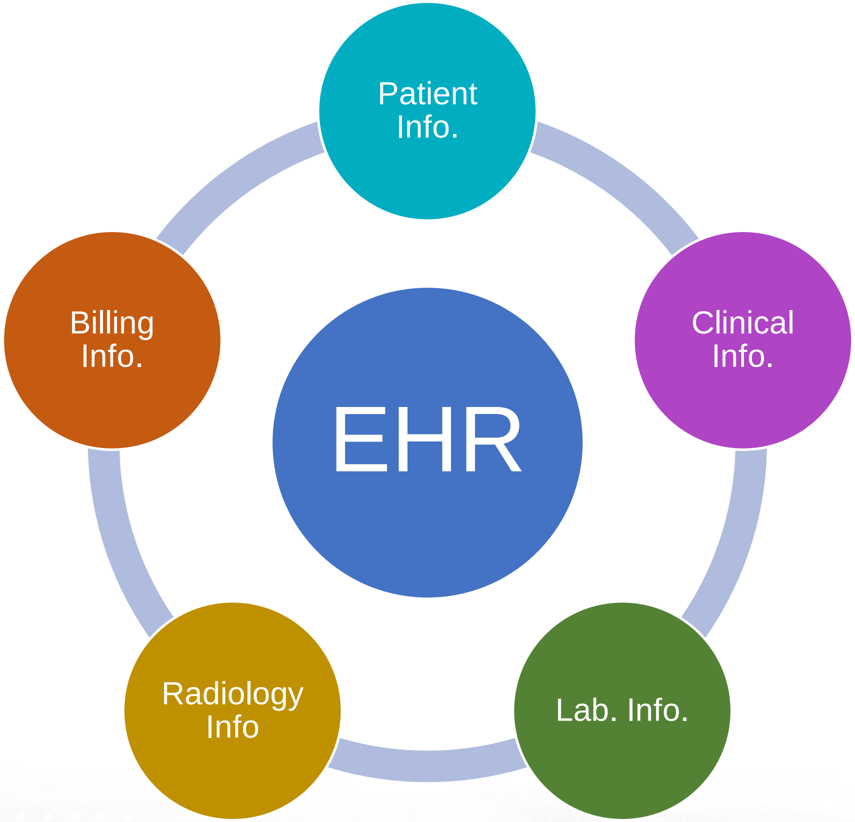
- Electronic health records (EHR)
  - **not require additional time or effort** for data collection
  - Increase the size of EHR data due to digitalization

U.S. DEPARTMENT OF ENERGY

COLUMBIA PSYCHIATRY

BROOKHAVEN NATIONAL LABORATORY

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)

Health Screening (HS) DB	Participant Insurance Eligibility (PIE) DB	Healthcare Utilization (HU) DB
21 Features: laboratory values, health profiles, history of family illness	2 Features: sex, age	6,412 features including ICD-10 codes and medication codes

U.S. DEPARTMENT OF ENERGY

COLUMBIA PSYCHIATRY

BROOKHAVEN NATIONAL LABORATORY

Machine learning analysis

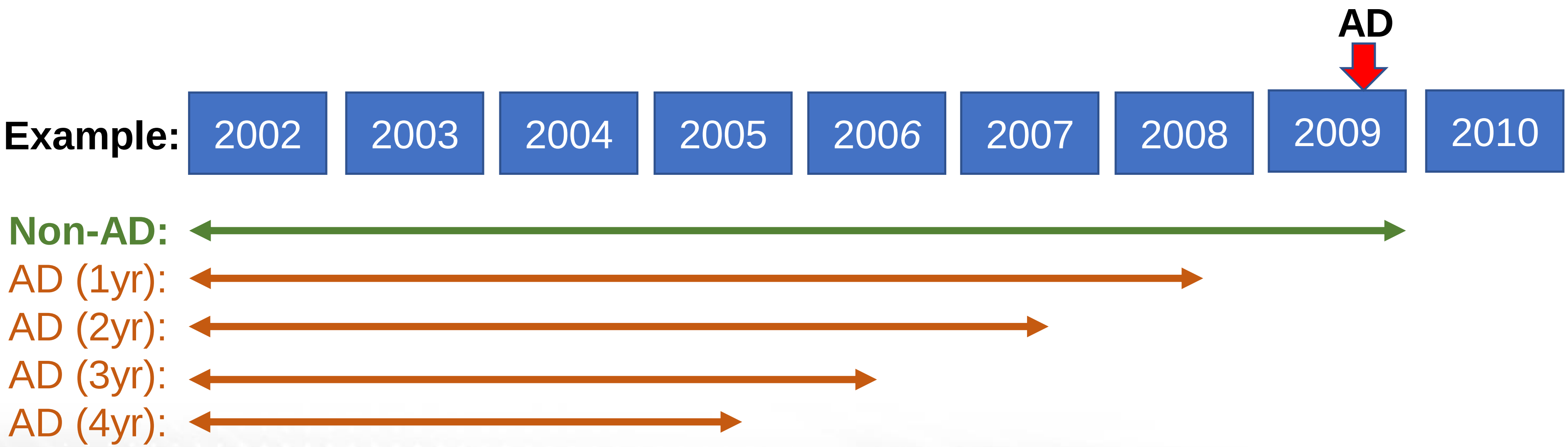
- Input: High-dimensional EHR data
- Methods
  - Random forest, support vector machine (SVM), logistic regression
- Task: Can machine learning be used to predict future incidence of Alzheimer's disease using electronic health records?

