

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

Ji Hwan Park, PhD<sup>1\*</sup>, Han Eol Cho, MD<sup>2\*</sup>, Jong Hun Kim, MD, PhD<sup>3</sup>, Melanie Wall, PhD<sup>4</sup>,  
Yaakov Stern, PhD<sup>4,5</sup>, Hyunsun Lim, PhD<sup>6</sup>, Shinjae Yoo, PhD<sup>1</sup>,  
Hyoung-Seop Kim, MD<sup>3</sup>, Jiook Cha, PhD<sup>4,7</sup> (\*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, Korea;
3. National Health Insurance Service Ilsan Hospital, Goyang, South Korea;
4. Department of Psychiatry;
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, Korea;
7. Data Science Institute, Columbia University, New York, USA.

## Correspondence to:

Hyoung-Seop Kim, MD

[rekhs@nhimc.or.kr](mailto:rekhs@nhimc.or.kr)

100 ilsan-ro ilsan-donggu Goyang City Gyeonggi Do, 10444, Republic of Korea

Jiook Cha, PhD

[jc4248@cumc.columbia.edu](mailto:jc4248@cumc.columbia.edu)

1051 Riverside Dr.

New York, NY, 10033, USA

30 **Key Points**

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

33

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

37

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

40

41 **Abstract**

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

48 **Methods:** We obtained de-identified EHR from Korean elders age above 65 years old  
49 (N=40,736) collected between 2002 and 2012 in the Korean National Health Insurance Service  
50 database system. Consisting of Participant Insurance Eligibility database, Healthcare Utilization  
51 database, and Health Screening database, this EHR contain 4,894 unique clinical features  
52 including ICD-9/10 codes, medication codes, laboratory values, history of personal and family  
53 illness, and socio-demographics. Our event of interest was new incidence of AD defined from  
54 the EHR based on both AD codes and prescription of anti-dementia medication. Two definitions  
55 were considered: a more stringent one requiring a diagnosis and dementia medication resulting  
56 in n=614 cases (“definite AD”) and a more liberal one requiring only diagnostic codes (n=2,026;  
57 “probable AD”). We trained and validated a random forest, support vector machine, and logistic  
58 regression to predict incident AD in 1,2,3, and 4 subsequent years using the EHR available  
59 since 2002. The length of the EHR used in the models ranged from 1,571 to 2,239 days. Data  
60 was randomly split into training (60%), validation (20%), and test sets (20%) so that AUC values  
61 represent true out of sample prediction are based on the test set.

62 **Results:** Average duration of EHR was 1,936 days in AD and 2,694 days in controls. For  
63 predicting future incidence of AD using the “definite AD” outcome, the machine learning models  
64 showed the best performance in 1 year prediction with AUC of 0.781; in 2 year, 0.739; in 3 year,

65 0.686; in 4 year, 0.662. Using “probable AD” outcome, the machine learning models showed the  
66 best performance in 1 year prediction with AUC of 0.730; in 2 year, 0.645; in 3 year, 0.575; in 4  
67 year, 0.602. Important clinical features selected in logistic regression included hemoglobin level  
68 (b=-0.902), age (b=0.689), urine protein level (b=0.303), prescription of Lodopin (antipsychotic  
69 drug) (b=0.303), and prescription of Nicametate Citrate (vasodilator) (b=-0.297).

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

## 72 **Introduction**

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

83

84 For population AD screening, prior models are based on predefined features including health  
85 profiles, such as sociodemographic (age, sex, education), lifestyle (physical activity), midlife  
86 health risk factors (systolic blood pressure, BMI and total cholesterol level)<sup>4,5</sup>; and cognitive  
87 profiles<sup>6,7</sup>. Despite of the demonstrated accuracy of these models, an important outstanding  
88 question is whether the several curated variables may sufficiently account for the  
89 heterogeneous etiology of multi-factorial AD. Indeed, a meta-analysis study shows that multi-  
90 factor models best predict risk for dementia, whereas single-factor models do poorly<sup>5</sup>,  
91 suggesting accurate AD screening with practical utility in large populations require sufficiently  
92 large feature space. An important new approach for developing individualized predictive  
93 modeling is the use of the rigorous data-driven machine learning that can harvest salient  
94 information from large-scale EHR to make an individual-specific predictions.

