Word count: 2,740

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

Ji Hwan Park, PhD[1]*, Han Eol Cho, MD[2]*, Jong Hun Kim, MD, PhD[3], Melanie Wall, PhD[4], Yaakov Stern, PhD[4,5], Hyunsun Lim, PhD[6], Shinjae Yoo, PhD[1], Hyoung-Seop Kim, MD[7], Jiook Cha, PhD[4,8] (*equally contributed)

1. Computational Science Initiative, Brookhaven National Laboratory, Upton, New York, USA; 2. Department of Rehabilitation Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea;
3. Department of Neurology, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
4. Department of Psychiatry, Columbia University, New York, USA;
5. Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University, New York, USA;
6. Research and Analysis Team, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
7. Department of Physical Medicine and Rehabilitation, Dementia Center, National Health Insurance Service Ilsan Hospital, Goyang, Republic of Korea;
8. Data Science Institute, Columbia University, New York, USA.

Correspondence to:

Hyoung-Seop Kim, MD
rekhs@nhimc.or.kr
100 ilsan-ro ilsan-donggu Goyang City Gyeoggi Do, 10444, Republic of Korea

Jiook Cha, PhD
jc4248@cumc.columbia.edu
1051 Riverside Dr.
New York, NY, 10033, USA

Key Points

Question Can machine learning be used to predict future incidence of Alzheimer's disease using electronic health records?

Findings We developed and validated supervised machine learning models using the EHR data from 40,736 South Korean elders (age above 65 years old). Our model showed acceptable accuracy in predicting up to four year subsequent incidence of AD.

Meaning This study shows the potential utility of the administrative EHR data in predicting risk for AD using data-driven machine learning to support physicians at the point of care.

Abstract

Background: Accurate prediction of future incidence of Alzheimer's disease may facilitate intervention strategy to delay disease onset. Existing AD risk prediction models require collection of biospecimen (genetic, CSF, or blood samples), cognitive testing, or brain imaging. Conversely, EHR provides an opportunity to build a completely automated risk prediction model based on individuals' history of health and healthcare. We tested machine learning models to predict future incidence of AD using administrative EHR in individuals aged 65 or older.

Methods: We obtained de-identified EHR from Korean elders age above 65 years old (N=40,736) collected between 2002 and 2010 in the Korean National Health Insurance Service database system. Consisting of Participant Insurance Eligibility database, Healthcare Utilization database, and Health Screening database, our EHR contain 4,894 unique clinical features including ICD-10 codes, medication codes, laboratory values, history of personal and family illness, and socio-demographics. Our event of interest was new incidence of AD defined from the EHR based on both AD codes and prescription of anti-dementia medication. Two definitions were considered: a more stringent one requiring a diagnosis and dementia medication resulting in n=614 cases ("definite AD") and a more liberal one requiring only diagnostic codes (n=2,026; "probable AD"). We trained and validated a random forest, support vector machine, and logistic regression to predict incident AD in 1,2,3, and 4 subsequent years using the EHR available since 2002. The length of the EHR used in the models ranged from 1,571 to 2,239 days. Model training, validation, and testing was done using iterative (5 times), nested, stratified 5-fold cross validation.

Results: Average duration of EHR was 1,936 days in AD and 2,694 days in controls. For predicting future incidence of AD using the "definite AD" outcome, the machine learning models showed the best performance in 1 year prediction with AUC of 0.781; in 2 year, 0.739; in 3 year, 0.686; in 4 year, 0.662. Using “probable AD” outcome, the machine learning models showed the best performance in 1 year prediction with AUC of 0.730; in 2 year, 0.645; in 3 year, 0.575; in 4 year, 0.602. Important clinical features selected in logistic regression included hemoglobin level (b=-0.902), age (b=0.689), urine protein level (b=0.303), prescription of Lodopin (antipsychotic drug) (b=0.303), and prescription of Nicametate Citrate (vasodilator) (b=-0.297).

Conclusion: This study demonstrates that EHR can detect risk for incident AD. This approach could enable risk-specific stratification of elders for better targeted clinical trials.

Introduction

Screening individuals at risk for Alzheimer's disease (AD) based on medical health records in preclinical stages may lead to more widespread early detection of AD pathology and ultimately to better therapeutic strategies for delaying the onset of AD.[1-3] In contrast to biomarkers requiring the collection of bio-specimen (e.g., serum or fluid) or imaging data, electronic health records (EHR) does not require additional time or effort for data collection. Furthermore, with advent of digitalization, the amounts of the EHR available for predictive modeling have exponentially increased.[4] Because it is ubiquitous and affordable, developing risk prediction of AD using the EHR will have a great impact on the AD research and clinical care. However, despite of the tremendous potential value of EHR-based predictive models, little is known about the utility of such models for AD screening.

For population AD screening, prior models are based on predefined features including health profiles, such as sociodemographic (age, sex, education), lifestyle (physical activity), midlife health risk factors (systolic blood pressure, BMI and total cholesterol level);[5,6] and cognitive profiles.[7,8] Despite of the demonstrated accuracy of these models, an important outstanding question is whether the several curated variables may sufficiently account for the heterogeneous etiology of multi-factorial AD. Indeed, a meta-analysis study shows that multi-factor models best predict risk for dementia, whereas single-factor models do poorly,[6] suggesting accurate AD screening with practical utility in large populations require sufficiently large feature space. An important new approach for developing individualized predictive modeling is the use of the rigorous data-driven machine learning that can harvest salient information from large-scale EHR to make an individual-specific prediction.

