Received: 18 February 2020 | Revised: 19 July 2020 | Accepted: 6 August 2020

DOI: 10.1002/lrh2.10246

RESEARCH REPORT

Learning Health Systems

Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records

Jie Xu[1] | Fei Wang[1] | Zhenxing Xu[1] | Prakash Adekkanattu[2] | Pascal Brandt[3] | Guoqian Jiang[4] | Richard C. Kiefer[4] | Yuan Luo[5] | Chengsheng Mao[5] | Jennifer A. Pacheco[5] | Luke V. Rasmussen[5] | Yiye Zhang[1] | Richard Isaacson[1] | Jyotishman Pathak[1]

[1]Department of Population Health Sciences, Information Technologies and Services, Weill Cornell Medicine, New York, New York, USA

[2]Information Technologies and Services, Weill Cornell Medicine, New York, New York, USA

[3]Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington, USA

[4]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

[5]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

Correspondence
Jyotishman Pathak, Weill Cornell Medicine, New York, NY 10065, USA.
Email: jyp2001@med.cornell.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Number: R01GM105688; National Institute of Mental Health, Grant/Award Number: R01 MH119177; National Institute of Mental Health, Grant/Award Number: R01MH105384

Abstract

Introduction: We sought to assess longitudinal electronic health records (EHRs) using machine learning (ML) methods to computationally derive probable Alzheimer's Disease (AD) and related dementia subphenotypes.

Methods: A retrospective analysis of EHR data from a cohort of 7587 patients seen at a large, multi-specialty urban academic medical center in New York was conducted. Subphenotypes were derived using hierarchical clustering from 792 probable AD patients (cases) who had received at least one diagnosis of AD using their clinical data. The other 6795 patients, labeled as controls, were matched on age and gender with the cases and randomly selected in the ratio of 9:1. Prediction models with multiple ML algorithms were trained on this cohort using 5-fold cross-validation. XGBoost was used to rank the variable importance.

Results: Four subphenotypes were computationally derived. Subphenotype A (n = 273; 28.2%) had more patients with cardiovascular diseases; subphenotype B (n = 221; 27.9%) had more patients with mental health illnesses, such as depression and anxiety; patients in subphenotype C (n = 183; 23.1%) were overall older (mean (SD) age, 79.5 (5.4) years) and had the most comorbidities including diabetes, cardiovascular diseases, and mental health disorders; and subphenotype D (n = 115; 14.5%) included patients who took anti-dementia drugs and had sensory problems, such as deafness and hearing impairment.
The 0-year prediction model for AD risk achieved an area under the receiver operating curve (AUC) of 0.764 (SD: 0.02); the 6-month model, 0.751 (SD: 0.02); the 1-year

model, 0.752 (SD: 0.02); the 2-year model, 0.749 (SD: 0.03); and the 3-year model, 0.735 (SD: 0.03), respectively. Based on variable importance, the top-ranked comorbidities included depression, stroke/transient ischemic attack, hypertension, anxiety, mobility impairments, and atrial fibrillation. The top-ranked medications included anti-dementia drugs, antipsychotics, antiepileptics, and antidepressants.

Conclusions: Four subphenotypes were computationally derived that correlated with cardiovascular diseases and mental health illnesses. ML algorithms based on patient demographics, diagnosis, and treatment demonstrated promising results in predicting the risk of developing AD at different time points across an individual's lifespan.

KEYWORDS

Alzheimer's disease, electronic health records, subphenotypes

1 | INTRODUCTION

Alzheimer's Disease (AD), a progressive and irreversible neurodegenerative disease with insidious onset, causes problems with memory, cognition, and behavior, and affects more than 5 million people in the United States ([1]; Lazarov and Tesco[2]). AD is a complex disease and the cause of 60% to 70% of cases of dementia. The biggest known risk factor for AD is age, but AD is not a normal part of aging. Depression, diabetes, head injuries, hearing loss, high cholesterol, inflammation, diabetes, and insulin resistance syndrome, oxidative damage, stress, and many other factors may lead to dementia ([3]; Lazarov and Tesco[2,4]). In the clinical diagnosis and treatment of AD, all of these risk factors need to be considered ([5,6]; Lazarov and Tesco[2]; Light and Lebowitz[7]).

Most existing data-driven research for AD progression, risk factors, and endophenotypes utilizes cohort study data, where rigorous and stricter inclusion-exclusion criteria are applied for participant enrollment. This leads to poor generalizability of the results in the real-world.[3,8,9] In addition, extensive follow-up to study the risk factors for AD is needed, because of the long prodrome before the suspicion of diagnosis and reverse causality. Certain symptoms or diseases may be a consequence of AD, rather than a risk factor. Moreover, it is likely that different factors influence the development of AD at different stages of life.[4] Prospective studies, in general, are resource- and time-intensive[10,11] and retrospective studies using existing data might offer complementary insights, improving our understanding for better diagnosis and treatment of AD. Electronic health records (EHRs), which capture a wide variety of important health events and include patients with the most severe manifestations of AD or disability, might be useful resources to identify new targets for intervention and help improve AD care.[8,12-14]

In this retrospective study, our goal was to identify distinct subphenotypes from routinely collected structured EHR data. Instead of targeting the precise biomarker-confirmed AD phenotype, we aimed to identify subgroups of patients given a clinical diagnosis of AD by their treating physician (which we categorize as “probable AD and related dementia”) using broader and more readily available EHR data to characterize their clinical features. The identification of subphenotypes might uncover “clusters” of patients with specific characteristics, allowing more precise treatment and improved care. Since it is currently uncertain whether EHRs are sufficient and complete to capture AD, we first assessed the effectiveness of variables extracted from longitudinal EHRs by training 0-year, 6-month, 1-year, and 2-year AD risk prediction models using multiple machine learning (ML) algorithms. Then, we applied the validated variables to computationally derive probable AD and related dementia subphenotypes.