95

96 Machine learning is an optimal choice of the analytic method for analyzing large-scale EHR  
97 containing thousands of descriptors in hundreds of thousands of individuals. Studies show  
98 successful application of machine learning to the EHR in predicting incident diseases (cancer,  
99 diabetes, schizophrenia, etc) or mortality<sup>8-11</sup>. Given the recent rapid growth of the machine  
100 learning technology, application of the AI technology to clinical predictive modeling is likely to  
101 have a deep impact on medicine<sup>12</sup>. But to our knowledge data-driven predictive modeling with  
102 EHR data has not been previously used to predict incident AD.

103

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

113

## 114 **Materials and Methods**

### 115 **Datasets**

116 We used the National Health Insurance Service (NHIS)-National Elderly cohort Database, a  
117 subsample of the National Health Insurance Service-national sample cohort<sup>13</sup>. This database  
118 contains for each individual features of services/diagnoses/prescriptions associated with all the  
119 health care services provided by the NHIS. All EHR was binned monthly. Clinical features  
120 include demographics and socioeconomics from the *Participant Insurance Eligibility database*;  
121 disease and medication codes from the *Healthcare Utilization database*; and laboratory values,  
122 health profiles, and history of personal and family illness from the *National Health Screening*  
123 *database* (from bi-annual health check-up required for elders with age above 40). The database  
124 consists of a 10% sample of randomly selected elderly individuals (430,133 individuals) over 65  
125 years of age containing health and insurance billing data of from 2002 to 2012 in South Korea.  
126 Individuals who died between 2002 and 2012 were not included in this cohort. This database is  
127 representative of the Korean population because for the years investigated in this study, the  
128 Korean NHIS covered over 96% of the entire 50-million South Korean population; thus, presents  
129 minimal selection bias (**Supplemental Figure 1**).

130

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

138

139 **Definition of AD**

140 Incident AD was the outcome variable. We used the two criteria to define AD: ICD-10 codes of  
141 AD<sup>14</sup> (F00, F00.0, F00.1, F00.2, F00.9, G30, G30.0, G30.1, G30.8, G30.9) and dementia  
142 medication prescribed with an initial AD diagnosis (e.g., donepezil, rivastigmine, galantamine,  
143 and memantine). When both criteria were used, we labeled it as *definite AD*. We also  
144 considered a broader definition of AD using only ICD-10 codes to minimize false negative cases  
145 (e.g. individuals with AD diagnose who did not take medication); this was labeled as *probable*  
146 *AD*. Within each individual with AD incidence, the EHR after the AD incidence was excluded.  
147 We conducted predictive modeling using both outcome variables.

148

149 **Data and Preprocessing**

150 We used the following variables from the EHR data: 21 features including laboratory values,  
151 health profiles, history of personal and family illness from the Health Screening database; three  
152 features including age, sex, income level from the Participant Insurance Eligibility database; and  
153 the 4,871 features including ICD 9/10 codes and medication codes. Descriptions of data coding  
154 and exclusion criteria for all the features except for ICD 9/10 codes and medication codes are  
155 available in **Supplementary Table 1**.

