Title. Predicting Alzheimer's Disease Diagnosis, a Decade or more Years before Onset using the Electronic Health Record and Random Forest Machine Learning Models

Authors. Sanya B. Taneja[1], Richard D. Boyce[1,2], Scott A. Malec[3], Steven M. Albert[4], C. Elizabeth Shaaban[5,6], Arthur S. Levine[7,8], Paul Munro[9,10], Jiang Bian[11], Jie Xu[11], Demetrius Maraganore[12], Karen Schliep[13], Jonathan C. Silverstein[2], Michelle Kienholz[14], Helmet T. Karim[15,16]

Affiliations.

1. Intelligent Systems Program, University of Pittsburgh
2. Department of Translational Informatics, University of Pittsburgh
3. Division of Translational Informatics, University of New Mexico, Albuquerque, NM
4. Department of Behavioral and Community Health Sciences, University of Pittsburgh
5. Department of Epidemiology, University of Pittsburgh
6. Alzheimer's Disease Research Center, University of Pittsburgh
7. Department of Neurobiology, University of Pittsburgh
8. Brain Institute, University of Pittsburgh
9. School of Computing and Information, University of Pittsburgh
10. Intelligent Systems Program, University of Pittsburgh
11. Department of Health Outcomes and Biomedical Informatics, University of Florida
12. Department of Neurology, Tulane University School of Medicine
13. Department of Family and Preventive Medicine, University of Utah Health
14. Clinical and Translational Science Institute, University of Pittsburgh
15. Department of Psychiatry, University of Pittsburgh
16. Department of Bioengineering, University of Pittsburgh

Corresponding Author:

Richard D. Boyce, PhD

Associate Professor of Biomedical Informatics

University of Pittsburgh

Pittsburgh, PA

E-mail: rdb20@pitt.edu

Keywords: Alzheimer's disease, prediction, machine learning, Electronic Health Record, Random Forest

Research in Context:

Systematic review: We reviewed the relevant literature using traditional methods (e.g., PubMed) and identified that previous studies have predicted Alzheimer's Disease (AD) early with the help of the Electronic Health Record (EHR). This may help identify patients early in the course of disease.

Interpretation: We demonstrate predicting AD 10 years prior to diagnosis using the EHR and identify important predictive factors.

Future Directions: More work is needed to build and validate these types of models across hospital systems over multiple timescales. These models could be directly embedded into EHR systems.

Abstract

Introduction: There is need to detect and intervene in pre-clinical phases of Alzheimer's disease (AD). Electronic health records (EHRs) may help predict AD using machine learning methods. **Methods:** We identified EHRs for 19,473 cases with AD and 111,922 controls. Records spanned 10 or more years prior to AD diagnosis. We trained a random forest model (employing 5-fold cross-validation with 2,499 features) to predict AD 10 years prior to its onset using a 75/25% train/test split and then computed permuted feature importance. **Results:** On the test set achieved, the model achieved an area under the ROC curve of 0.80 and area under the precision-recall curve of 0.55. Feature importance identified factors associated with AD, including age, sex, race, ethnicity, BMI, cardiovascular diseases, inflammation, pain, sleep, trauma, other neurodegenerative disorders, diuretics, colon-related disorders and procedures, seizures, and vitamin B12. **Discussion:** This work contributes to knowledge about EHR-based prediction of AD 10 years prior to onset, which could help predict AD and inform prevention/early intervention. The value of this work is not necessarily predicting AD, given what is now known about its biology and blood biomarkers. Rather, the work inspires further examination of informatics methods for predicting disease diagnoses a decade or more prior to clinical diagnoses. The code used to build and evaluate the model is located at https://github.com/dbmi-pitt/ad_prediction_PLP/tree/main

Introduction[a]

Alzheimer's disease (AD) affects 6.5 million people in the United States, growing to an expected 13.8 million in 2060, and is the seventh leading cause of death (1). There are currently only limited treatment options for individuals with AD that have modest effects on cognitive function (2-4). There is a critical need to develop tools to detect AD earlier, identify protective and risk factors of AD, and intervene earlier, ideally prior to the onset of AD pathology. Indeed, sufficient evidence indicates that amyloid positivity, levels of some experimental biomarkers(5-8), and even some clinical signs precede clinical diagnosis of AD by decades (9, 10). The electronic health record (EHR) is a promising source of patient data that, combined with machine learning, can inform or enable clinical interventions that could help improve the health of patients at high risk of AD.

For example, many clinical trials are currently focusing on amyloid-clearing treatments. These studies have shown adequate clearance of amyloid but only relatively minor improvements in cognitive function. Early detection could also help inform more holistic treatment or prevention strategies, where teams of healthcare workers can work in tandem to treat several risk factors with machine-guided and personalized precision(11). Numerous factors contribute to cognitive impairment and rapid cognitive decline – either as part of the central mechanisms (e.g., affecting amyloid or tau) or by reducing cognitive reserve. The Electronic Health Record (EHR) is highly effective in facilitating predictive modeling, given that it accumulates data in real time over extended periods as an integral component of clinical practice. The nature of EHR data makes it exceptionally compatible with machine learning techniques, enabling researchers to construct models that pinpoint factors associated with AD.

Several studies have focused on predicting AD diagnosis using EHR data(12-14). One study identified 2,640 AD patients and 40,736 records to predict AD at 1, 2, 3, and 4 years using EHR data; these models achieved an area under the receiver operating characteristic curves (AUCs) of 0.68-0.78(13). Another group identified patients with mild cognitive impairment (MCI) using clinical notes and deep learning, achieving an AUC of 0.75(12). Yet another group used graph neural networks to predict AD in a sample of 1.64 million records with 8,174 AD patients and achieved an area under the precision-recall curve (AUPRC) of 0.39-0.71(14).