Machine learning is an optimal choice of the analytic method for analyzing large-scale EHR containing thousands of descriptors in hundreds of thousands of individuals. Studies show successful application of machine learning to the EHR in predicting incident diseases (cancer, diabetes, schizophrenia, etc) or mortality.[9-12] Given the recent rapid growth of the machine learning technology, application of the AI technology to clinical predictive modeling is likely to have a deep impact on medicine.[13-15] But to our knowledge data-driven predictive modeling with EHR data has not been previously used to predict incident AD.

When developing machine learning models, it is important to use sufficiently large data representative of a target population of interest. The size and breadth of the data is important for model precision, while the representativeness of the data is important for minimizing potential bias an improving generalizability. In the present study, we use a large nationally representative (South Korea) sample cohort taken from the Korean National Health Insurance Service database.[16] We construct and validate data-driven machine learning models to predict future incidence of AD using the extensive measures collected within the EHR. We demonstrate the feasibility of developing accurate prediction models for AD which may then provide a starting point for future.

Materials and Methods

Datasets

We used the National Health Insurance Service (NHIS)-National Elderly cohort Database, a subsample of the National Health Insurance Service-national sample cohort.[17] This database contains for each individual features of services, diagnoses, and prescriptions associated with all the health care services provided by the NHIS. All EHR was binned monthly. Clinical features include demographics and socioeconomics from the *Participant Insurance Eligibility database*; disease and medication codes from the *Healthcare Utilization database*; and laboratory values, health profiles, and history of personal and family illness from the *National Health Screening database* (from bi-annual health check-up required for elders with age above 40). The database consists of a 10% sample of randomly selected elderly individuals (430,133 individuals) over 65 years of age containing health and insurance billing data of from 2002 to 2010 in South Korea. Individuals who died between 2002 and 2010 were not included in this cohort. This database is representative of the Korean population because for the years investigated in this study, the Korean NHIS covered over 96% of the entire 50-million South Korean population; thus, presents minimal selection bias (**Supplemental Figure 1**).

Of those samples, 40,736 elders were selected in this study, whose records exist in all the three databases (Participant Insurance Eligibility database, Healthcare Utilization database, and National Health Screening database). The Korean NHIS Electronic Health Records Detailed description of the EHR including access is available elsewhere (https://nhiss.nhis.or.kr/bd/ab/bdaba000eng.do). Ethics review and institutional review boards approved the study with exemption of informed consent (for retrospective, de-identified, publicly available data) (IRB number NHIMC 2018-12-006).

Definition of AD

Incident AD was the outcome variable. We used the two criteria to define AD: ICD-10 codes of AD[18] (F00, F00.0, F00.1, F00.2, F00.9, G30, G30.0, G30.1, G30.8, G30.9) and dementia medication prescribed with an initial AD diagnosis (e.g., donepezil, rivastigmine, galantamine, and memantine). When both criteria were used, we labeled it as *definite AD*. We also considered a broader definition of AD using only ICD-10 codes to minimize false negative cases (e.g. individuals with AD diagnose who did not take medication); this was labeled as *probable AD*. Within each individual with AD incidence, the EHR after the AD incidence was excluded. We conducted predictive modeling using both outcome variables.

Data and Preprocessing

We used the following variables from the EHR data: 21 features including laboratory values, health profiles, history of family illness from the Health Screening database; 2 features including age and sex from the Participant Insurance Eligibility database; and 6,412 features including ICD-10 codes and medication codes. Descriptions of data coding and exclusion criteria for all the features except for ICD-10 codes and medication codes are available in **Supplementary Table 1.**

Our data preprocessing steps are as follows. (i) EHR alignment: We aligned the EHRs to each individual’s initial AD diagnosis (event-centric ordering). (ii) ICD-10 and medication coding: Since ICD-10 and medication codes have hierarchical structures, we used the first disease category codes (e.g., F00 [Dementia in Alzheimer’s disease] including F00.0 [Dementia in Alzheimer’s disease with early onset], F00.1 [Dementia in Alzheimer’s disease with late onset], F00.2 [Dementia in Alzheimer’s disease, atypical or mixed type], and F00.9 [Dementia in Alzheimer´s disease, unspecified]), and the first 4 characters for the medication codes representing main ingredients. (iii) Rare disease or medication codes found less than five times in the entire data were excluded from the analysis (1,179 disease and 362 medication codes). (iv) if a participant has no health screening data (laboratory values, health profiles, and history of personal and family illness from the National Health Screening database) during the last two years of the processed data (in Korea an biannual health screening is required for every elder), we excluded that participant from the analysis. This preprocessing procedure yielded 4,894 unique variables used in the models (see **Table 3** for detailed information).