156

157 Our data preprocessing steps are as follows. (i) EHR alignment: We aligned the EHRs to each  
158 individual's initial AD diagnosis (event-centric ordering<sup>29</sup>). (ii) ICD 9/10 and medication coding:  
159 Since ICD9/10 and medication codes have hierarchical structures, we used the first disease  
160 category codes (e.g., F00 [Dementia in Alzheimer's disease] including F00.0 [Dementia in  
161 Alzheimer's disease with early onset], F00.1 [Dementia in Alzheimer's disease with late onset],  
162 F00.2 [Dementia in Alzheimer's disease, atypical or mixed type], and F00.9 [Dementia in  
163 Alzheimer's disease, unspecified]), and the first 4 characters for the medication codes  
164 representing main ingredients. (iii) Rare disease or medication codes found less than five times



165 in the entire data were excluded from the analysis (1,179 disease and 362 medication codes).  
166 (iv) if a participant has no health screening data (laboratory values, health profiles, and history of  
167 personal and family illness from the National Health Screening database) during the last two  
168 years of the processed data (in Korea an biannual health screening is required for every elder),  
169 we excluded that participant from the analysis. After preprocessing, we identified 4,894 unique  
170 variables used in the models (see **Table 3** for detailed information).

171

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

177

### 178 **Machine learning analysis**

179 We implemented three machine learning algorithms: random forest, support vector machine  
180 with linear kernel, and logistic regression. Data was randomly split into training (60%), validation  
181 (20%), and test sets (20%) in a stratified manner. Feature selection was done within train sets  
182 using the variance threshold method<sup>15</sup>. Hyper-parameters optimization was done within  
183 validation sets. The following parameters were tuned: for random forest, the minimum number  
184 of samples required at a leaf node and the number of trees in the forest; for support vector  
185 machine, regularization strength; for logistic regression, the inverse of regularization strength. In  
186 logistic regression L2 regularization was used. Generalizability of model performance was  
187 assessed on the test sets. We measured the following model performance metrics in the test  
188 set: The area under the receiver operating characteristic curve (ROC), sensitivity and specificity.  
189 We comply with the Transparent Reporting of a Multivariable Prediction Model for Individual

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

192

## 193 **Results**

### 194 **Sample characteristics**

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

201

### 202 **Model prediction**

203 Classifiers were trained on these to predict 0,1,2,3, and 4 subsequent-year incidence of AD.  
204 When using the definite AD definition (based on ICD-10 codes and dementia prescription), in  
205 predicting 0yr incidence of AD, random forest (RF) showed the best performance with AUC of  
206 0.887 (**Table 2** and **Figure 2**). When using the probable AD definition (based on ICD-10 codes),  
207 classification performance was slightly lower with AUC of 0.805 (RF). Classification  
208 performance decreased in predicting future incident AD of later years: using the definite AD  
209 definition, AUC of 0.781 (1 year), 0.739 (2 year), 0.686 (3 year), and 0.662 (4 year); using the  
210 probable AD definition, AUC of 0.730 (1 year), 0.645 (2 year), 0.575 (3 year), and 0.602 (4  
211 year). Numbers of features and look-back periods also decreased in later year (**Table 3**).

212

### 213 **Important features**

214 Logistic regression identified the features positively related to incident AD. These included age  
215 (b value = 0.689), elevated urine protein (0.303), prescription of Zolpidem (antipsychotic drug)  
216 (0.303), and the features negatively related to incident AD, such as, decreased hemoglobin (-

217 0.902), prescription of Nicametate Citrate (-0.297), diagnosis of other degenerative disorders of  
218 nervous systems (-0.292), and disorders of the external ear (-0.292) (**Table 4**).

219

## 220 **Discussion**

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

228

229 Our model performance with AUC of 0.89, 0.78, and 0.66 in predicting baseline, subsequent  
230 one-year, and four-year incident AD is relatively accurate compared with the literature. In all-  
231 cause dementia risk prediction based on genetic (ApoE) or neuropsychological evaluations,  
232 MRI, health indices (diabetes, hypertension, lifestyle), and demographic (age, sex, education)  
233 variables, prior models show accuracy ranging from 0.5 to 0.78 in AUC (reviewed in <sup>16</sup>). Of note,  
234 compared with these studies, our approach is solely based on administrative EHR without  
235 neuropsychological, genetic testing, or brain imaging. This has important implications for the  
236 practical utility of the EHR-based risk prediction, in that it can provide an early indication of AD  
237 risk to clinicians. Together with existing screening tools (e.g., MMSE), this may assist deciding  
238 when to seek a further clinical assessment to a given patient in an individual-specific manner.