Two more recent studies have evaluated data up to 5 years in advance. One study used 35,866 cases and 179,330 controls in an administrative claims dataset and predicted AD 4-5 years in advance

[a] **Abbreviations**: AUPRC=area under the precision-recall curve; CDM=Common Data Model; OHDSI=Observational Data Health Sciences and Informatics; PFI= permutation feature importance

with an AUC of 0.69(15). Another evaluated the effectiveness of predicting AD diagnosis 5 years prior to onset, which is the furthest back to date(16), using 23,835 cases and 1,038,643 controls, making it one of the largest to date. They achieved high prediction accuracy – AUC of 0.85 for 5 years prior to onset – and identified several predictive factors. An even more recent study identified predictors up to 7 years prior to onset(17) achieving an AUC of 0.72 in individuals with more accurately defined AD diagnosis through a memory and aging center.

In this study, we identified a large set of patient records spanning more than a decade of continuous care for 19,473 cases and 111,922 controls. The goal of the study was to use machine learning to predict AD 10 or more years prior to diagnosis (e.g., a patient with a diagnosis in 2016 had to have EHR data from at least 2006 or earlier). To the best of our knowledge, this is the first study to investigate the EHR at such a timescale with such a large set of cases.

Methods

Extracting Electronic Health Record Data

An EHR dataset was extracted from the UPMC EHR using the Health Record Research Request service(18). The University of Pittsburgh Institutional Review Board approved this study as exempt. Data were de-identified to a HIPAA-limited dataset and contained information on patient demographics, allergies, diagnoses, family history, immunizations, lab results, medications, medical history, procedures, social history on alcohol and tobacco use, and vitals. The data included all types of visits (e.g., inpatient, outpatient, emergency, etc.). We used the following diagnostic codes to identify diagnoses of dementia or AD (International Classification Code of Disease (ICD)-9 and 10 codes 290.0, 290.1, 290.10, 290.11, 290.12, 290.13, 290.20, 290.21, 290.3, 294.20, 294.21, F03.90, F03.91, G30, G30.0, G30.1, G30.8, and G30.9). The case cohort was patients diagnosed with AD or dementia from 1/1/2016 to 12/31/2021 who had at least one UPMC EHR encounter 10 or more years before this diagnosis.

This extraction resulted in 19,473 cases and 111,922 controls matched on the index date—or the date of diagnosis in AD cases. In the final pull of data, controls were excluded for exposure to acetylcholinesterase inhibitor or memantine or any of these diagnostic codes: 331.0, G30, G30.0, G30.1, G30.8, G30.9, F00, F00.0, F00.1, or F00.2 (Alzheimer’s Disease).

Preprocessing of the EHR data

All EHR data were transformed to the Observational Data Health Sciences and Informatics (OHDSI)’s Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) (version 5) with standardized vocabularies for each data domain. The final data were stored in the CDM person (demographics), condition occurrence, drug exposure, measurement, observation, procedure occurrence, visit occurrence, condition era, and drug era tables. The condition era and drug era tables are chronological periods of condition occurrence and drug exposure, respectively. We

manually mapped codes for family history concepts, allergies, and encounter types to standardized vocabulary concepts as these were not originally present in the data. We also used term-matching to find standardized concepts for unmapped National Drug Code drugs, which resulted in a standardized representation of the data using this and other vocabularies, such as the Systematized Nomenclature of Medicine, RxNORM, and the Logical Observation Identifiers Names and Codes. It also enabled using the OHDSI Patient-Level Prediction R package for predictive analysis(19).

Training of Random Forest Machine Learning Model

We trained a random forest model to predict AD diagnosis (cases vs. controls) 10 or more years prior to the AD diagnosis date (observation window). We chose a random forest model as random forests have been shown to outperform several other methods on tabular data including deep nets(20). The data from the start of a person's EHR history to 10 years prior to diagnosis of AD were collapsed into features for the random forest model. For each measurement, condition, procedure, drug, visit, and observation record in the data, the values in the observation window were converted to binary values representing present/absent in the EHR. The values for each measurement were included along with an indicator as to whether the measurement value was below, within, or above the normal range. We also included as features in the model the age groups, race and ethnicity categories, number of visits, Charlson Comorbidity Index score, Diabetes Complication Severity Index, $CHADS_2$, and CHA_2DS_2-VASc scores.

We used a 75% training and 25% test split of the data to train the model and evaluate performance, respectively. The training was conducted via 5-fold cross-validation on the training set only. The number of features available for prediction was 20,931. Using a minimum covariate occurrence of >= 0.001 in target population reduced the feature set to 2,499 included in the final model. We used the OHDSI Patient-Level Prediction R package (version 4.3.1) to train and test the random forest model(19). The R package is specifically designed to train and validate models for data represented in the OMOP CDM. We conducted hyper-parameter optimization over the following: number of trees (50, 100, 500, 1000), depth of trees (4, 10, 17, 25, 50), and number of features per tree (5, 20, 100, 200, and square root of the total number of features). The model was optimized using the AUC value.

Evaluation of Random Forest Model

We evaluated the random forest model over the test set using AUC and AUPRC, accuracy (# predicted accurately/total), positive predictive value (PPV, # AD predicted accurately/total # predicted AD), negative predictive value (NPV, # controls predicted accurately/total # predicted controls), sensitivity (# AD predicted accurately/total AD), and specificity (# controls predicted accurately/total controls). The optimal threshold was chosen to maximize balanced F-measure. The test set was only used to evaluate these after completing all training.