For each *n*-year prediction, within the AD group, we used the EHR between 2002 and the year of incident AD – *n* because it requires at least *n* years prior to the incident AD. Within the non-AD group, we used the EHR from 2002 to 2010 – *n*. For example, for 1 year prediction, if a patient was diagnosed with AD at 2009, we used the EHR between 2002 and 2008; for 2 year prediction, 2002-2007; for 3 year, 2002-2006; and for 4 year, 2002-2005.

Machine learning analysis

We implemented three machine learning algorithms: random forest, support vector machine with linear kernel, and logistic regression. Model training, validation, and testing was done using nested stratified 5-fold cross validation with 5 iterations. Feature selection was done within train sets using the variance threshold method.[19] Hyper-parameters optimization was done within validation sets. The following parameters were tuned: for random forest, the minimum number of samples required at a leaf node and the number of trees in the forest; for support vector machine, regularization strength; for logistic regression, the inverse of regularization strength. In logistic regression L2 regularization was used. Generalizability of model performance was assessed on the test sets. We measured the following model performance metrics in the test set: The area under the receiver operating characteristic curve (ROC), sensitivity and specificity. We comply with the Transparent Reporting of a Multivariable Prediction Model for Individual

Prognosis or Diagnosis (TRIPOD) reporting guideline. Codes are available at https://github.com/a011095/koreanEHR.

Results

Sample characteristics

Of 40,736 individuals with age above 65 years in 2002, we identified 614 unique individuals with AD incidence using the definite AD outcome, 2,026 with AD incidence using the probable AD definition, and 38,710 elders with no AD incidence. The rate of AD in this cohort was 1.56% using the definite AD definition, and 4.97% using the probable AD definition. Demographic characteristics showed significant differences in age between both AD groups and non-AD groups and non-significant differences in income and sex (**Table 1**).

Model prediction

Classifiers were trained on these to predict 0,1,2,3, and 4 subsequent-year incidence of AD. When using the definite AD definition (based on ICD-10 codes and dementia prescription), in predicting 0yr incidence of AD, random forest (RF) showed the best performance with AUC of 0.887 (**Table 2** and **Figure 2**). When using the probable AD definition (based on ICD-10 codes), classification performance was slightly lower with AUC of 0.805 (RF). Classification performance decreased in predicting future incident AD of later years: using the definite AD definition, AUC of 0.781 (1 year), 0.739 (2 year), 0.686 (3 year), and 0.662 (4 year); using the probable AD definition, AUC of 0.730 (1 year), 0.645 (2 year), 0.575 (3 year), and 0.602 (4 year). Numbers of features and look-back periods also decreased in later year (**Table 3**).

Important features

Logistic regression identified the features positively related to incident AD. These included age (b value = 0.689), elevated urine protein (0.303), prescription of Zotepine (antipsychotic drug) (0.303), and the features negatively related to incident AD, such as, decreased hemoglobin (-

0.902), prescription of Nicametate Citrate (-0.297), diagnosis of other degenerative disorders of nervous systems (-0.292), and disorders of the external ear (-0.292) (**Table 4**).

Discussion

This study assessed the utility of the EHR in predicting the future incidence of AD. Using machine learning, we predicted future incidence of AD with acceptable accuracy in terms of AUC (0.781 in one-year prediction). The high accuracy of our models based on large nation-wide samples may lend a support to the potential utility of the EHR-based predictive modeling in AD. Despite of the limitations inherent to the use of administrative EHR, such as the inability to directly ascertain clinical phenotypes, this study demonstrates the potential utility of the EHR for AD screening, when combined with rigorous data-driven machine learning.

Our model performance with AUC of 0.887, 0.781, and 0.662 in predicting baseline, subsequent one-year, and four-year incident AD is relatively accurate compared with the literature. In all-cause dementia risk prediction based on genetic (ApoE) or neuropsychological evaluations, MRI, health indices (diabetes, hypertension, lifestyle), and demographic (age, sex, education) variables, prior models show accuracy ranging from 0.5 to 0.78 in AUC (reviewed in [20]). Of note, compared with these studies, our approach is solely based on administrative EHR without neuropsychological, genetic testing, or brain imaging. This has important implications for the practical utility of the EHR-based risk prediction, in that it can provide an early indication of AD risk to clinicians. Together with existing screening tools (e.g., MMSE), this may assist deciding when to seek a further clinical assessment to a given patient in an individual-specific manner.

Our model detected interesting EHR-based features associated with incident AD. The data-driven selection of features is consistent with risk factors found in the literature. A decrease in hemoglobin level was selected as the feature most strongly associated with incident AD. Indeed, anemia is known as an important risk factor for dementia.[21-23] A study using National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS), the NHIS health screening data in Korea, not only found that anemia was associated with dementia, but also revealed a dose-dependent relationship between anemia and dementia.[24] Likewise, our data-driven model shows the hemoglobin level as the most significant predictor. This finding has implications for public health because anemia is a modifiable factor. Given our finding and the consistent literature on the large association between hemoglobin level and AD and other dementia, future research may investigate the biological pathway of anemia's contribution to AD pathology and cognitive decline.