2 | METHODS

2.1 | Data sources

This study used de-identified EHR data from Weill Cornell Medicine (WCM)/NewYork-Presbyterian Hospital (NYP)—a large, urban academic medical center located in the Upper East Side of Manhattan, New York. The study dataset was standardized using the Observational Medical Outcomes Partnership common data model (OMOP CDM). The current OMOP database at WCM/NYP has over 2.7 million patients mapped from both outpatient (EpicCare) and inpatient (Allscript Sunrise Clinical Manager) clinical care systems. The mapping process normalizes conditions to standardized medical vocabularies, including SNOMED-CT for conditions (diagnoses), RxNorm for medications, and LOINC for laboratory results. The OMOP vocabulary provides mappings to other classification systems, such as Anatomical Therapeutic Chemical (ATC) and VA Classes for medications. The study was approved by the WCM Institutional Review Board (protocol# 1408015423).

2.2 | Study cohort

An incidence-based, matched case-control design was used in this study. We identified a cohort of study patients older than 65 years who had a diagnosis of AD after January 2012 based on the presence of one or more AD related SNOMED-CT codes in our OMOP database (see Table 1). These patients were labeled as “Probable AD” (cases). We use the term “probable” since a definite diagnosis of AD would require pathologic confirmation of disease biomarkers at autopsy or the use of an amyloid positron emission tomography (PET) scan and/or cerebrospinal fluid (CSF) markers as part of the clinical diagnostic process. These information were not routinely available via the OMOP database. The inclusion-exclusion criteria used in the study are shown in Figure 1.

We randomly selected nine controls for each probable AD patient with age and gender matching from control candidate sample as shown in Figure 1, where control candidate samples were specified by excluding the patients without a diagnosis of AD or mild cognitive impairment (MCI) and those who had less than 5 years EHR data. This ratio was determined based on an estimated prevalence of 10% of AD in the general population of 65 years and older [1].

TABLE 1 Concept code description

SNOMED	Concept Name
378419	Alzheimer's disease
4218017	Primary degenerative dementia of the Alzheimer type, presenile onset
4220313	Primary degenerative dementia of the Alzheimer type, senile onset
4297400	Mild cognitive disorder
439795	Minimal cognitive impairment (MCI)

2.3 | Cohort characteristics

The probable AD cohort included 792 patients (487 female [61.5%]; mean [SD] age, 78.4 (5.4) years) who were matched at a ratio of 1:9 with 6795 control patients (4175 female [61.4%]; mean [SD] age, 78.5 [5.4] years). The characteristics of the study cohort are shown in Table 2.

2.4 | Data preparation

For probable AD patients, we extracted EHR data for over 8 years prior to their AD diagnosis. For the controls, we extracted the latest 8 years or more worth of EHR data. The extracted data included demographics, comorbidities, and medications, where comorbidities were derived based on the Chronic Condition Data Warehouse (CCW) algorithms (Medicare & Medicaid Services, Centers. 2017). Medications were mapped to drug classes derived from the ATC 3rd level. The CCW and ATC mappings to SNOMED are maintained at the latest version. Uniform 10-Bins Discretizer and one-hot encoding were chosen to encode age and gender variables. By determining whether a patient has received diagnosis or medication in those categories, the final study CCW and ATC categories included 32 different diagnoses and 171 medication classes, respectively. Therefore, data