Ranking Feature Importance

We conducted a permutation feature importance (PFI) analysis on both train and test data. PFI permutes each feature, computes the AUC, and then compares the result with the original model's AUC to the permuted model's AUC. The PFI for each feature is then calculated as the original AUC minus permuted AUC. Larger reductions in AUC indicate a more important feature. We then sorted the PFI score of each feature used by the model in descending order and report only those features that were above the rank above the optimal number of features to sample at each split.

Results

The characteristics of AD cases and controls are shown in Table 1. The cases and controls are predominantly non-Hispanic white, with >60% Females and a large age distribution.

The performance of the random forest model on the training data was AUC of 0.85 and AUPRC of 0.68. On the test set, performance was AUROC 0.80 and AUPRC of 0.55. The optimal prediction threshold on the test dataset was 0.25 which maximized the balanced F measure at a value of 0.47. At this threshold, the model achieved a PPV of 0.53, NPV of 0.90, sensitivity of 0.43, and specificity of 0.93 (see Figures 1 and 2). The optimal random forest model used for prediction included 1000 trees, a maximum depth of 25, and a number of features to sample at each split equal to 200. Since the number of features sampled for each split was 200, Table 2 shows the same number of features after ranked ordering by PFI score in descending order. This included a several categories of variables:

- Demographics (age, sex, race, ethnicity, and height);
- Measurements (body height, weight, BMI, blood pressure, heart rate)
- Co-morbidity scores (Charlson index, Diabetes comorbidity score, CHADS2)
- Procedures and health encounter (colonoscopy, mammography, emergency department visit, etc.)
- Cardiovascular factors, procedures, and related drugs (e.g., aneurysm, arthropathy, hypertensive disorder, beta-blocking agents, etc.);
- Neurologic and mental health disorders (seizure; psychotic disorder)
- Physical trauma (e.g., head injury, fall, traumatic injury);
- Gastrointestinal disorders and related drugs (e.g., colonic lesion, diverticulosis, drugs for peptic ulcer and gastro-oesophageal reflux disease, etc.);
- Inflammatory disorders and related drugs;
- Miscellaneous (flu and cholera vaccines, opioids, cough suppressants, B12, etc.).

Discussion

We achieved a relatively high-performance random forest model with an AUC of 0.80 on the test set that is comparable with previous EHR prediction models but extends the prediction window to 10 years or more prior to diagnosis. This is the first demonstration of such performance over such a lengthy time period of EHR. Other models have achieved similar performance(12-14), and a recent study showed an AUC of 0.85 for 5 years before diagnosis(16).

The most comparable analysis was conducted by Li et al.(16), which evaluated multiple time windows from 0 to 5 years before onset and predicted AD with high accuracy (AUC 0.94 and 0.85 for 0 and 5 years, respectively). Comparing their model with ours, we achieved similar PPV (0.53 compared to their model 0.49) and NPV (0.90 compared to their model 0.93) showing that even at 10 years out, we can achieve similar predictive values. As in their study, we can pinpoint controls with a fair degree of accuracy, but it is more difficult to identify those with AD. In part, this may be due to the EHR's capacity to accurately diagnose AD cases, where past studies have shown high accuracy but not perfect accuracy with PPV of ~90% for EHR vs. gold standard(21). One recent study used expert guided diagnosis of AD and found similar performance when using the EHR to predict AD onset 7 years prior(17). Either way, EHR case data are important as they help identify potentially high-risk patients for monitoring (e.g., cognitive screening), lifestyle interventions that could delay progression, and screening or confirmatory biomarker and neuroimaging studies. Much work remains prior to implementing such models in clinical settings, including improving the predictive capacity, but this is a first step in that direction.

In general, many factors were predictive, including demographic, cardiovascular, inflammatory, and neurological as well as certain procedures. These are well in line with past models of AD that have demonstrated many of these as risk factors(22, 23), including age, weight/BMI, height (which may be an indicator of frailty in old age(24)), head injury, cardiovascular disease (and related factors), hearing or vision impairment, untreated depression and psychosis, and sedentary lifestyle.

A recent study identified many of these factors, including features that overlapped, such as age, cerebrovascular disease, mood disorder, diabetes, ischemia, weight/BMI, race, and confusion(16). They additionally found factors we did not find, including fatigue, malaise, debility, and epilepsy (16). Using a deep phenotyping approach combined with EHR data, another group showed similar factors that predicted risk, including hypertension, hyperlipidemia, osteoporosis, and mental health disorders(25). It is important to note that some of these may be related to social determinants of health – for example, those who undergo preventative screenings are generally wealthier, educated, have insurance, and may be more likely to adopt healthy behaviors. Some screenings are also associated with age, which itself is predictive of AD. Li et al.(16) similarly identified features found using our modeling approach, including age, diabetes, mammography, blood pressure, hypertension, mood disorders, seizure, vascular disorders, confusion, sleep apnea, and race. They additionally identified routine medical exams and naproxen use, which we did not. While they

looked at the top 20 most important features, we listed a larger set of features based on the feature importance. In concurrent research conducted by our team, as reported in a recent study by Malec et al. (26), sleep apnea has been recognized not only as a confounding variable that complicates the relationship between depression and AD but also as a potentially modifiable risk factor for AD. Furthermore, sleep apnea may independently contribute to the risk profile for AD.