We also noted a positive association between urine protein level and incident AD. In the EHR, protein in urine is typically measured using urine dip stick. This approach is not a quantitative measure of urine protein, but it is useful as a screening method for proteinuria.[25,26] Literature shows association between albuminuria and dementia.[27] Our finding suggests the potential utility of a urine test as part of the routine health check-up in AD risk prediction.

Four medications were also associated with incident dementia within top ten features. We found that Zotepine, Eperisone hydrochloride had a positive association and Nicametate Citrate and Tolfenamic acid had a negative association with incident AD. It is interesting that patients prescribed tolfenamic acid showed lower incidence of AD. This drug used in Korea for pain control in conditioner such as rheumatoid arthritis. It is known to lower the gene expression of Amyloid precursor protein 1(APP1) and beta-site APP cleaving enzyme 1(BACE1) by promoting the degradation of specificity protein 1(Sp1).[28-30] As a potential modifier of tau protein, Tolfenamic acid is under investigation as a potential drug to prevent and modify the progression of AD.[31] The results of this study support the above experimental result and show that tolfenamic acid may be a potential anti-dementia medication.

Zotepine is an atypical antipsychotic drug with proven efficacy for treatment of schizophrenia. Our model showed the use of zotepine positively correlated with incident AD. There are two possible interpretations. Some studies indicate that individuals with schizophrenia may have an increased risk for the development of dementia.[32] It is possible that the incident AD was high in patients with schizophrenia using zotepine. Alternatively, zotepine may have been used to control behavioral and psychological symptoms before incident AD.[33] Further research is required to address why other schizophrenia drugs or other drugs used to treat behavioral and psychological symptoms of dementia (BPSD) were not detected.

Nicametate Citrate, a vasodilator, was also negatively associated with incident AD. This may be in line with the literature showing effects of vasodilators on increasing cognitive function and reducing the risk of vascular dementia, although the exact mechanism remains unclear.[34,35] Further research is required.

Limitations

One of the limitations of this study is that diagnose of AD in our EHR is not clinically ascertained. This is inevitable in nation-wide administrative data. Nevertheless, some aspects may worth noting. Firstly, we confirmed the comparable prediction outcomes using definitions of incident AD, that is, “probable AD” based on AD disease codes and “definite AD” based on both AD disease codes and anti-dementia medication, separately. Secondly, in South Korea, every elder with age 60 years old is required to have complementary dementia screening supported by the National Health Insurance Service at public healthcare centers, where individuals that high-risk for dementia get referred to physicians for further clinical examination. This healthcare system may help reduce false negative cases. These aspects may alleviate potential concerns of the validity of AD diagnoses in terms of false positive and negative cases. Lastly, the health insurance system and policies unique to Korea support the reliability of the AD diagnoses. In

Korea, the Health Insurance Review and Assessment Service (HIRA) of NHIS reviews and supervises the medical claims of drugs to treat AD. For example, HIRA requires the following conditions to consider the insurance coverage of dementia medication: for donepezil and rivastigmine patches, MMSE (Mini-Mental State Examination) =< 26 and CDR (Clinical Dementia Rating) = 1~3 or GDS (Global Deterioration Scale)= 3~7; for galantamine and rivastigmine capsules, MMSE = 10 ~ 26 and CDR = 1~2 or GDS = 3~5; for memantine, MMSE =< 20 and CDR = 2~3 or GDS = 4~7. Furthermore, these medications can be only refilled when the patients meet the same criteria on follow-up neurocognitive tests every 12 months (**Supplementary Figure 2**). Thus, it is highly likely that individuals with records of receiving dementia medication meet strong diagnostic criteria.

Another limitation of this study is that generalizability of our findings to ethnicities other than Asian or to different healthcare systems remains to be tested.

Conclusions

In sum, this study presents the first data in predicting future incident AD using data-driven machine learning based on large-scale EHR. Our results lend 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.

Acknowledgement

Ilsan Hospital Research Support Program (PI: Kim, HS); NIMH K01 MH109836 (PI: Cha); Young Investigator Award, Brain Behavior Research Foundation (PI: Cha); Young Investigator Grant, Korean Scientists and Engineers Association (PI: Cha).

References

1. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am J Public Health.* 1998;88(9):1337-1342.
2. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary costs of dementia in the United States. *N Engl J Med.* 2013;368(14):1326-1334.
3. Zissimopoulos J, Crimmins E, St Clair P. The Value of Delaying Alzheimer's Disease Onset. *Forum Health Econ Policy.* 2014;18(1):25-39.
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst.* 2014;2:3.
5. Kivipelto M, Ngandu T, Laatikainen T, Winblad B, Soininen H, Tuomilehto J. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. *Lancet Neurol.* 2006;5(9):735-741.
6. Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate? *Nat Rev Neurol.* 2010;6(6):318-326.
7. Backman L, Jones S, Berger AK, Laukka EJ, Small BJ. Multiple cognitive deficits during the transition to Alzheimer's disease. *J Intern Med.* 2004;256(3):195-204.
8. Jorm AF, Masaki KH, Petrovitch H, Ross GW, White LR. Cognitive deficits 3 to 6 years before dementia onset in a population sample: the Honolulu-Asia aging study. *J Am Geriatr Soc.* 2005;53(3):452-455.
9. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine.* 2018;1(1):18.
10. Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. 2018;1(8):e185097-e185097.

11. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley DJJNO. Development and validation of an electronic health record–based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. 2018;1(4):e181018-e181018.

12. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep.* 2016;6:26094.

13. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med.* 2016;375(13):1216-1219.

14. Naylor CD. On the Prospects for a (Deep) Learning Health Care System. *JAMA.* 2018;320(11):1099-1100.

15. Hinton G. Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA.* 2018;320(11):1101-1102.

16. Shin DW, Cho B, Guallar E. Korean National Health Insurance Database. *JAMA Intern Med.* 2016;176(1):138.

17. Lee J, Lee JS, Park S-H, Shin SA, Kim K. Cohort profile: The national health insurance service–national sample cohort (NHIS-NSC), South Korea. *International journal of epidemiology.* 2016;46(2):e15-e15.

18. WHO. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)-WHO Version for 2016. http://apps.who.int/classifications/icd10/browse/2016/en#/F00. Accessed.

19. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3(Mar):1157-1182.

20. Tang EY, Harrison SL, Errington L, et al. Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. *PLoS One.* 2015;10(9):e0136181.

21. Atti AR, Palmer K, Volpato S, Zuliani G, Winblad B, Fratiglioni L. Anaemia increases the risk of dementia in cognitively intact elderly. *Neurobiology of aging.* 2006;27(2):278-284.

22. Shah RC, Buchman AS, Wilson RS, Leurgans SE, Bennett DA. Hemoglobin level in older persons and incident Alzheimer disease: prospective cohort analysis. *Neurology.* 2011;77(3):219-226.

23. Hong CH, Falvey C, Harris TB, et al. Anemia and risk of dementia in older adults: findings from the Health ABC study. *Neurology.* 2013;81(6):528-533.

24. Jeong SM, Shin DW, Lee JE, Hyeon JH, Lee J, Kim S. Anemia is associated with incidence of dementia: a national health screening study in Korea involving 37,900 persons. *Alzheimer's research & therapy.* 2017;9(1):94.

25. Chotayaporn T, Kasitanon N, Sukitawut W, Louthrenoo W. Comparison of proteinuria determination by urine dipstick, spot urine protein creatinine index, and urine protein 24 hours in lupus patients. *Journal of clinical rheumatology : practical reports on rheumatic & musculoskeletal diseases.* 2011;17(3):124-129.

26. White SL, Yu R, Craig JC, Polkinghorne KR, Atkins RC, Chadban SJ. Diagnostic accuracy of urine dipsticks for detection of albuminuria in the general community. *American journal of kidney diseases : the official journal of the National Kidney Foundation.* 2011;58(1):19-28.

27. Deckers K, Camerino I, van Boxtel MP, et al. Dementia risk in renal dysfunction: A systematic review and meta-analysis of prospective studies. *Neurology.* 2017;88(2):198-208.

28. Subaiea GM, Adwan LI, Ahmed AH, Stevens KE, Zawia NH. Short-term treatment with tolfenamic acid improves cognitive functions in Alzheimer's disease mice. *Neurobiology of aging.* 2013;34(10):2421-2430.

29. Adwan L, Subaiea GM, Basha R, Zawia NH. Tolfenamic acid reduces tau and CDK5 levels: implications for dementia and tauopathies. *Journal of neurochemistry.* 2015;133(2):266-272.

30. Adwan L, Subaiea GM, Zawia NH. Tolfenamic acid downregulates BACE1 and protects against lead-induced upregulation of Alzheimer's disease related biomarkers. *Neuropharmacology.* 2014;79:596-602.

31. Chang JK, Leso A, Subaiea GM, et al. Tolfenamic Acid: A Modifier of the Tau Protein and its Role in Cognition and Tauopathy. *Current Alzheimer research.* 2018;15(7):655-663.

32. Cai L, Huang J. Schizophrenia and risk of dementia: a meta-analysis study. *Neuropsychiatric disease and treatment.* 2018;14:2047-2055.

33. Rhee Y, Csernansky JG, Emanuel LL, Chang CG, Shega JW. Psychotropic medication burden and factors associated with antipsychotic use: an analysis of a population-based sample of community-dwelling older persons with dementia. *Journal of the American Geriatrics Society.* 2011;59(11):2100-2107.

34. Perng CH, Chang YC, Tzang RF. The treatment of cognitive dysfunction in dementia: a multiple treatments meta-analysis. *Psychopharmacology.* 2018;235(5):1571-1580.

35. McLennan SN, Lam AK, Mathias JL, Koblar SA, Hamilton-Bruce MA, Jannes J. Role of vasodilation in cognitive impairment. *International journal of stroke : official journal of the International Stroke Society.* 2011;6(3):280.

Table 1. Sample characteristics

	Definite AD	**Probable AD**	**Non-AD**
Number	614	2,026	38,710
Income	6.00 (5.73-6.27)	5.90 (5.87-5.93)	6.02 (5.87-6.17)
Age	80.67 (80.2-81.1)	79.2 (79.0-79.5)	74.5 (74.4-74.5)
sex	Male:229 Female:285	Male:733 Female:1,293	Male:18,200 Female:20,510

*Based on the 0-year prediction model.