239

240 Our model detected interesting EHR-based features associated with incident AD. The data-  
241 driven selection of features is consistent with risk factors found in the literature. A decrease in  
242 hemoglobin level was selected as the feature most strongly associated with incident AD.  
243 Indeed, anemia is known as an important risk factor for dementia<sup>17-19</sup>. A study using National  
244 Health Insurance Service-National Health Screening Cohort (NHIS-HEALS), the NHIS health

245 screening data in Korea, not only found that anemia was associated with dementia, but also  
246 revealed a dose-dependent relationship between anemia and dementia<sup>20</sup>. Likewise, our data-  
247 driven model shows the hemoglobin level as the most significant predictor. This finding has  
248 implications for public health because anemia is a modifiable factor. Given our finding and the  
249 consistent literature on the large association between hemoglobin level and AD and other  
250 dementia, future research may investigate the biological pathway of anemia's contribution to AD  
251 pathology and cognitive decline.

252

253 We also noted a positive association between urine protein level and incident AD. In the EHR,  
254 protein in urine is typically measured using urine dip stick. This approach is not a quantitative  
255 measure of urine protein, but it is useful as a screening method for proteinuria<sup>21,22</sup>. Literature  
256 shows association between albuminuria and dementia<sup>23</sup>. Our finding suggests the potential  
257 utility of a urine test as part of the routine health check-up in AD risk prediction.

258

259 Four medications were also associated with incident dementia within top ten features. We found  
260 that Zolpidem, Eperisone hydrochloride had a positive association and Nicametate Citrate and  
261 Tolfenamic acid had a negative association with incident AD. It is interesting that patients  
262 prescribed tolfenamic acid showed lower incidence of AD. This drug used in Korea for pain  
263 control in conditions such as rheumatoid arthritis. It is known to lower the gene expression of  
264 Amyloid precursor protein 1 (APP1) and beta-site APP cleaving enzyme 1 (BACE1) by promoting  
265 the degradation of specificity protein 1 (Sp1)<sup>24-26</sup>. As a potential modifier of tau protein,  
266 Tolfenamic acid is under investigation as a potential drug to prevent and modify the progression  
267 of AD<sup>27</sup>. The results of this study support the above experimental result and show that  
268 tolfenamic acid may be a potential anti-dementia medication.

269

270 Zotepine is an atypical antipsychotic drug with proven efficacy for treatment of schizophrenia.  
271 Our model showed the use of zotepine positively correlated with incident AD. There are two  
272 possible interpretations. Some studies indicate that individuals with schizophrenia may have an  
273 increased risk for the development of dementia<sup>28</sup>. It is possible that the incident AD was high in  
274 patients with schizophrenia using zotepine. Alternatively, zotepine may have been used to  
275 control behavioral and psychological symptoms before incident AD<sup>29</sup>. Further research is  
276 required to address why other schizophrenia drugs or other drugs used to treat behavioral and  
277 psychological symptoms of dementia (BPSD) were not detected.

278  
279 Nicametate Citrate, a vasodilator, was also negatively associated with incident AD. This may be  
280 in line with the literature showing effects of vasodilators on increasing cognitive function and  
281 reducing the risk of vascular dementia, although the exact mechanism remains unclear<sup>30,31</sup>.  
282 Further research is required.

283

#### 284 **Limitations**

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

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

306

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

309

310

### 311 **Conclusions**

312 In sum, this study presents the first data in predicting future incident AD using data-driven  
313 machine learning based on large-scale EHR. Our results lend support to the development of  
314 EHR-based AD risk prediction that may enable better selection of individuals at risk for AD in  
315 clinical trials or early detection in clinical settings.

316



317 **Acknowledgement**

318 Ilsan Hospital Research Support Program (PI: Kim, HS); NIMH K01 MH109836 (PI: Cha);

319 Young Investigator Award, Brain Behavior Research Foundation (PI: Cha); Young Investigator

320 Grant, Korean Scientists and Engineers Association (PI: Cha).

321 **References**

322 1. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the  
323 public health impact of delaying disease onset. *Am J Public Health.* 1998;88(9):1337-1342.