This convergence helps identify features that may be important for predicting AD. However, prediction is not causal, and none of these models, including ours, identify causal mechanisms for any identified features. Indeed, we present the features and their PFI data simply to demonstrate the feasibility of applying machine learning tools to EHR data to identify and predict early risk; we cannot rank these features or discuss their relevance to future AD diagnosis as part of this project. In a recent study(17) that used EHR to predict AD 7 years prior to onset, they additionally conducted matched analysis and prediction and found that certain features were non-specific, e.g., hypertension, as they were no longer predictive after conducting matching.

Various approaches have been suggested for increasing the time window for preventive brain health services(27, 28), such as early risk assessment through *APOE* genotyping and lifestyle factors that may contribute to risk (28). Still, a major challenge is identifying *who* should receive these risk assessments. Embedding machine learning and risk prediction models directly into the EHR could facilitate this process. The EHR could help identify high-risk patients so that they could undergo cognitive testing and eventually biomarker screening, with those showing early changes in brain health referred to specialty clinics. An embedded EHR tool could also aid in continuous passive monitoring of risks to brain health. Our results complement two recent findings that estimated the timing of various AD markers to time of onset in the US(29) and China(30) that showed peripheral and central markers of amyloid accumulate 15 to 19 years prior to diagnosis, markers of tau, (peripheral) neurodegeneration, and inflammation 10 to 14 years prior, and markers of neural atrophy 7 to 9 year prior. This is especially relevant with the advent of highly accurate plasma markers of AD.

It is important to note that while certain features identified may be confusing, these data are 10 years prior to the onset of AD. Thus, features like ‘impaired cognition’ are relevant predictive factors. In addition, features like ‘psychotic disorder’ were not excluded as our goal was to predict AD diagnosis even if psychotic disorders may be a sign of existing AD. Ultimately, the EHR is not perfect, and diagnostic codes are not always accurate, but our model conducts these predictions under such assumptions.

There are several limitations of the current work. The model should be validated in independent data and be further validated clinically for utility. While the current model uses data from a single large health network that is remarkable for having a large number of patients who have remained within the same geographic area for many years, many health system EHRs have a very limited number of samples or years of continuous records, which could limit generalizability as an EHR-embedded tool. In addition, overfitting could be problematic; it is possible that the current

model is more predictive of the current data compared with another dataset. The EHR also does not have certain types of information embedded in it, which is a limitation of the data itself. We used records from 2016 through 2021 to identify AD patients to leverage the more robust AD dementia diagnostic guidelines available in that time frame to improve case identification. In addition, EHR diagnoses are thought to fairly accurately be associated with gold-standard diagnostics(21). While the current model predicts AD, extending this to other dementias may be important for performance validation. The ICD codes used here did not correspond fully with validated EHR phenotypes, which may be a limitation. By only looking at those patients with at least 10 years of data and by using a case-control design, there may be issues related to bias of the features and samples. The primary way to address such problems is to test these models in multiple EHR systems.

In conclusion, our work is significant for showing that it is feasible to predict AD diagnosis 10 years prior to onset using EHR data. The random forest model achieved a relatively high AUC and identified risk factors that predicted future AD diagnosis. The AD risk factors set included known and potentially novel risk factors. While work remains to be done on the refinement and generalizability of these models, the value of this work is not necessarily predicting AD, given what is now known about its biology and blood biomarkers. Rather, the work inspires further examination of informatics methods for predicting disease diagnoses a decade or more prior to clinical diagnoses.

Acknowledgments. We would like to thank Elizabeth Wu for her contribution to this manuscript.

Funding. US National Institutes of Health grants supported this research, including T15LM007059, K99LM013367, R00LM013367, UL1TR001857, K01AG071849, and T32AG055381, as did a Pitt Momentum Fund Teaming Grant from the University of Pittsburgh Office of the Provost. The views expressed here do not necessarily reflect the views of the National Institutes of Health.

Conflicts of Interest. CES reports that she is Chair of the ISTAART PIA to Elevate Early Career Researchers (PEERs) and Co-chair of the Sex and Gender Special Interest Group of the Diversity and Disparities PIA in ISTAART. All other authors report no conflicts of interest.

References

1. 2022 Alzheimer's disease facts and figures. Alzheimers Dement 2022; 18:700-789

2. Effect of reductions in amyloid levels on cognitive change in randomized trials: instrumental variable meta-analysis. BMJ 2022; 378:o2094

3. Budd Haeberlein S, Aisen PS, Barkhof F, et al: Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease. J Prev Alzheimers Dis 2022; 9:197-210

4. van Dyck CH, Swanson CJ, Aisen P, et al: Lecanemab in Early Alzheimer's Disease. N Engl J Med 2023; 388:9-21

5. Hansson O, Blennow K, Zetterberg H,Dage J: Blood biomarkers for Alzheimer's disease in clinical practice and trials. Nat Aging 2023; 3:506-519

6. Haque R, Watson CM, Liu J, et al: A protein panel in cerebrospinal fluid for diagnostic and predictive assessment of Alzheimer's disease. Sci Transl Med 2023; 15:eadg4122

7. Rafii MS,Aisen PS: Detection and treatment of Alzheimer's disease in its preclinical stage. Nat Aging 2023; 3:520-531

8. Walker KA, Chen J, Shi L, et al: Proteomics analysis of plasma from middle-aged adults identifies protein markers of dementia risk in later life. Sci Transl Med 2023; 15:eadf5681