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*



U.S. DEPARTMENT OF ENERGY
COLUMBIA PSYCHIATRY
BROOKHAVEN NATIONAL LABORATORY

Data Preprocessing

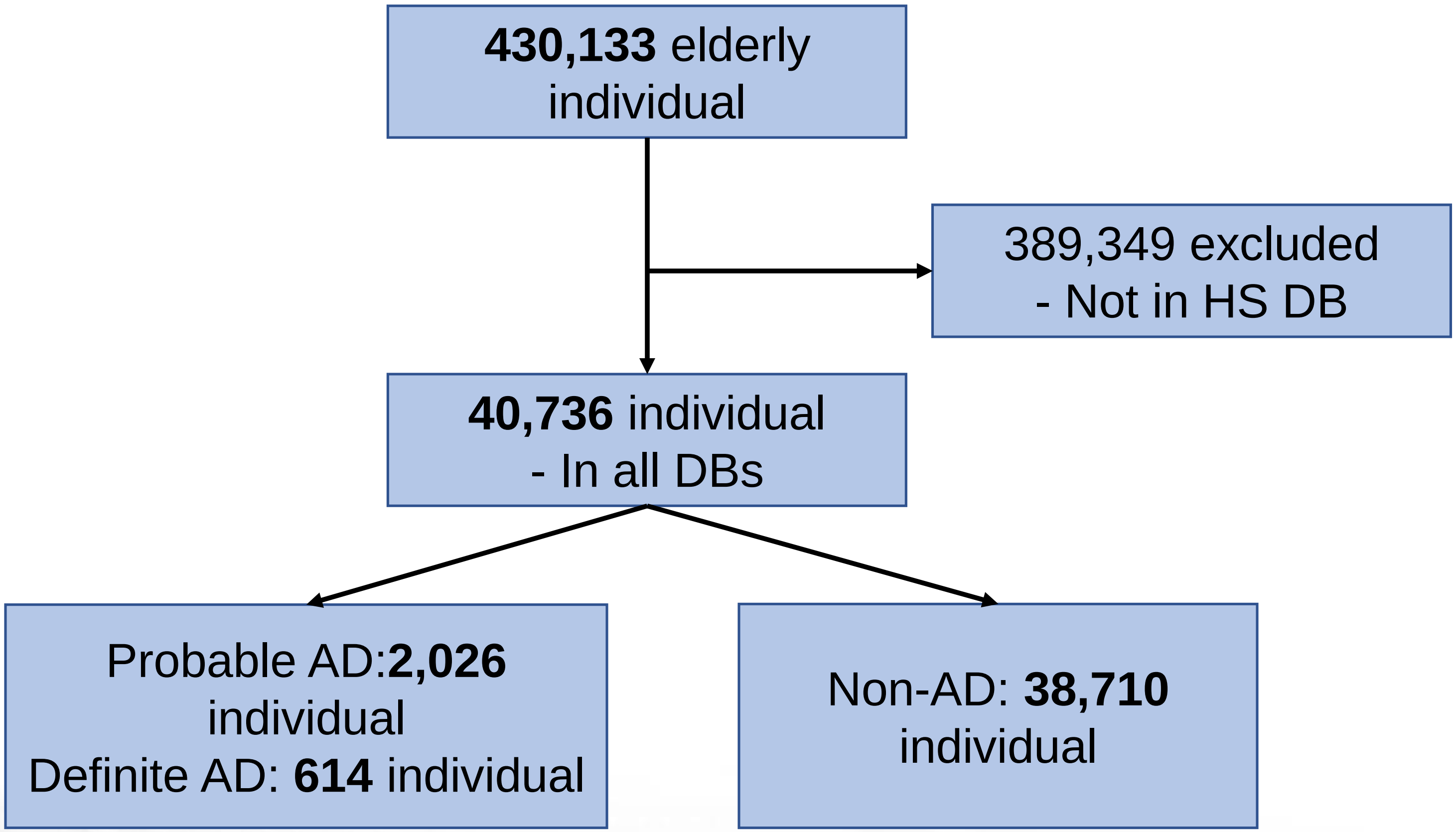
- EHR alignment
- ICD-10 and medication coding
  - the first disease category codes: e.g., **F00**.0
  - the first 4 characters for the medication codes representing main ingredients: e.g., **1498**01ATB
- Rare disease exclusion (≤ 5)
- Records exist in all the three databases (HS, PIE,HU)