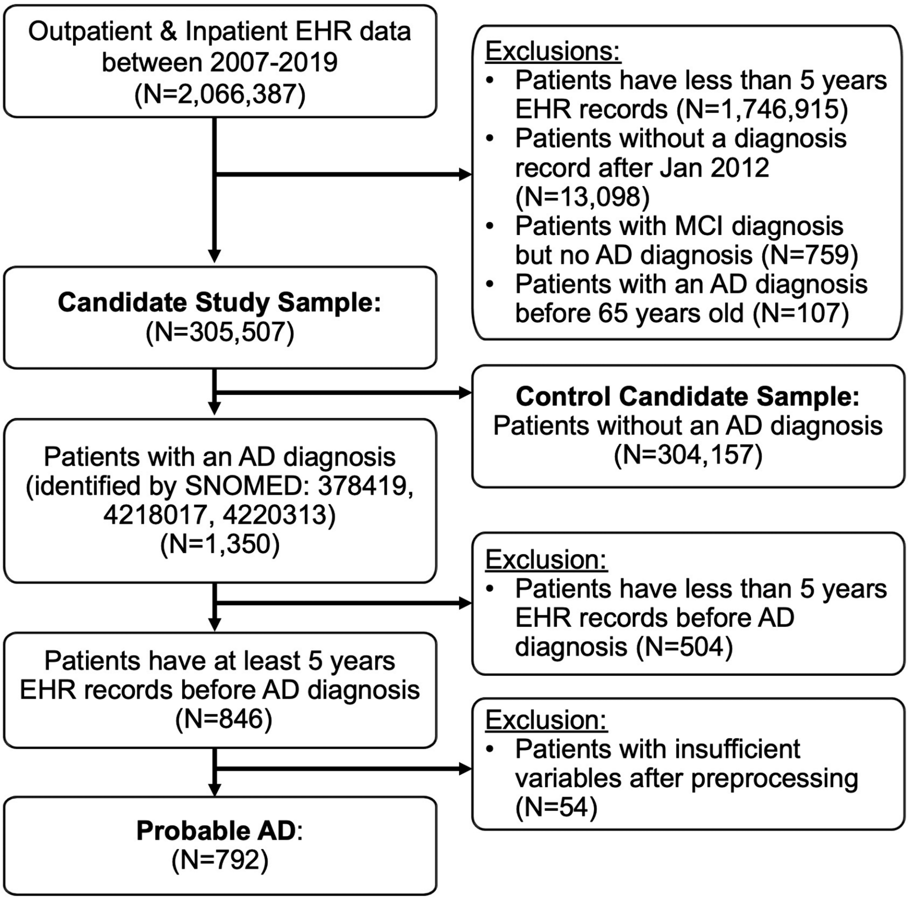

Outpatient & Inpatient EHR data between 2007-2019 (N=2,066,387)
Exclusions:
• Patients have less than 5 years EHR records (N=1,746,915)
• Patients without a diagnosis record after Jan 2012 (N=13,098)
• Patients with MCI diagnosis but no AD diagnosis (N=759)
• Patients with an AD diagnosis before 65 years old (N=107)
Candidate Study Sample: (N=305,507)
Control Candidate Sample: Patients without an AD diagnosis (N=304,157)
Patients with an AD diagnosis (identified by SNOMED: 378419, 4218017, 4220313) (N=1,350)
Exclusion:
• Patients have less than 5 years EHR records before AD diagnosis (N=504)
Patients have at least 5 years EHR records before AD diagnosis (N=846)
Exclusion:
• Patients with insufficient variables after preprocessing (N=54)
Probable AD: (N=792)

FIGURE 1 Cascade flow identifying Probable AD patients from EHR data

TABLE 2 Characteristics of the study cohort

	Study cohort	
Characteristic	**Probable AD (n = 792)**	**Control (n = 6795)**
Age, Mean (SD), y	78.4 (5.4)	78.5 (5.4)
Sex, No. (%)		
Female	487 (61.5)	4175 (61.4)
Race, No. (%)		
White	434 (54.8)	3805 (56.0)
Black	70 (8.8)	426 (6.3)
Asian	15 (1.9)	308 (4.5)
Others and Unknown	273 (34.5)	2256 (33.2)
# Visits (avg.)		
	70.6 (before AD onset)	59.1 (across entire EHRs)

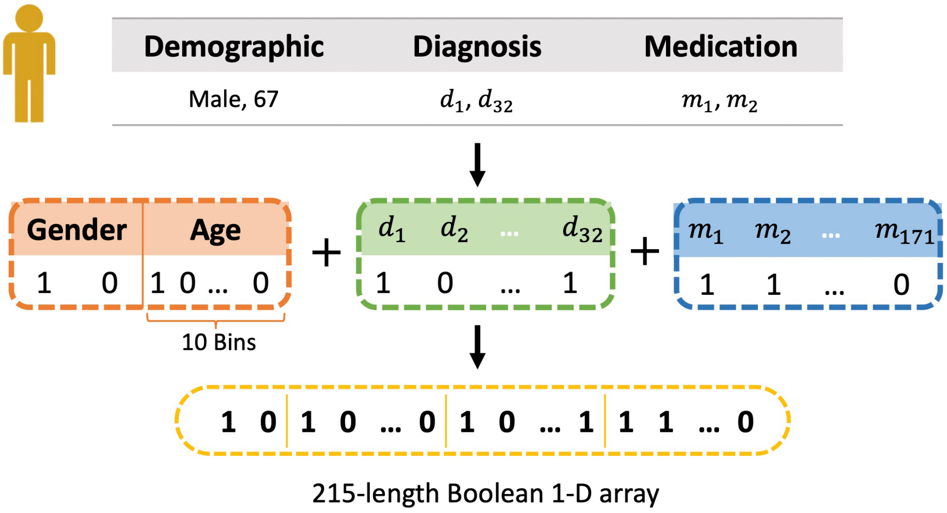


FIGURE 2 Example of feature construction. Diagnosis and medication are denoted by d and m, respectively

derived for each patient could be modeled as a 215-length Boolean 1-D array (Figure 2).

2.5 | Modeling

To derive the subphenotypes, we first assessed the effectiveness of variables extracted from the EHRs by training 0-year, 6-month, 1-year, and 2-year AD risk prediction models on 7587 patients using five times 5-fold cross-validation. Figure 3 shows the prediction window setting in these experiments, where we randomly chose a subset of controls to minimize class imbalance between the probable AD patients and the controls. We compared the prediction performance of four traditional ML algorithms including logistic regression (LR), LASSO, random forest (RF), and XGBoost. In addition, a decision-tree-based ensemble ML algorithm XGBoost was used to rank the variable importance in distinguishing probable AD patients and controls.

After validating the effectiveness of the extracted variables using the prediction task, we arbitrarily chose 5 years of EHR data for identifying probable AD patients before their first AD diagnosis to derive subphenotypes. After representing the patient data as binary codes as previously described, Ward's hierarchical agglomerative clustering method was performed to derive probable AD and related dementia subphenotypes by minimizing within-group dispersion at each binary fusion.[15] Dice distance matrices, calculated from these binary codes, were chosen as input dissimilarities.[16]

2.6 | Measurements

To assess the ML algorithm's prediction performance, we chose the area under the receiver operating characteristic curve (AUC) as our primary evaluation metric. In addition to AUC, we also calculated sensitivity, specificity, and F1 score for 0-year, 6-month, 1-year, 2-year, and 3-year prediction models. We assessed the variable importance using F-score, which reflects how much the variables contribute to the prediction task. After deriving the subphenotypes, we reported the frequency of each variable in each subphenotype. We used the Chi-square test to compare the subphenotypes. A significance level of 5% was used for deriving the main inferences.