9. Bateman RJ, Xiong C, Benzinger TL, et al: Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med 2012; 367:795-804

10. Perl DP: Neuropathology of Alzheimer's disease. Mt Sinai J Med 2010; 77:32-42

11. Fosnacht AM, Patel S, Yucus C, et al: From Brain Disease to Brain Health: Primary Prevention of Alzheimer's Disease and Related Disorders in a Health System Using an Electronic Medical Record-Based Approach. J Prev Alzheimers Dis 2017; 4:157-164

12. Fouladvand S, Mielke MM, Vassilaki M, et al: Deep Learning Prediction of Mild Cognitive Impairment using Electronic Health Records. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2019; 2019:799-806

13. Park JH, Cho HE, Kim JH, et al: Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ Digit Med 2020; 3:46

14. Zhu W,Razavian N: Variationally regularized graph-based representation learning for electronic health records, 2021

15. Nori VS, Hane CA, Martin DC, et al: Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS One 2019; 14:e0203246

16. Li Q, Yang X, Xu J, et al: Early prediction of Alzheimer's disease and related dementias using real-world electronic health records. Alzheimers Dement 2023;

17. Tang AS, Rankin KP, Cerono G, et al: Leveraging electronic health records and knowledge networks for Alzheimer's disease prediction and sex-specific biological insights. Nat Aging 2024; 4:379-395

18. Visweswaran S, McLay B, Cappella N, et al: An atomic approach to the design and implementation of a research data warehouse. J Am Med Inform Assoc 2022; 29:601-608

19. Reps JM, Schuemie MJ, Suchard MA, et al: Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc 2018; 25:969-975

20. Shwartz-Ziv R,Armon A: Tabular data: Deep learning is not all you need. Information Fusion 2022; 81:84-90

21. Ponjoan A, Garre-Olmo J, Blanch J, et al: How well can electronic health records from primary care identify Alzheimer's disease cases? Clin Epidemiol 2019; 11:509-518

22. Anstey KJ, Ee N, Eramudugolla R, et al: A Systematic Review of Meta-Analyses that Evaluate Risk Factors for Dementia to Evaluate the Quantity, Quality, and Global Representativeness of Evidence. J Alzheimers Dis 2019; 70:S165-S186

23. Hersi M, Irvine B, Gupta P, et al: Risk factors associated with the onset and progression of Alzheimer's disease: A systematic review of the evidence. Neurotoxicology 2017; 61:143-187

24. Ji S, Lee E, Kim BJ, et al: Height loss as an indicator of ageing through its association with frailty and sarcopenia: An observational cohort study. Arch Gerontol Geriatr 2023; 110:104916

25. Tang AS, Oskotsky T, Havaldar S, et al: Deep phenotyping of Alzheimer's disease leveraging electronic medical records identifies sex-specific clinical associations. Nat Commun 2022; 13:675

26. Malec SA, Taneja SB, Albert SM, et al: Causal feature selection using a knowledge graph combining structured knowledge from the biomedical literature and ontologies: A use case studying depression as a risk factor for Alzheimer's disease. J Biomed Inform 2023; 142:104368

27. Solomon A, Stephen R, Altomare D, et al: Multidomain interventions: state-of-the-art and future directions for protocols to implement precision dementia risk reduction. A user manual for Brain Health Services-part 4 of 6. Alzheimers Res Ther 2021; 13:171

28. Altomare D, Molinuevo JL, Ritchie C, et al: Brain Health Services: organization, structure, and challenges for implementation. A user manual for Brain Health Services-part 1 of 6. Alzheimers Res Ther 2021; 13:168

29. Li Y, Yen D, Hendrix RD, et al: Timing of Biomarker Changes in Sporadic Alzheimer's Disease in Estimated Years from Symptom Onset. Ann Neurol 2024;

30. Jia J, Ning Y, Chen M, et al: Biomarker Changes during 20 Years Preceding Alzheimer's Disease. N Engl J Med 2024; 390:712-722

Table 1. Demographics, medical history, and medication use for the cases vs. controls. Age groups are currently calculated from date of Alzheimer's diagnosis for cases and from the first visit after 2016 for controls.

Characteristic	% in controls (n =111,922)	% in cases (n = 19,473)
Age group		
45 - 49	0.31	0
50 - 54	1.45	0.74
55 - 59	2.96	1.53
60 - 64	5.40	2.86
65 - 69	9.89	5.40
70 - 74	16.21	9.15
75 - 79	22.04	14.96
80 - 84	20.49	19.90
85 - 89	13.71	23.14
90 - 94	5.94	16.83
95 - 99	1.41	4.87
100 - 104	0.18	0.58
105 - 109	0	0.03
Gender: female	60.38	64.12
Race		
race = Asian	0.14	0.11
race = Black or African American	5.43	11.53
race = White	91.86	87.50
race = Unknown	2.21	0.53
race = Native Hawaiian or Other Pacific Islander	0.01	0.01
race = American Indian	0.09	0.08
race = Asian Indian	0.09	0.08
race = Chinese	0.13	0.10
race = Japanese	0.02	0.03
race = Korean	0.02	0.03
race = Vietnamese	0.01	0.01
Ethnicity		
ethnicity = Hispanic or Latino	0.17	0.22
ethnicity = Not Hispanic or Latino	94.81	95.84
ethnicity = Not specified, declined, or unknown	5.02	3.94
Medical History		
Dementia	0.41	100.00