324 2. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary costs of dementia in the  
325 United States. *N Engl J Med.* 2013;368(14):1326-1334.

326 3. Zissimopoulos J, Crimmins E, St Clair P. The Value of Delaying Alzheimer's Disease Onset. *Forum*  
327 *Health Econ Policy.* 2014;18(1):25-39.

328 4. Kivipelto M, Ngandu T, Laatikainen T, Winblad B, Soininen H, Tuomilehto J. Risk score for the  
329 prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-  
330 based study. *Lancet Neurol.* 2006;5(9):735-741.

331 5. Stephan BC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the  
332 population: are screening models accurate? *Nat Rev Neurol.* 2010;6(6):318-326.

333 6. Backman L, Jones S, Berger AK, Laukka EJ, Small BJ. Multiple cognitive deficits during the  
334 transition to Alzheimer's disease. *J Intern Med.* 2004;256(3):195-204.

335 7. Jorm AF, Masaki KH, Petrovitch H, Ross GW, White LR. Cognitive deficits 3 to 6 years before  
336 dementia onset in a population sample: the Honolulu-Asia aging study. *J Am Geriatr Soc.*  
337 2005;53(3):452-455.

338 8. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health  
339 records. *npj Digital Medicine.* 2018;1(1):18.

340 9. Marafino BJ, Park M, Davies JM, et al. Validation of prediction models for critical care outcomes  
341 using natural language processing of electronic health record data. 2018;1(8):e185097-e185097.

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

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

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

350 13. Lee J, Lee JS, Park S-H, Shin SA, Kim K. Cohort profile: The national health insurance service–  
351 national sample cohort (NHIS-NSC), South Korea. *International journal of epidemiology.*  
352 2016;46(2):e15-e15.

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

356 15. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.*  
357 2003;3(Mar):1157-1182.

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

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

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

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

- 366 20. Jeong SM, Shin DW, Lee JE, Hyeon JH, Lee J, Kim S. Anemia is associated with incidence of  
367 dementia: a national health screening study in Korea involving 37,900 persons. *Alzheimer's*  
368 *research & therapy*. 2017;9(1):94.
- 369 21. Chotayaporn T, Kasitanon N, Sukitawut W, Louthrenoo W. Comparison of proteinuria  
370 determination by urine dipstick, spot urine protein creatinine index, and urine protein 24 hours  
371 in lupus patients. *Journal of clinical rheumatology : practical reports on rheumatic &*  
372 *musculoskeletal diseases*. 2011;17(3):124-129.
- 373 22. White SL, Yu R, Craig JC, Polkinghorne KR, Atkins RC, Chadban SJ. Diagnostic accuracy of urine  
374 dipsticks for detection of albuminuria in the general community. *American journal of kidney*  
375 *diseases : the official journal of the National Kidney Foundation*. 2011;58(1):19-28.
- 376 23. Deckers K, Camerino I, van Boxtel MP, et al. Dementia risk in renal dysfunction: A systematic  
377 review and meta-analysis of prospective studies. *Neurology*. 2017;88(2):198-208.
- 378 24. Subaiea GM, Adwan LI, Ahmed AH, Stevens KE, Zawia NH. Short-term treatment with tolfenamic  
379 acid improves cognitive functions in Alzheimer's disease mice. *Neurobiology of aging*.  
380 2013;34(10):2421-2430.
- 381 25. Adwan L, Subaiea GM, Basha R, Zawia NH. Tolfenamic acid reduces tau and CDK5 levels:  
382 implications for dementia and tauopathies. *Journal of neurochemistry*. 2015;133(2):266-272.
- 383 26. Adwan L, Subaiea GM, Zawia NH. Tolfenamic acid downregulates BACE1 and protects against  
384 lead-induced upregulation of Alzheimer's disease related biomarkers. *Neuropharmacology*.  
385 2014;79:596-602.
- 386 27. Chang JK, Leso A, Subaiea GM, et al. Tolfenamic Acid: A Modifier of the Tau Protein and its Role  
387 in Cognition and Tauopathy. *Current Alzheimer research*. 2018;15(7):655-663.
- 388 28. Cai L, Huang J. Schizophrenia and risk of dementia: a meta-analysis study. *Neuropsychiatric*  
389 *disease and treatment*. 2018;14:2047-2055.
- 390 29. Rhee Y, Csernansky JG, Emanuel LL, Chang CG, Shega JW. Psychotropic medication burden and  
391 factors associated with antipsychotic use: an analysis of a population-based sample of  
392 community-dwelling older persons with dementia. *Journal of the American Geriatrics Society*.  
393 2011;59(11):2100-2107.
- 394 30. Perng CH, Chang YC, Tzang RF. The treatment of cognitive dysfunction in dementia: a multiple  
395 treatments meta-analysis. *Psychopharmacology*. 2018;235(5):1571-1580.
- 396 31. McLennan SN, Lam AK, Mathias JL, Koblar SA, Hamilton-Bruce MA, Jannes J. Role of vasodilation  
397 in cognitive impairment. *International journal of stroke : official journal of the International*  
398 *Stroke Society*. 2011;6(3):280.