FIGURE 3 Prediction window setting

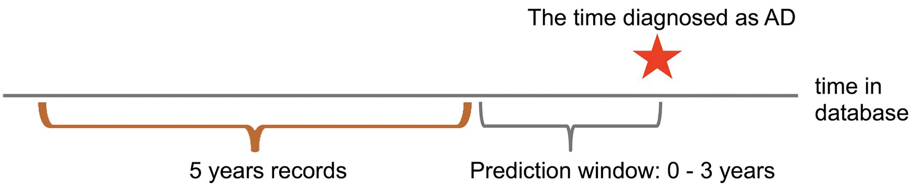


TABLE 3 Prediction results associated with 0-y, 6-mo, 1-y, 2-y, and 3-y prediction window

	Probable AD vs Control			
	AUC	Sensitivity	Specificity	F1 score
0-y				
LR	0.764 (0.02)	0.678 (0.04)	0.851 (0.03)	0.744 (0.03)
LASSO	0.747 (0.02)	0.661 (0.05)	0.833 (0.03)	0.726 (0.03)
RF	0.738 (0.02)	0.724 (0.03)	0.751 (0.05)	0.738 (0.03)
XGBoost	0.759 (0.03)	0.667 (0.05)	0.851 (0.03)	0.737 (0.04)
6-mo				
LR	0.743 (0.02)	0.636 (0.03)	0.850 (0.02)	0.714 (0.03)
LASSO	0.731 (0.01)	0.625 (0.03)	0.838 (0.03)	0.702 (0.02)
RF	0.745 (0.02)	0.727 (0.03)	0.762 (0.04)	0.744 (0.02)
XGBoost	0.751 (0.02)	0.638 (0.03)	0.863 (0.03)	0.721 (0.02)
1-y				
LR	0.745 (0.02)	0.642 (0.05)	0.848 (0.02)	0.718 (0.03)
LASSO	0.724 (0.02)	0.623 (0.04)	0.825 (0.05)	0.695 (0.02)
RF	0.736 (0.02)	0.715 (0.02)	0.756 (0.03)	0.734 (0.02)
XGBoost	0.752 (0.02)	0.636 (0.05)	0.868 (0.02)	0.721 (0.04)
2-y				
LR	0.745 (0.03)	0.643 (0.06)	0.848 (0.02)	0.717 (0.04)
LASSO	0.728 (0.04)	0.631 (0.05)	0.825 (0.03)	0.701 (0.04)
RF	0.725 (0.02)	0.697 (0.04)	0.754 (0.02)	0.720 (0.03)
XGBoost	0.749 (0.03)	0.631 (0.05)	0.867 (0.02)	0.716 (0.03)
3-y				
LR	0.735 (0.03)	0.636 (0.02)	0.833 (0.05)	0.708 (0.02)
LASSO	0.710 (0.02)	0.602 (0.03)	0.818 (0.05)	0.676 (0.02)
RF	0.716 (0.02)	0.685 (0.03)	0.747 (0.04)	0.709 (0.02)
XGBoost	0.733 (0.02)	0.615 (0.03)	0.850 (0.04)	0.699 (0.03)

Abbreviations: LR, logistic regression; RF, random forest.

3 | RESULTS

3.1 | Variable assessing results

Table 3 shows the results for various prediction models studied. The 0-year prediction model for probable AD achieved an AUC (SD) of 0.764 (0.02); the 6-month model, 0.751 (0.02); the 1-year model, 0.752 (0.02); the 2-year model, 0.749 (0.03); and the 3-year model, 0.735 (0.03). Patient records from the 0-year prediction window before AD diagnosis showed the best results, suggesting that the closer the AD diagnosis, the more pronounced the AD symptoms were.

Table 4 shows the top features and their importance (F score) associated with different times of various prediction windows. There are some differences between the top important features among the different prediction windows. Compared to the controls, depression became more distinguishable in probable AD patients closer to AD diagnosis. Hypertension as a common chronic disease played a more important role several years before the patients received an AD diagnosis. The top-ranked comorbidities verified by XGBoost included depression, stroke/transient ischemic attack, hypertension, anxiety, mobility impairments, and atrial fibrillation. The top-ranked medications included anti-dementia drugs, antipsychotics, antiepileptics, angiotensin II antagonists, adrenergics, and antidepressants.

TABLE 4 Top features importance (F score) associated with 0-y, 6-mo, 1-y, 2-y, and 3-y prediction windows

	Prediction window (y)				
Feature Name	**0**	**0.5**	**1**	**2**	**3**
Conditions					
Depression	33	33	31	20	16
Stroke/Transient Ischemic Attack	22	22	20	15	15
Hypertension	19	19	23	33	42
Anxiety	17	17	18	20	10
Mobility Impairments	17	17	15	12	6
Acquired Hypothyroidism	14	14	9	6	4
Epilepsy	12	12	12	15	9
Asthma	10	10	12	12	10
Heart Failure	8	8	7	15	16
Anemia	7	7	1	4	1
Migraine and Chronic Headache	6	6	7	6	6
Atrial Fibrillation	6	6	16	18	15
Hip/Pelvic Fracture	5	5	1	1	2
Bipolar Disorder	3	3	7	7	7
Sensory—Deafness and Hearing Impairment	3	3	8	5	8
Medications					
Anti-dementia Drugs	40	40	40	41	40
Antipsychotics	21	21	25	24	29
Antiepileptics	17	17	16	15	11
Angiotensin II Antagonists, Plain	16	16	15	15	18
Adrenergics, Inhalants	16	16	25	20	20
Dopaminergic Agents	16	16	13	13	14
Vasodilators used in Cardiac Diseases	13	13	8	6	7
Anti-inflammatory Agents	13	13	13	10	11
Other Mineral Supplements	11	11	10	7	6
Direct Acting Antivirals	11	11	16	13	10
Vitamin B12 and folic acid	10	10	14	10	5
Beta Blocking Agents	8	8	12	11	13
Antidepressants	6	6	9	15	19