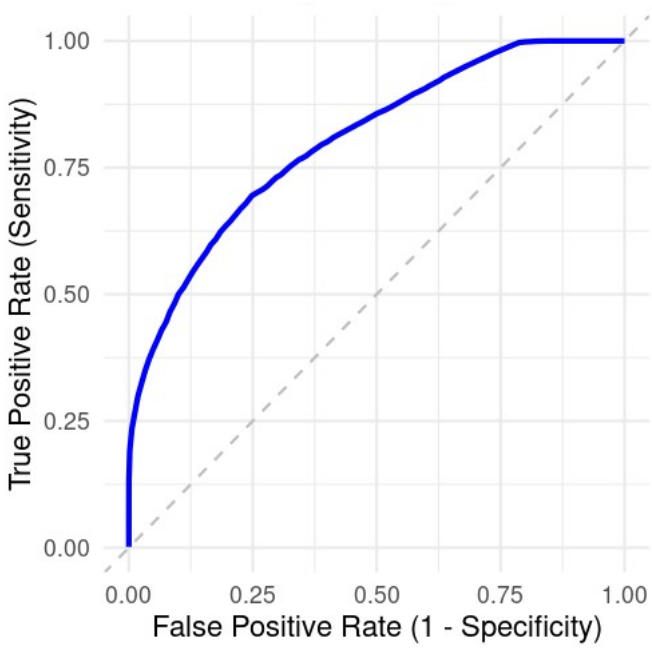


Figure 1. ROC curve for the AD random forest machine learning model when ran on the test set, plotting the false positive to true positive rate at various thresholds. The threshold that optimized the balanced F-measure was 0.25 which led to an F-measure of 0.47, sensitivity of 0.43, specificity of 0.93, PPV of 0.52, and NPV of 0.90.

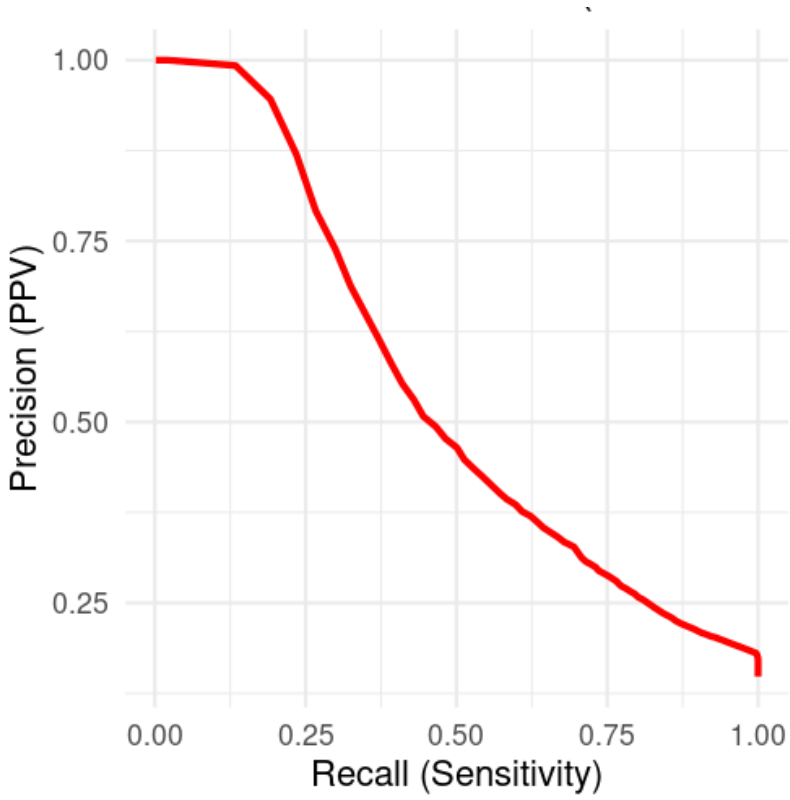


Figure 2. Precision-recall curve for the AD random forest machine learning model plotting the precision and recall at various thresholds.

Table 2. Top 200 features after sorting the Permuted feature importance (PFI) in descending order. Note that the PFI does not indicate directionality of importance.

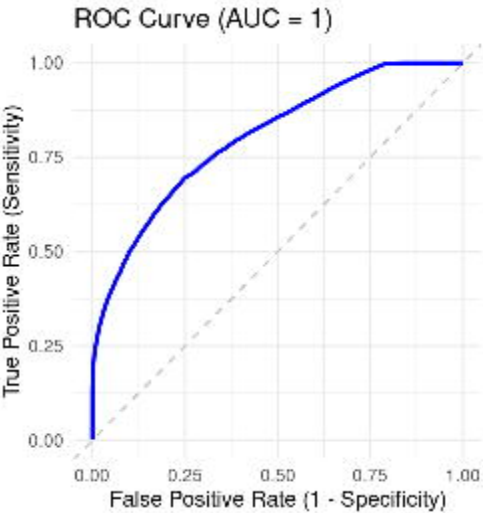
PFI	Covariate Name
0.321183	ethnicity = not hispanic or latino
0.086534	age group: 75 - 79
0.060969	age group: 70 - 74
0.049533	age group: 85 - 89
0.039225	age group: 65 - 69
0.029863	drug_exposure: seasonal trivalent influenza vaccine, adjuvanted, preservative free
0.022280	age group: 90 - 94
0.020494	age group: 60 - 64
0.018827	age in years
0.018546	race = black or african american
0.011382	age group: 55 - 59
0.011277	race = unknown
0.010758	condition_era: mammography
0.010471	condition_occurrence: mammography
0.007273	age group: 95 - 99
0.006660	measurement value: body weight (pound (us))
0.005640	chads2vasc
0.005214	measurement: body weight
0.004854	measurement: body weight measured
0.004816	chads2
0.004500	measurement value: body weight measured (pound (us))
0.004282	drug_exposure: immune globulin, unspecified formulation
0.004257	condition_occurrence: inflammatory disorder
0.004145	drug_exposure: influenza virus vaccine, unspecified formulation
0.003542	condition_era: inflammatory disorder
0.003512	condition_era group: pain
0.003258	condition_era: disorder of soft tissue
0.003117	condition_occurrence: disorder of soft tissue
0.002711	gender = male
0.002542	condition_era group: pain finding at anatomical site
0.002533	condition_era group: peripheral nerve disease
0.002287	drug_exposure: cholera, bivwc
0.002148	diabetes comorbidity severity index (dcsi)
0.001492	condition_era group: measurement finding outside reference range
0.001474	condition_era group: arthropathy
0.001444	procedure_occurrence: emergency department visit for the evaluation and management of a patient, which requires these 3 key components: an expanded problem focused history; an expanded problem focused examination; and medical decision making of moderate complexity. counseling