399

400 **Table 1. Sample characteristics**

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

\*Based on the 0-year prediction model.

401

402

403 **Table 2. Performance of predictive models trained on EHR.**

<b>Definite AD (AD codes and dementia prescription)</b>					
	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
<b>Probable AD (AD codes)</b>					
	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%

404

405

406

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

	Number of features	Definite AD		Probable AD		Non-AD	
		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)

407

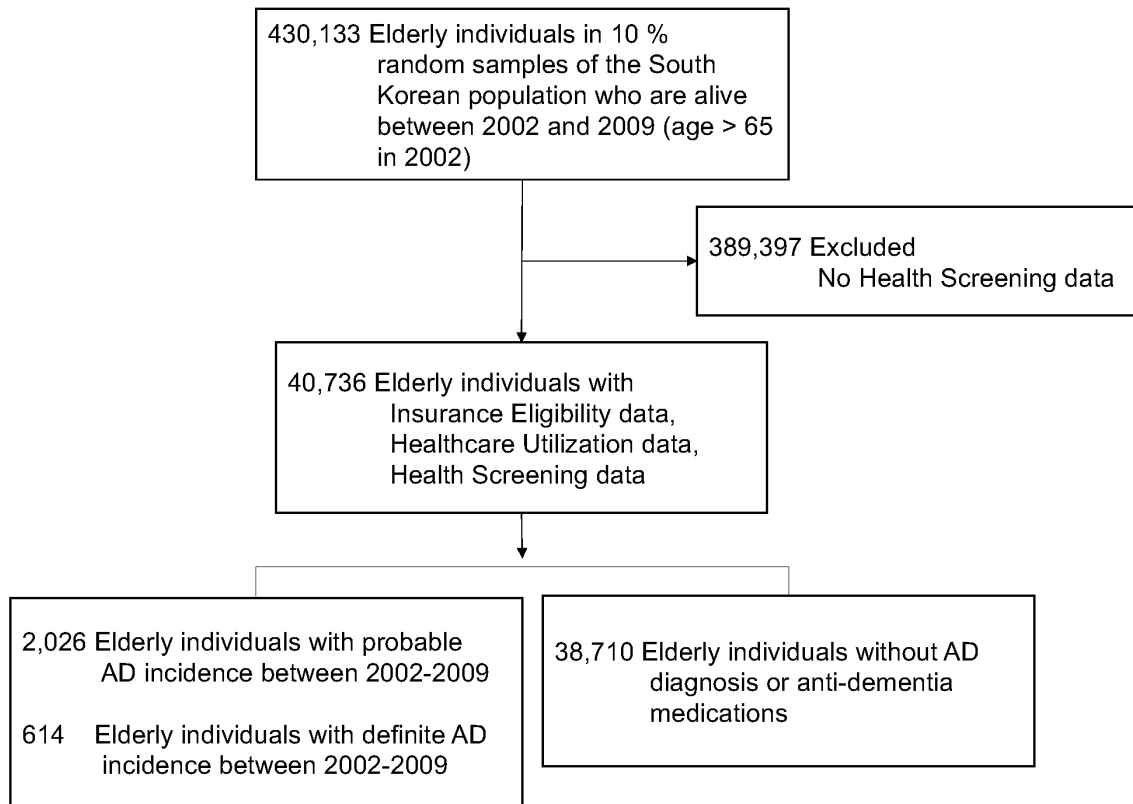
408

409 **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

410

411 **Figure 1. Consort Diagram.**

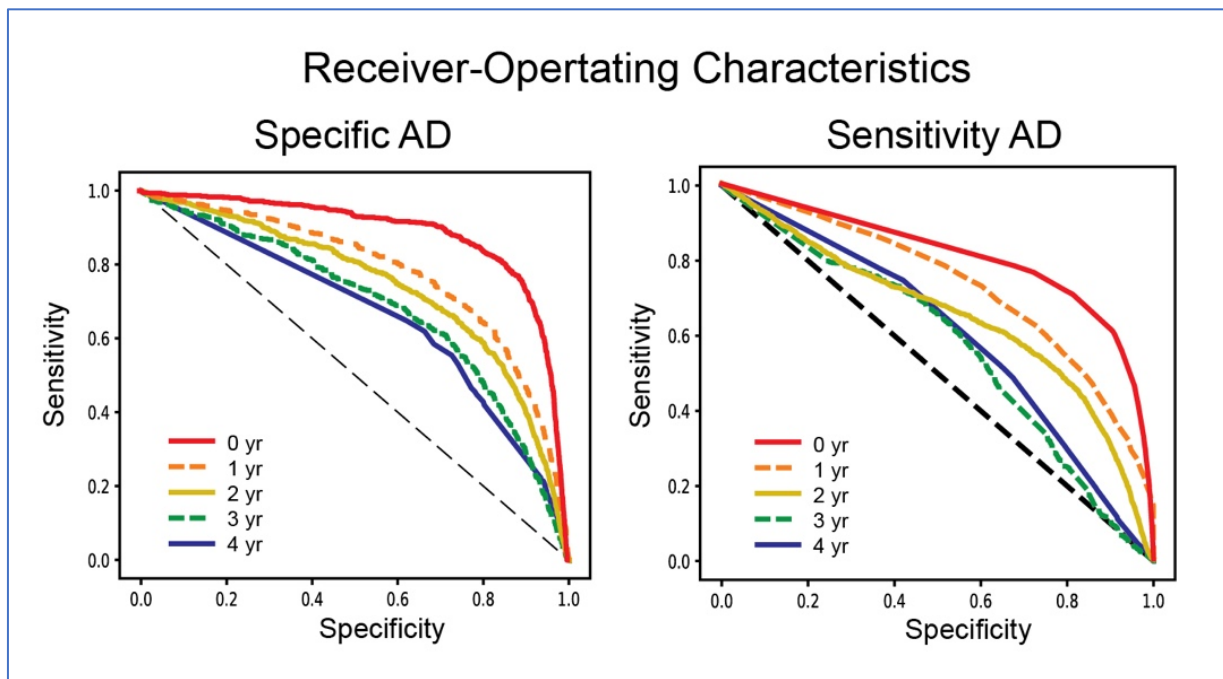


412

413



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

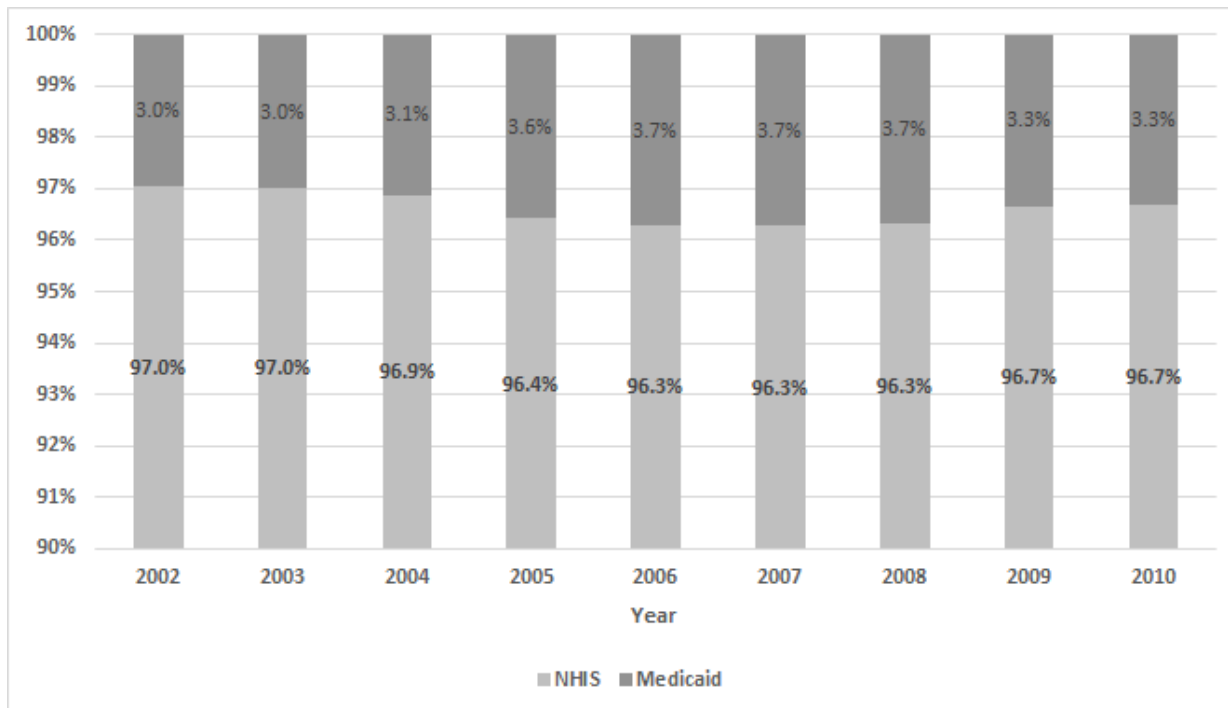


418

419

420 **Supplementary Materials**

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