U.S. DEPARTMENT OF ENERGY

COLUMBIA PSYCHIATRY

BROOKHAVEN NATIONAL LABORATORY

of data samples

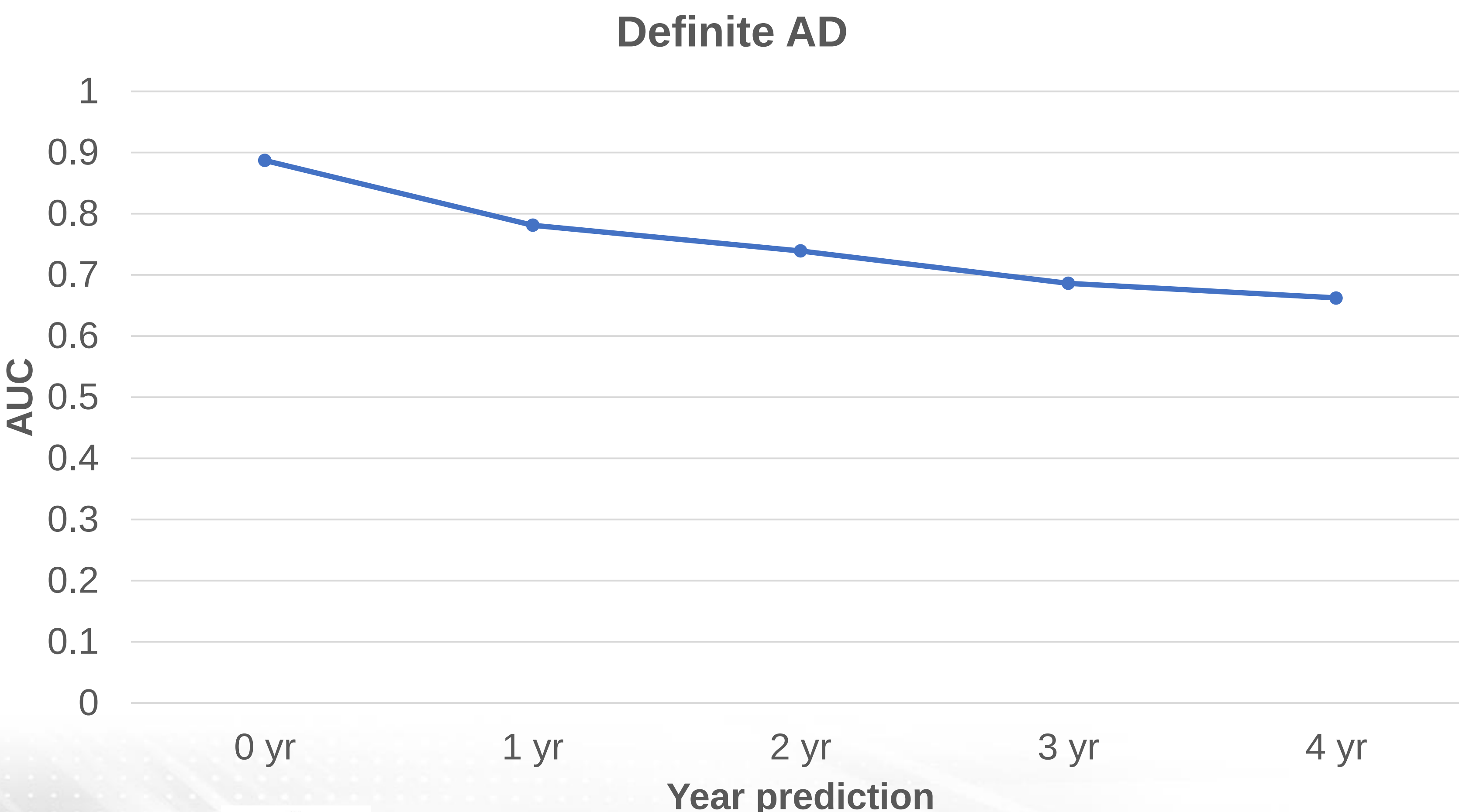


Sample characteristics

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

*Based on the 0-year prediction model.