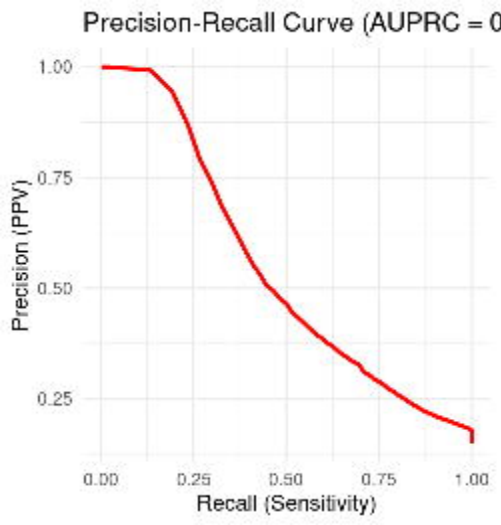
0.001419	condition_occurrence: essential hypertension
0.001344	condition_occurrence: sleep disorder
0.001341	condition_era group: hypertensive disorder
0.001326	condition_era: sleep disorder
0.001144	condition_era group: essential hypertension
0.001121	condition_era: essential hypertension
0.001106	condition_occurrence: injury of head
0.001095	age group: 100 - 104
0.001055	condition_era: injury of head
0.001007	drug_exposure: dtp-haemophilus influenzae type b conjugate vaccine
0.000964	condition_occurrence: arthropathy
0.000947	condition_occurrence: pneumonia
0.000911	condition_occurrence: skin finding
0.000903	condition_era: skin finding
0.000850	condition_era group: peripheral neuritis
0.000838	condition_era: pneumonia
0.000790	procedure_occurrence: mammography; bilateral (deprecated)
0.000770	condition_era group: pain of truncal structure
0.000767	measurement: diastolic blood pressure
0.000763	condition_era: arthropathy
0.000692	condition_era group: inflammation of specific body systems
0.000684	procedure_occurrence: unlisted diagnostic radiographic procedure
0.000638	condition_era group: organic mental disorder
0.000605	measurement: systolic blood pressure
0.000601	condition_era group: traumatic injury
0.000589	condition_era group: vascular disorder
0.000584	condition_era group: measurement finding above reference range
0.000555	condition_era group: impaired cognition
0.000518	observation: average systolic blood pressure
0.000513	observation: average diastolic blood pressure
0.000495	measurement: heart rate
0.000490	condition_era group: mononeuropathy
0.000475	condition_era group: pneumonitis
0.000452	drug_era group: nervous system
0.000450	condition_era group: inflammatory disorder of head
0.000433	drug_era group: cardiovascular system
0.000426	charlson index - romano adaptation
0.000404	procedure_occurrence: emergency department visit for the evaluation and management of a patient, which requires these 3 key components: a detailed history; a detailed examination; and medical decision making of moderate complexity. counseling and/or coordination of care with o
0.000403	condition_era group: neuropathy
0.000358	condition_era group: pneumonia
0.000357	condition_era group: psychotic disorder
0.000354	condition_era group: heart disease
0.000346	condition_era group: inflammatory neuropathy

0.000336	condition_era group: inflammatory disorder of the respiratory system
0.000334	condition_era group: lung consolidation
0.000330	drug_era group: alimentary tract and metabolism
0.000321	condition_era: sleep apnea
0.000296	condition_era group: inflammatory disorder of the respiratory tract
0.000294	condition_era group: sleep apnea
0.000286	drug_era group: anesthetics
0.000285	drug_era group: anesthetics, general
0.000279	condition_occurrence: altered sensation of skin
0.000278	condition_era group: apnea
0.000271	condition_occurrence: sleep apnea
0.000262	drug_era group: salicylic acid and derivatives
0.000259	condition_era group: complication associated with device
0.000256	drug_era group: agents for treatment of hemorrhoids and anal fissures for topical use
0.000255	drug_era group: thiazides, combinations with other drugs
0.000255	condition_era group: inflammation of specific body organs
0.000252	drug_era group: otologicals
0.000251	measurement: body mass index (bmi) [ratio]
0.000251	condition_era group: altered sensation of skin
0.000250	drug_era group: hypnotics and sedatives in combination, excl. barbiturates
0.000244	condition_occurrence: cough
0.000243	drug_era group: other otologicals
0.000241	condition_era group: increased lipid
0.000240	condition_occurrence: hyperlipidemia
0.000237	condition_era: altered sensation of skin
0.000228	drug_era group: hypnotics and sedatives
0.000228	drug_era group: other analgesics and antipyretics
0.000225	condition_era: cough
0.000222	condition_era group: colonic lesion
0.000221	drug_era group: other agents against amoebiasis and other protozoal diseases
0.000217	condition_era: seizure
0.000216	condition_era: hyperlipidemia
0.000215	drug_era group: analgesics and anesthetics
0.000212	drug_era group: synthetic anticholinergic agents in combination with analgesics
0.000208	drug_era group: antispasmodics in combination with analgesics
0.000207	procedure_occurrence: screening mammography, bilateral (two view film study of each breast) (deprecated)
0.000205	drug_era group: belladonna and derivatives in combination with analgesics
0.000203	condition_era group: seizure
0.000203	condition_era group: contusion
0.000200	condition_era group: finding related to sleep
0.000193	drug_era group: sensory organs
0.000192	condition_era group: structural disorder of heart