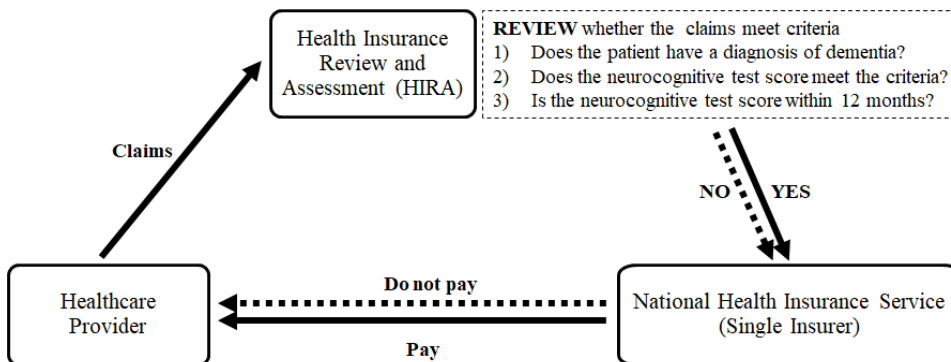
423

424

425

426

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



428

429  
430

**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
Income level	ordinal	1 : 2 : 3 : 4 : 5 :
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	NUM(3)	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

431

		2 : below 60g of alcohol 3 : below 90g of alcohol 4 : over 120g of alcohol
--	--	----------------------------------------------------------------------------------