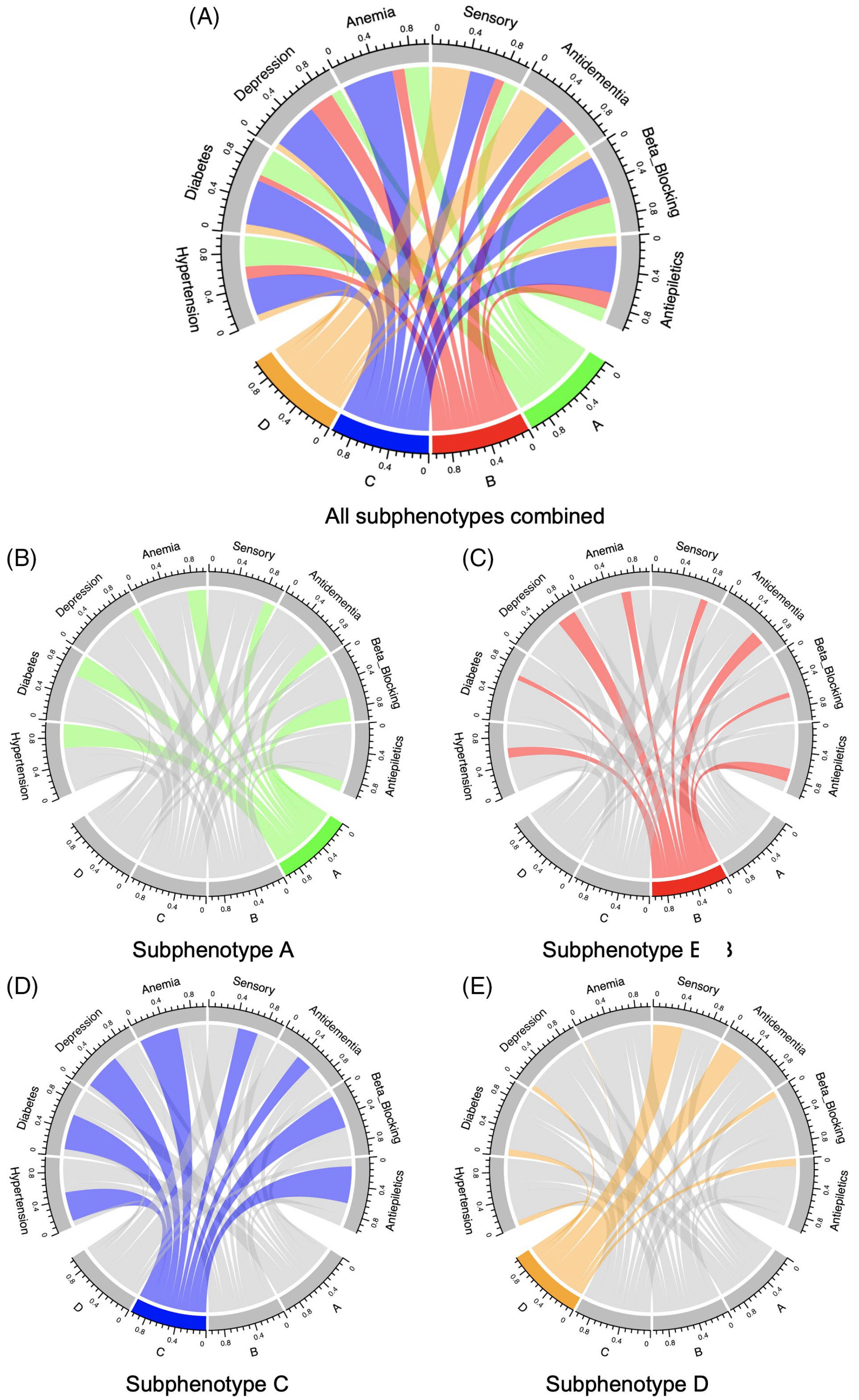
3.2 | Subphenotypes

After assessing the effectiveness of variables extracted from longitudinal EHRs, we then leveraged these variables to identify distinct subphenotypes from the 792 probable AD patients. Four subphenotypes were computationally derived. We plotted chord diagrams in Figure 4 to show the major differences across the subphenotypes. (Table 5 shows the characteristics of each subphenotype). To further compare subphenotypes, we conducted Chi-square tests to identify variables whose *P* value was larger than .05 between two subphenotypes, as shown in Table 6.

Among the four derived subphenotypes, Subphenotype A (n = 273; 28.2%) was mainly characterized by cardiovascular diseases. 70.7% of the patients in subphenotype A had hypertension, and 75.1% of the patients took lipid-modifying agents. Subphenotype B (n = 221; 27.9%) was mainly characterized by mental health illnesses where 31.7% of the patients had depression and 45.7% of the patients took antidepressants, 21.7% of the patients had anxiety and 24% of the patients took anxiolytics. Patients in Subphenotype C (n = 183; 23.1%) were overall older (mean [SD] age, 79.5 (5.4) years) and had the highest number of comorbidities, including diabetes, cardiovascular diseases, and mental health diseases. 86.3% of the patients had hypertension, 62.3% had diabetes, and 60.1% had depression. Accordingly, 74.9% of the patients took lipid-modifying agents, 79.2% of the patients took antidepressants, and 60.3% of the patients took anxiolytics. Subphenotype D (n = 115; 14.5%) included patients who took anti-dementia drugs and had sensory problems, such as deafness and hearing impairment. 75.7% of the patients took anti-dementia drugs and 45.2% of the patients had sensory problems.