0.000191	condition_era group: cough
0.000187	drug_era group: analgesics
0.000187	condition_occurrence: disorder of lung
0.000185	condition_era group: diverticular disease
0.000182	condition_era group: hyperlipidemia
0.000181	drug_era group: other agents for treatment of hemorrhoids and anal fissures for topical use
0.000179	drug_era group: agents against amoebiasis and other protozoal diseases
0.000173	condition_era group: aneurysm of peripheral artery
0.000170	measurement: body height
0.000170	condition_era group: diabetes mellitus
0.000169	condition_occurrence: seizure
0.000167	condition_era group: ear, nose and throat disorder
0.000157	procedure_occurrence: colonoscopy, flexible; diagnostic, including collection of specimen(s) by brushing or washing, when performed (separate procedure)
0.000157	condition_era group: lipids abnormal
0.000154	observation: median bmi yearly
0.000154	condition_occurrence: vehicle accident
0.000153	condition_era group: mood disorder
0.000152	condition_occurrence: pain in limb
0.000151	drug_era group: oxazol, thiazine, and triazine derivatives
0.000150	condition_era group: closed wound
0.000145	drug_era group: cough suppressants and expectorants, combinations
0.000144	drug_era group: respiratory system
0.000139	condition_era: disorder of lung
0.000138	condition_era group: soft tissue lesion
0.000136	condition_era: pain in limb
0.000135	condition_era: vehicle accident
0.000134	drug_era group: acetic acid derivatives and related substances
0.000132	drug_era group: blood and blood forming organs
0.000132	drug_era group: opium derivatives and expectorants
0.000129	drug_era group: opioids
0.000128	drug_era group: drugs for peptic ulcer and gastro-oesophageal reflux disease (gord)
0.000128	condition_era group: pain in limb
0.000127	drug_era group: opioids in combination with non-opioid analgesics
0.000125	measurement value: body height (inch (us))
0.000122	drug_era group: antiinflammatory and antirheumatic products, non-steroids
0.000120	condition_era group: open wound
0.000120	drug_era group: corticosteroids and antiinfectives in combination
0.000120	condition_era group: type 2 diabetes mellitus
0.000117	condition_era: fall
0.000116	condition_era: organic mental disorder

0.000114	condition_era group: coronary arteriosclerosis
0.000114	condition_era: peripheral neuritis
0.000114	drug_era group: other cough suppressants and expectorants
0.000109	condition_era: coronary arteriosclerosis
0.000108	condition_era group: visual system disorder
0.000108	condition_era group: head and neck injury
0.000108	procedure_occurrence: neuroplasty and/or transposition; median nerve at carpal tunnel
0.000107	drug_era group: vitamin b12 (cyanocobalamin and analogues)
0.000107	condition_era group: victim of vehicular and/or traffic accident
0.000107	condition_occurrence: peripheral neuritis
0.000107	drug_era group: corticosteroids and antiinfectives in combination
0.000106	drug_era group: vitamin b12 and folic acid
0.000106	condition_era group: victim status
0.000104	condition_occurrence: backache
0.000104	condition_era group: sinusitis
0.000101	condition_era group: aneurysm of artery of trunk
0.000101	condition_era group: distention of artery
0.000101	condition_era: victim of vehicular and/or traffic accident
0.000099	drug_era group: beta blocking agents and other diuretics
0.000099	observation: history of clinical finding in subject
0.000098	condition_era group: arterial aneurysm
0.000097	condition_occurrence: accidental collision with stationary object
0.000097	condition_occurrence: disorder of bone
0.000097	condition_era group: infectious disease of lung
0.000097	condition_occurrence: infective pneumonia
0.000096	drug_era group: adrenergics in combination with corticosteroids or other drugs, excl. anticholinergics
0.000096	condition_era: disorder of bone
0.000095	condition_occurrence: type 2 diabetes mellitus
0.000095	condition_occurrence: fall
0.000094	drug_era group: ace inhibitors, combinations
0.000093	drug_era group: nicotinic acid and derivatives
0.000092	condition_era group: lower respiratory tract infection
0.000092	condition_era group: open wound of head
0.000092	condition_era group: distention of blood vessel
0.000091	condition_era: cataract
0.000091	condition_era group: mass of trunk
0.000090	condition_era group: diverticulosis of colon
0.000090	drug_era group: antiemetics and antinauseants
0.000089	condition_era group: diverticular disease of colon
0.000089	drug_era group: antiemetics and antinauseants


ROC Curve (AUC = 1)
1.00
0.75
0.50
0.25
0.00
True Positive Rate (Sensitivity)
0.00
0.25
0.50
0.75
1.00
False Positive Rate (1 - Specificity)


Precision-Recall Curve (AUPRC = 0
1.00
0.75
0.50
0.25
Precision (PPV)
0.00
0.25
0.50
0.75
1.00
Recall (Sensitivity)