Table 2. Performance of predictive models trained on EHR.

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
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

Table 3. Lengths of EHR (look-back periods) and number of features

		Definite AD		**Probable AD**		**Non-AD**	
	Number of features	Average EHR length per subject in days	Average number of non-zero features per subject	Average EHR length per subject in days	Average number of non-zero features per subject	Average EHR length per subject in days	Average number of non-zero features per subject
0 yr	4,894	1936 (1906-1967)	162 (156-167)	2239 (2205-2273)	185 (179-192)	3033 (3028-3038)	176 (174-177)
1 yr	4,722	1851 (1800-1902)	172 (161-182)	1936 (1906-1967)	162 (156-167)	2694 (2690-2698)	164 (163-165)
2 yr	4,622	1571 (1524-1619)	141 (133-149)	1656 (1627-1684)	139 (134-144)	2381 (2378-2384)	151 (150-152)
3 yr	4,494	1666 (1622-1710)	146 (138-154)	1736 (1709-1763)	144 (139-150)	2045 (2042-2047)	135 (134-136)
4 yr	4,353	1736 (1691-1781)	158 (147-169)	1822 (1796-1848)	152 (146-158)	1711 (1708-1714)	116 (114-117)

Table 4. Top ten features and weights from logistic regression (0-yr prediction).

Type of data	Name	b value
health checkup	hemoglobin	-0.902
demography	age	0.689
health checkup	urine protein	0.303
medication	Zotepine (antipsychotic drug)	0.303
medication	Nicametate Citrate (vasodilator)	-0.297
disease code	other degenerative disorders of nervous system in diseases classified elsewhere	-0.292
disease code	disorders of external ear in diseases classified elsewhere	-0.274
medication	Tolfenamic acid 200mg (pain killer)	-0.266
disease code	adult respiratory distress syndrome	-0.259
medication	Eperisone Hydrochloride (antispasmodic drug)	0.255

Figure 1. Consort Diagram.