We observed significant gender differences across the four subphenotypes (*P* ≤.001). While, in general, there were more females across the study population, 73.8% of patients were females in

FIGURE 4 Chord diagrams showing relatively frequent variables by subphenotypes



All subphenotypes combined

Subphenotype A

Subphenotype B

Subphenotype C

Subphenotype D

Subphenotype B. After adjusting significance in terms of age, no significant differences were observed for race. This could be potentially attributed to missing race information on >30% of the patients in our study cohort. Significant differences between the resulting four subphenotypes were found for diabetes (P≤.001), chronic kidney disease (P ≤ .05), lipid-modifying agents (P ≤.001), antidepressants (P≤.05), anxiolytics (P ≤.05), beta-blocking agents (P ≤.05), and antiepileptics (P ≤.05).

TABLE 5 Characteristics of the four subphenotypes

		Subphenotypes				χ^2	
Characteristic	**Total**	**A**	**B**	**C**	**D**	***P* value**	**Adjust age**
No. of patients (%)	792	273 (34.5)	221 (27.9)	183 (23.1)	115 (14.5)		
Age, Mean (SD), y	78.4 (5.4)	78.8 (5.2)	77.6 (5.5)	79.5 (5.4)	77.7 (5.4)		
Sex, No. (%)							
Female	487 (61.5)	145 (53.1)	163 (73.8)	116 (63.4)	63 (54.8)		
Male	305 (38.5)	128 (46.9)	58 (26.2)	67 (36.6)	52 (45.2)		
Race, No. (%)							
White	434 (54.8)	146 (53.5)	121 (54.8)	102 (55.7)	65 (56.5)		
Black	70 (8.8)	33 (12.1)	12 (5.4)	18 (9.8)	7 (6.1)		
Asian	15 (1.9)	9 (3.3)	3 (1.4)	3 (1.6)	0 (0)		
Others and Unknown	273 (34.5)	85 (31.1)	85 (38.5)	60 (32.8)	43 (37.4)		
Conditions, No. (%)							
Hypertension	433 (54.7)	193 (70.7)	64 (29)	158 (86.3)	18 (15.7)	≤.001	0.289
Diabetes	249 (31.4)	102 (37.4)	19 (8.6)	114 (62.3)	14 (12.2)	≤.001	≤0.001
Depression	224 (28.3)	35 (12.8)	70 (31.7)	110 (60.1)	9 (7.8)	≤.001	0.289
Hip/Pelvic Fracture	196 (24.7)	48 (17.6)	57 (25.8)	80 (43.7)	11 (9.6)	≤.001	0.368
Anemia	182 (23)	65 (23.8)	26 (11.8)	90 (49.2)	1 (0.9)	≤.001	0.985
Sensory—Deafness and Hearing Impairment	181 (22.9)	48 (17.6)	25 (11.3)	56 (30.6)	52 (45.2)	≤.001	0.335
Stroke/Transient Ischemic Attack	173 (21.8)	70 (25.6)	28 (12.7)	72 (39.3)	3 (2.6)	≤.001	0.326
Anxiety	154 (19.4)	32 (11.7)	48 (21.7)	66 (36.1)	8 (7)	≤.001	0.129
Chronic Kidney Disease	153 (19.3)	60 (22)	5 (2.3)	84 (45.9)	4 (3.5)	≤.001	0.003
Bipolar Disorder	135 (17)	21 (7.7)	38 (17.2)	73 (39.9)	3 (2.6)	≤.001	0.685
Atrial Fibrillation	128 (16.2)	48 (17.6)	11 (5)	66 (36.1)	3 (2.6)	≤.001	0.166
Acquired Hypothyroidism	127 (16)	47 (17.2)	16 (7.2)	57 (31.1)	7 (6.1)	≤.001	0.956
Migraine and Chronic Headache	120 (15.2)	43 (15.8)	24 (10.9)	53 (29)	0 (0)	≤.001	0.327
Heart Failure	106 (13.4)	44 (16.1)	1 (0.5)	61 (33.3)	0 (0)	≤.001	0.733
Medications, No. (%)							
Lipid Modifying Agents, Plain	484 (61.1)	205 (75.1)	68 (30.8)	137 (74.9)	74 (64.3)	≤.001	≤0.001
Anti-Dementia Drugs	389 (49.1)	117 (42.9)	95 (43)	90 (49.2)	87 (75.7)	≤.001	0.135
Antidepressants	366 (46.2)	79 (28.9)	101 (45.7)	145 (79.2)	41 (35.7)	≤.001	0.003
Beta Blocking Agents	310 (39.1)	142 (52)	23 (10.4)	130 (71)	15 (13)	≤.001	0.013
Antiepileptics	196 (24.7)	42 (15.4)	42 (19)	99 (54.1)	13 (11.3)	≤.001	0.027
Anxiolytics	191 (24.1)	41 (15)	53 (24)	92 (50.3)	5 (4.3)	≤.001	0.001
Antipsychotics	140 (17.7)	37 (13.6)	28 (12.7)	67 (36.6)	8 (7)	≤.001	0.899
Angiotensin II Antagonists, Plain	134 (16.9)	61 (22.3)	20 (9)	51 (27.9)	2 (1.7)	≤.001	0.899
Vitamin B12 and folic acid	172 (21.7)	63 (23.1)	36 (16.3)	66 (36.1)	7 (6.1)	≤.001	0.068
Other Mineral Supplements	171 (21.6)	56 (20.5)	8 (3.6)	105 (57.4)	2 (1.7)	≤.001	0.164
Adrenergics, Inhalants	107 (13.5)	39 (14.3)	7 (3.2)	58 (31.7)	3 (2.6)	≤.001	0.265

4 | DISCUSSION

The purpose of this study was to computationally derive probable AD and related dementia subphenotypes using routinely collected data from EHRs to potentially enhance our understanding of AD, and help develop better diagnosis and treatment pathways. In particular, by applying multiple "off-the-shelf" ML algorithms on EHR data, including patient demographics, comorbidities, and medication history, we assessed the effectiveness of variables extracted from longitudinal EHRs. Recent work[17] also used ML to predict AD with large-scale administrative claims data. While our study only used EHR data, we observed comparable performance at the 0-year prediction task, and significantly improved performance at 1-year, 2-year, and 3-year prediction tasks.

