Received: 12 March 2024 | Revised: 13 May 2024 | Accepted: 18 May 2024

DOI: 10.1002/dad2.12613

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Develop and validate a computable phenotype for the identification of Alzheimer's disease patients using electronic health record data

Xing He[1] | Ruoqi Wei[1] | Yu Huang[1] | Zhaoyi Chen[2] | Tianchen Lyu[1] | Sarah Bost[1] | Jiayi Tong[3] | Lu Li[3] | Yujia Zhou[4] | Zhao Li[5] | Jingchuan Guo[6] | Huilin Tang[6] | Fei Wang[7] | Steven DeKosky[8] | Hua Xu[4] | Yong Chen[3] | Rui Zhang[9] | Jie Xu[1] | Yi Guo[1] | Yonghui Wu[1] | Jiang Bian[1]

[1]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA

[2]Center for Biomedical Informatics & Information Technology, National Cancer Institute, Rockville, Maryland, USA

[3]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[4]Biomedical Informatics and Data Science, School of Medicine, Yale, New Haven, Connecticut, USA

[5]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA

[6]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, Florida, USA

[7]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA

[8]Department of Neurology, College of Medicine, University of Florida, Gainesville, Florida, USA

[9]Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, Minnesota, USA

Correspondence
Jiang Bian, PhD, Health Outcomes & Biomedical Informatics, University of Florida, 1889 Museum Rd, Suite 7000, Gainesville, FL 32611, USA.
Email: bianjiang@ufl.edu

Funding information
National Institute on Aging, Grant/Award Numbers: R01AG080991, R01AG080624, R01AG076234, R01AG078154, R56AG069880; Ed and Ethel Moore Alzheimer's Disease Research Program from the Florida Department of Health, Grant/Award Number: FL DOH #23A09; Patient-Centered Outcomes Research Institute® (PCORI®) Award, Grant/Award Number: ME-2018C3-14754

Abstract

INTRODUCTION: Alzheimer's disease (AD) is often misclassified in electronic health records (EHRs) when relying solely on diagnosis codes. This study aimed to develop a more accurate, computable phenotype (CP) for identifying AD patients using structured and unstructured EHR data.

METHODS: We used EHRs from the University of Florida Health (UFHealth) system and created rule-based CPs iteratively through manual chart reviews. The CPs were then validated using data from the University of Texas Health Science Center at Houston (UTHealth) and the University of Minnesota (UMN).

RESULTS: Our best-performing CP was "*patient has at least 2 AD diagnoses and AD-related keywords in AD encounters*," with an F1-score of 0.817 at UF, 0.961 at UTHealth, and 0.623 at UMN, respectively.

DISCUSSION: We developed and validated rule-based CPs for AD identification with good performance, which will be crucial for studies that aim to use real-world data like EHRs.

Xing He and Ruoqi Wei contributed equally and were co-first authors.

KEYWORDS
Alzheimer's disease, computable phenotype, electronic health record

Highlights

- Developed a computable phenotype (CP) to identify Alzheimer's disease (AD) patients using EHR data.
- Utilized both structured and unstructured EHR data to enhance CP accuracy.
- Achieved a high F1-score of 0.817 at UFHealth, and 0.961 and 0.623 at UTHealth and UMN.
- Validated the CP across different demographics, ensuring robustness and fairness.

1 | BACKGROUND

Alzheimer's disease (AD) and AD-related dementias (AD/ADRD) represent complex neurodegenerative diseases affecting approximately 6.7 million Americans over 65 and over 40 million people worldwide.[1–3] Significant efforts have been made to better understand AD/ADRD, seek effective treatments and prevention strategies, and address the needs of AD/ADRD patients. The United States (US) National Alzheimer's Project Act (NAPA) has recommended a $2 billion annual budget and calls for an aggressive and coordinated national plan to accelerate AD/ADRD research and improve patient care.[4]