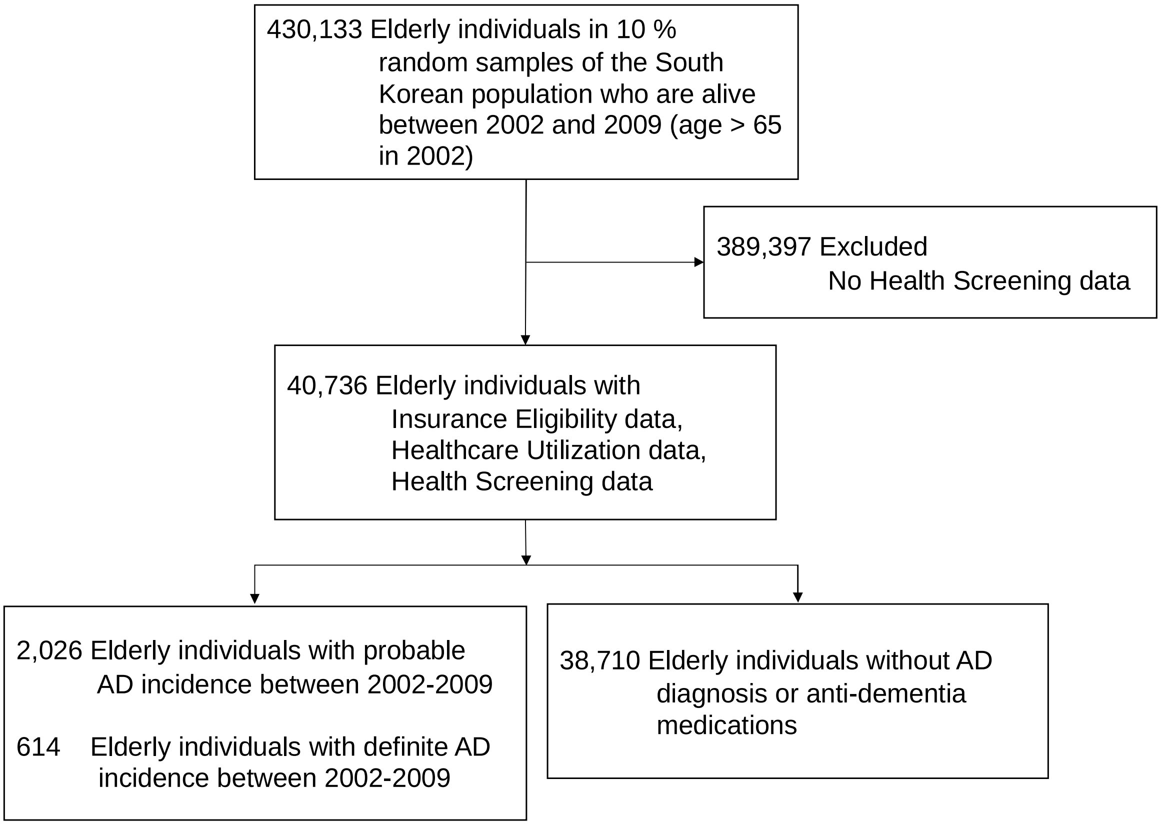
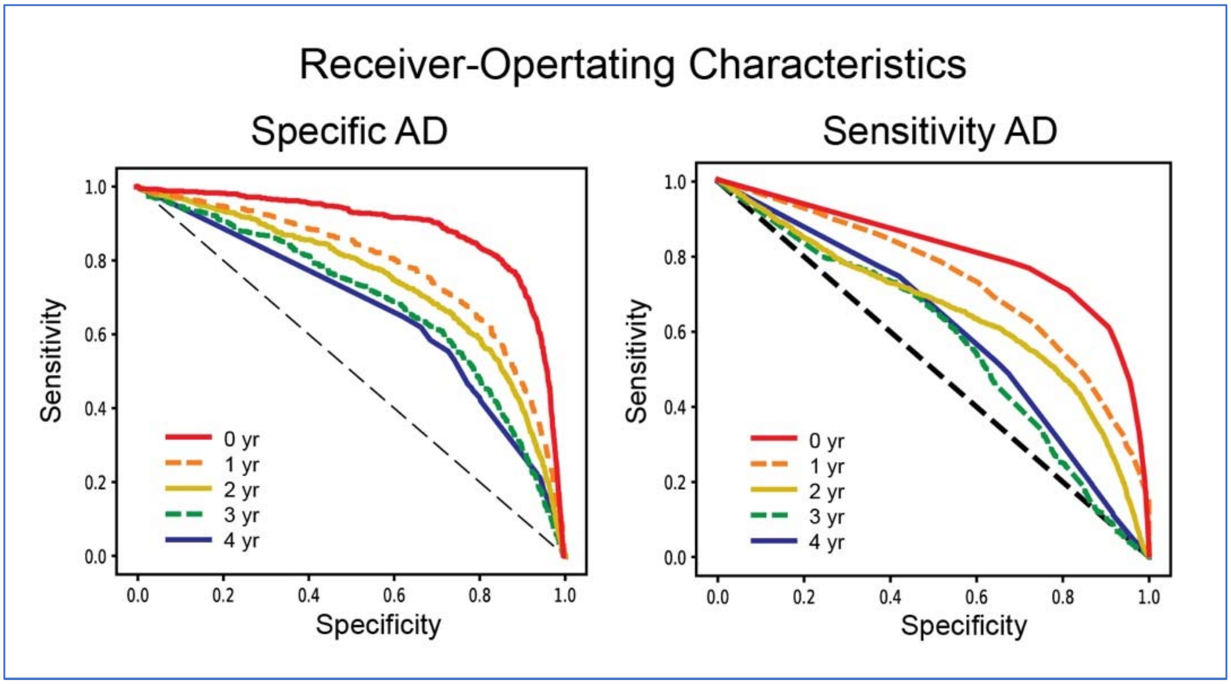
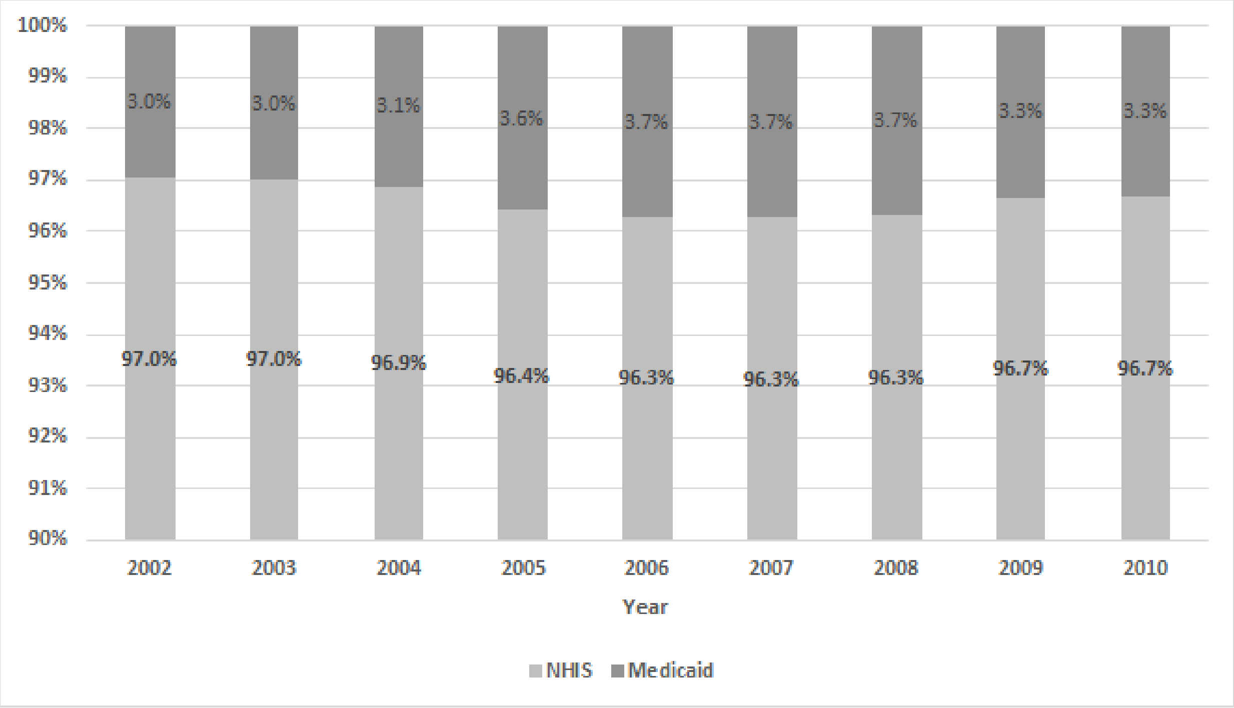