TABLE 6 Variables with *P* value > .05 (Chi-square test) between subphenotypes

	*P*value*					
Variables	**A vs B**	**A vs C**	**A vs D**	**B vs C**	**B vs D**	**C vs D**
Diabetes	-	-	-	-	0.296	-
Depression	-	-	0.157	-	-	-
Anxiety	-	-	0.159	-	-	-
Chronic Kidney Disease	-	-	-	-	0.513	-
Atrial Fibrillation	-	-	-	-	0.303	-
Acquired Hypothyroidism	-	-	-	-	0.691	-
Migraine and Chronic Headache	0.114	-	-	-	-	-
Lipid Modifying Agents, Plain	-	0.956	-	-	-	0.052
Anti-Dementia Drugs	0.977	0.184	-	0.214	-	-
Antidepressants	-	-	0.191	-	0.077	-
Beta Blocking Agents	-	-	-	-	0.469	-
Antiepileptics	0.287	-	0.293	-	0.070	-
Antipsychotics	0.773	-	0.064	-	0.108	-
Angiotensin II Antagonists, Plain	-	0.179	-	-	-	-
Vitamin B12 and folic acid	0.061	-	-	-	-	-
Other Mineral Supplements	-	-	-	-	0.336	-
Adrenergics, Inhalants	-	-	-	-	0.775	-

*P value ≤.05.

Four subphenotypes were computationally derived. Subphenotype A included more patients with cardiovascular diseases; Subphenotype B included more patients with mental health illnesses like depression and anxiety; Subphenotype C was relatively older (mean [SD] age, 79.5 [5.4] years) and had the most comorbidities including diabetes, cardiovascular diseases, and mental health disorders; and Subphenotype included patients who took anti-dementia drugs and had sensory problems, such as deafness and hearing impairment. Patients in Subphenotypes, B and D, were younger overall and had fewer comorbidities.

After adjusting for significance in terms of age, we observed statistically significant differences across the subphenotypes for comorbidities, including history of diabetes and chronic kidney disease, and treatment, including the use of lipid-modifying agents, antidepressants, anxiolytics, beta-blocking agents, and antiepileptics. Prior studies indicate that the prevalence of diabetes and AD is increasing in our aging population.[18,19] Clinical studies also demonstrated that patients with chronic kidney disease are more prone to cognitive impairment and AD.[20] There is some evidence from experimental studies that lowering cholesterol may slow the expression of AD.[21] Depression is very common among people with AD.[22,23] Studies have also reported that anxiety may be an early sign of increased risk for AD.[23] In addition, while prior studies have reported on environmental risk factors for AD, including the role of aluminum in drinking water and occupational exposure to solvents and pesticides,[24] studying the associations between built environment (eg, local air pollution, proximity to open space, access to public transportation) and risk of AD is beyond the scope of the current work.

This study has several limitations. First, the findings from this study have not been replicated using external EHR data sets. Given that this study was conducted using EHR data from patients at a single large, urban academic medical center in New York, the study population may not be representative of the general AD population. Conducting a replication study will be a critical next step. Moreover, to our knowledge, there is a dearth of existing work in data-driven identification of probable AD and related dementia subphenotypes using EHR data to compare our results. In addition, due to the insufficient number of patients who have longitudinal EHR data, we arbitrarily chose 5 years of EHR data for identifying probable AD patients before their first AD diagnosis to derive subphenotypes. This assumption might not completely reflect the characteristics of AD patients since patients might have early signs and symptoms of the disease prior to the 5-year time window. There exists a substantial level of overdiagnosis and underdiagnosis of AD (Sodenaga[25]). Moreover, several presenting symptoms of vascular dementia are similar to symptoms of AD—and become more similar as dementia progresses. It is sometimes difficult to tell whether a person has AD or vascular dementia[26] and it is not unusual to have a mixed form of dementia, meaning the person has both vascular dementia and AD together.[27] A definite diagnosis of AD would require pathologic confirmation of disease biomarkers at autopsy, or the use of an amyloid imaging scan and/or CSF markers as part of the clinical diagnostic process. Due to the high cost of amyloid imaging and lack of reimbursement by insurance companies, and the cost and invasive nature of CSF samples, most healthcare providers do not incorporate these measures in their routine clinical practice. This is an important limitation of an EHR-based approach. This study also did not analyze unstructured clinical text from EHRs, including clinician encounter notes, which may contain additional information about an individual's comorbidities and treatment. In the future, we will apply our extensive work in natural language processing to analyze such data.[28,29]

5 | CONCLUSION

Using routinely collected longitudinal EHR data and ML algorithms, we computationally derived probable AD and related dementia subphenotypes that can potentially guide improved diagnosis and treatment of AD patients. The derived subphenotypes had statistically significant differences with respect to patient demographics, comorbidities, and treatment, suggesting that despite converging to a final common clinicopathological endpoint, AD is a heterogeneous disorder with multiple phenotypes.