N-year prediction for definite AD

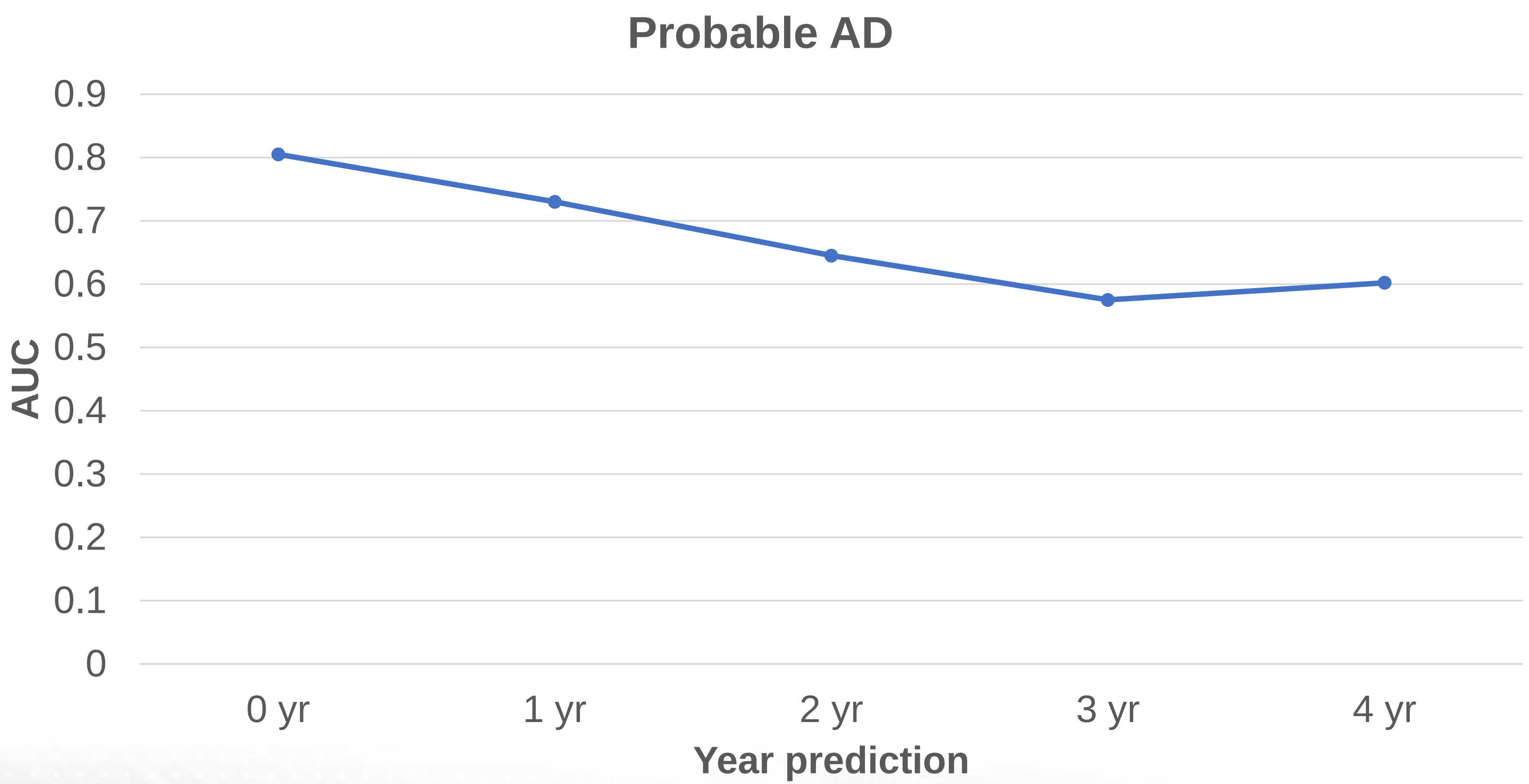


U.S. DEPARTMENT OF ENERGY

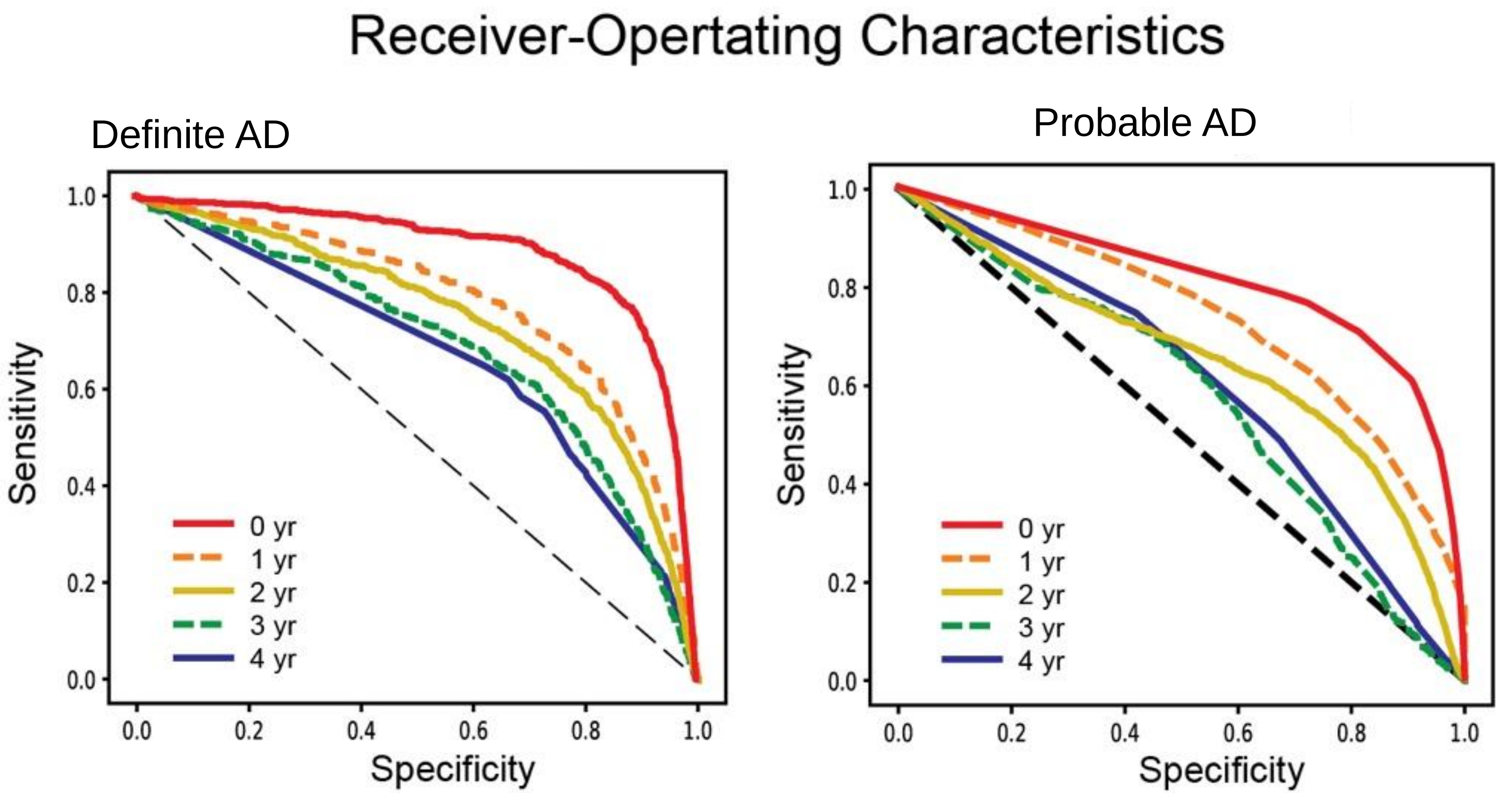
COLUMBIA PSYCHIATRY

BROOKHAVEN NATIONAL LABORATORY

N-year prediction for probable AD



Model prediction result - ROC



Important features

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

(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

U.S. DEPARTMENT OF ENERGY

COLUMBIA PSYCHIATRY

BROOKHAVEN NATIONAL LABORATORY

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)

Definite AD (AD codes and dementia prescription)					
	Classifier*	AD/non-AD	AUC	Sensitivity** (when 90% specificity)	Specificity** (when 90% Sensitivity)
0 yr	RF	614/38,710	**0.887**	0.687	0.737
1 yr	SVM	672/38,967	**0.781**	0.380	0.475
2 yr	SVM	640/38,605	**0.739**	0.281	0.400
3 yr	SVM	605/29,983	**0.686**	0.227	0.291
4 yr	RF	491/14,196	**0.662**	0.000	0.151

*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)

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

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