430,133 Elderly individuals in 10 % random samples of the South Korean population who are alive between 2002 and 2009 (age > 65 in 2002)
389,397 Excluded No Health Screening data
40,736 Elderly individuals with Insurance Eligibility data, Healthcare Utilization data, Health Screening data
2,026 Elderly individuals with probable AD incidence between 2002-2009
614 Elderly individuals with definite AD incidence between 2002-2009
38,710 Elderly individuals without AD diagnosis or anti-dementia medications

Figure 2. Performance of machine learning models in predicting incident AD. Receiver-Operating Characteristic plots are shown for 0,1,2,3,4-year prediction**.** Incident AD was defined based on ICD-10 AD codes and anti-dementia medication for AD, "Definite AD", or based on AD codes only, "Probable AD".

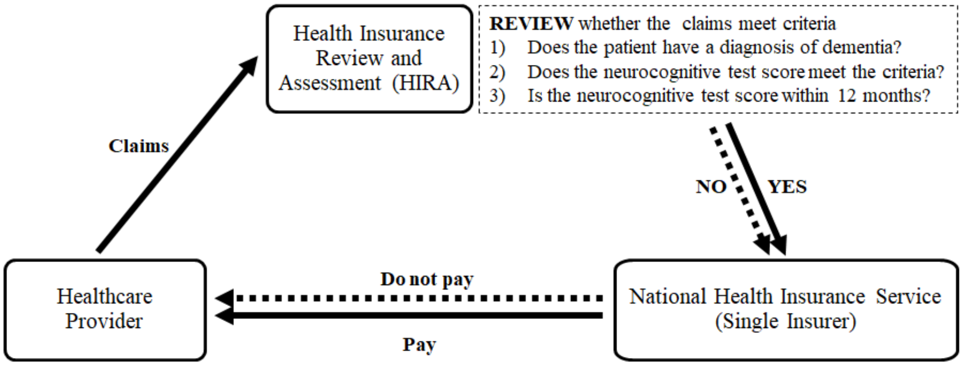


Supplementary Materials

Supplementary Figure 1. For the years investigated in this study, the Korean NHIS covered more than 96% of the South Korean population (50 millions).



Supplementary Figure 2. Medical insurance system dementia medication in Korea.


Health Insurance Review and Assessment (HIRA)
REVIEW whether the claims meet criteria
1) Does the patient have a diagnosis of dementia?
2) Does the neurocognitive test score meet the criteria?
3) Is the neurocognitive test score within 12 months?
Claims
NO
YES
Healthcare Provider
Do not pay
Pay
National Health Insurance Service (Single Insurer)

Supplementary Table1. Sociodemographic and Health Profile Variables Use in The Model.

Variables	Type of variable	Explanation
Age	continuous	In years
Sex	binary	0: Female; 1 : Male
Body mass index	continuous	Weight(kg) / (Height*Height)(m2)
Systolic blood pressure	continuous	mmHg Below 60mmHg or Above 400mmHg : Treated as null
Diastolic blood pressure	continuous	mmHg Below 30mmHg or Above 250mmHg : Treated as null
Fasting glucose	continuous	mg/dL Below 25mg/dL or Above 999mg/dL : Treated as null
Hemoglobin	continuous	Measured from 2009 g/dL Above 25.0g/dL : Treated as null ~
Urine protein	ordinal	Measured from 2009 1 : negative (-) 2 : weak positive (±) 3 : positive (1+) 4 : positive (2+) 5 : positive (3+) 6 : positive (4+)
Serum creatinine	continuous	mg/dL
Serum AST	continuous	U/L
Serum ALT	continuous	U/L
r-GTP	continuous	U/L
Family history of liver disease	binary	1 : no 2 : yes
Family history of hypertension	binary	
Family history of stroke	binary	
Family history of cardiac disease	binary	
Family history of diabetes mellitus	binary	
Family history of cancer	binary	
Smoking status	continuous	1 : Never smoked 2 : Not current smoker but smoked in the past 3 : Current smoker
Total smoking period	ordinal	1 : below 5 years 2 : 5-9 years 3 : 10-19 years 4 : 20-29 years 5 : over 30 years
Current daily amount of smoking	ordinal	1 : 1~ 12 cigarettes 2: 13-24 cigarettes 3 : 25~48 cigarettes 4 : over 49 cigarettes
Frequency of drinking alcohol	ordinal	1 : almost none 2 : 2~3 per month 3: 1~2 per week 4 : 3~4 per week 5 : almost everyday
Amount of alcohol intake in one day	ordinal	1 : below 30g of alcohol 2 : below 60g of alcohol 3 : below 90g of alcohol 4 : over 120g of alcohol