ACKNOWLEDGEMENT

This study was supported, in part, by funding from NIH R01GM105688, R01MH119177, and R01MH105384.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jie Xu https://orcid.org/0000-0001-5291-5198
Pascal Brandt https://orcid.org/0000-0001-5116-0555

REFERENCES

1. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimer's Dementia*. 2019;15:321-387. https://doi.org/10.1016/j.jalz.2019.01.010.
2. Lazarov O, Tesco G. *Genes, Environment and Alzheimer's Disease*. United Kingdom: Academic Press; 2016.
3. Zhang, Rui, Gyorgy Simon, and Fang Yu. 2017. "Advancing Alzheimer's research: a review of big data promises." *Int J Med Inform*. 106:48–56. https://doi.org/10.1016/j.ijmedinf.2017.07.002
4. Pujades-Rodriguez M, Assi V, Gonzalez-Izquierdo A, et al. The diagnosis, burden and prognosis of dementia: a record-linkage cohort study in England. *PLoS One*. 2018;13(6):e0199026.
5. Manson A, Ciro C, Williams KN, Maliski SL. Identity and perceptions of quality of life in Alzheimer's disease. *Appl Nurs Res*. 2019;52: 151225.
6. Wilkins JM, Forester BP. Informed consent, therapeutic misconception, and clinical trials for Alzheimer's disease. *Int J Geriatr Psychiatry*. 2020;35(5):430–435. https://doi.org/10.1002/gps.5262.
7. Light E, Lebowitz B. *Alzheimer's Disease Treatment and Family Stress: Directions for Research*. United Kingdom: Taylor & Francis; 1990.
8. Geifman, Nophar, Richard E. Kennedy, Lon S. Schneider, Iain Buchan, and Roberta Diaz Brinton. 2018. "Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions." *Alzheimer's research & therapy*. 10(1):1–7. https://doi.org/10.1186/s13195-017-0332-0.
9. Oxtoby NP, Young AL, Cash DM, et al. Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain*. 2018; 141(5):1529-1544.
10. Chatfield, Mark D., Carol E. Brayne, and Fiona E. Matthews. 2005. "A systematic literature review of attrition between waves in longitudinal studies in the elderly shows a consistent pattern of dropout between differing studies." *J Clin Epidemiol*. 58(1):13–19. https://doi.org/10.1016/j.jclinepi.2004.05.006.
11. Matthews FE, Chatfield M, Freeman C, McCracken C, Brayne C, MRC CFAS. Attrition and bias in the MRC cognitive function and ageing study: an epidemiological investigation. *BMC Public Health*. 2004;4 (April):12.
12. Maserejian, Nancy Nairi, Jin Wang, Henry Krzywy, Maneesh Juneja, and Judith Jaeger. 2018. "Cognitive and other neuropsychological assessments documented in electronic health records prior to or at Alzheimer's disease diagnosis." *Alzheimer's Dement*. 14(7):423. https://doi.org/10.1016/j.jalz.2018.06.346.
13. Park J-H, Cho H-E, Cha JM, et al. Machine learning prediction of future incidence of Alzheimer's disease using population-wide electronic health records. *Alzheimer's Dement*. 2019;15(7):342–343. https://doi.org/10.1016/j.jalz.2019.06.825.
14. Tjandra, Donna, Raymond Migrino, Bruno Giordani, and Jenna Wiens. 2019. "An EHR-based cohort Discovery tool for identifying probable AD." *Alzheimer's Dement*. 15(7):1530–1531. https://doi.org/10.1016/j.jalz.2019.08.101.
15. Murtagh, Fionn, and Pierre Legendre. 2014. "Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?" *J Classif*. 31(3):274–295. https://doi.org/10.1007/s00357-014-9161-z.
16. Jimenez S, Gonzalez FA, Gelbukh A. Mathematical properties of soft cardinality: enhancing Jaccard, Dice and cosine similarity measures with element-wise distance. *Inf Sci*. 2016;367: 373-389.
17. 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(1):1-7.
18. Brookmeyer R, Abdalla N, Kawas CH, Corrada MM. Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. *Alzheimers Dement*. 2018;14(2):121-129.
19. Rojas-Carranza CA, Bustos-Cruz RH, Pino-Pinzon CJ, Ariza-Marquez YV, Gomez-Bello RM, Canadas-Garre M. Diabetes-related neurological implications and pharmacogenomics. *Curr Pharm Des*. 2018;24(15):1695-1710.
20. Shi, Yan, Zhangsuo Liu, Yong Shen, and Hanyu Zhu. 2018. "A novel perspective linkage between kidney function and Alzheimer's disease." *Front Cell Neurosci*. 12:384–392. https://doi.org/10.3389/fncel.2018.00384.
21. Ancelin ML, Carrière I, Barberger-Gateau P, et al. Lipid lowering agents, cognitive decline, and dementia: the three-city study. *J Alzheimers Dis*. 2012;30(3):629-637.
22. Henriques-Calado J, Duarte-Silva ME. Personality disorders characterized by anxiety predict Alzheimer's disease in women: a case-control studies. *J Gen Psychol*. 2019;147(4):414-431.
23. Novais F, Starkstein S. Phenomenology of depression in Alzheimer's disease. *J Alzheimers Dis*. 2015;47(4):845-855.
24. Campdelacreu, J. 2014. "Parkinson's disease and Alzheimer disease: environmental risk factors." *Neurología*. 29(9):541–549. https://doi.org/10.1016/j.nrleng.2012.04.022.
25. Rui, Z, Gyorgy, S. Advancing Alzheimer's research: a review of big data promises. *International journal of medical informatics*. 2017;106: 48–56.
26. Ballard C, O'brien J, Morris CM, et al. The progression of cognitive impairment in dementia with Lewy bodies, vascular dementia and Alzheimer's disease. *Int J Geriatr Psychiatry*. 2001;16(5):499-503.

27. Korczyn AD. Mixed dementia—the most common cause of dementia. *Ann N Y Acad Sci*. 2002;977(1):129-134.
28. Sharma H, Mao C, Zhang Y, et al. Developing a portable natural language processing based phenotyping system. *BMC Med Inform Decis Mak*. 2019;19(Suppl 3):78.
29. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak*. 2019;19(Suppl 3):71.

How to cite this article: Xu J, Wang F, Xu Z, et al. Data-driven discovery of probable Alzheimer's disease and related dementia subphenotypes using electronic health records. *Learn Health Sys*. 2020;4:e10246. https://doi.org/10.1002/lrh2.10246