The widespread adoption of electronic health record (EHR) systems has made large-scale, longitudinal clinical datasets available for research. As an important real-world data (RWD) source,[5] EHRs have become increasingly important for generating real-world evidence (RWE)[6] in AD/ADRD research reflecting the patient population treated in real-world clinical settings. For example, Miller et al. examined the prevalence of AD/ADRD in the state of Florida and characterized the demographic characteristics of the AD/ADRD population using EHR data from the OneFlorida (now OneFlorida+) Clinical Research Consortium.[7] Many studies have also developed AD/ADRD prediction models using diagnosis, medication history, and biomarker data from RWD like EHRs and administrative claims.[8–11] However, identifying target populations manually from large collections of RWD sources (e.g., OneFlorida+ and others) for AD/ADRD research is notably difficult. Algorithms that can accurately and automatically identify patients with required phenotype characteristics (e.g., AD differentiating from other ADRD) are essential in constructing research-grade cohorts to support AD/ADRD research.

Previously, AD/ADRD cohorts were often identified solely by diagnosis codes (e.g., International Classification of Diseases [ICD]), leading to significant misclassification errors. High variations in classification accuracies have been reported in validation studies when using diagnosis codes to define dementia, including AD/ADRD.[12] In a study that used two Swedish national RWD registers and six population-based studies, Rizzuto et al.[13] found that relying solely on diagnosis codes yields a positive predictive value (PPV) of 0.82. In two other studies using Danish nationwide hospital registers, diagnosis codes accurately identified AD in only 60-80% of cases, with a PPV ranging from 0.78 to 0.81.[14,15] In a US-based study, Taylor et al. found that the AD diagnosis codes in Medicare claims data only have a sensitivity of 0.64 and a specificity of 0.96 for identifying AD.[16] In addition to diagnosis codes, EHR data elements like AD-related medications have also been used to identify AD patients. Tjandra et al. developed and validated an AD cohort discovery tool using a rule set that included encounters, diagnosis codes, medications, and procedure codes (e.g., for psychological/cognitive testing), and achieved moderate performance with an F1-score of 0.73, a PPV of 0.77, and a sensitivity of 0.70 in a Michigan Alzheimer's Disease Research Center (ADRC) cohort.[17]

Identifying patients with a particular condition, for example, AD, within the context of EHRs, is accomplished through a computable phenotype (CP) or simply phenotype (traditionally often called cohort identification or case-finding algorithms), which is defined as "*clinical conditions, characteristics, or sets of clinical features that can be determined solely from EHRs and ancillary data sources and does not require chart review or interpretation by a clinician*."[18] CPs have gained popularity for their high specificity and sensitivity in EHR-based cohort identification, demonstrating success in various domains such as the identification of HIV prevalent cases, transgender and gender nonconforming individuals, and resistant hypertension, among others.[19–21] Traditionally, EHR-based CPs only considered structured information (e.g., diagnoses, medications), while EHR contains rich unstructured clinical narratives (e.g., progression notes, discharge summaries).[22] In fact, over 80% of patient information in EHRs is documented in free-text clinical narratives,[22] which contain more detailed patient information, including important variables such as cognitive assessments that can facilitate the identification of AD patients. Prior studies across different disease domains have shown that leveraging both structured EHR data and unstructured narratives in CPs can significantly enhance their performance.[23,24]

In this study, aimed at addressing the challenges of accurately identifying individuals with AD, we developed and validated a CP that utilizes both structured and unstructured data from the University of Florida Health (UFHealth) EHR. We assessed the prevalence of AD in the

UFHealth patient cohort, detailing the characteristics of these patients. Additionally, to ensure the CP's applicability and generalizability, we conducted validation studies in other sites' EHRs, including the University of Texas Health Science Center at Houston (UTHealth) and the University of Minnesota (UMN). Resources such as the diagnosis codes and keywords used in this study are available on GitHub[25] and also in the supplemental material "Supplement_AD_CP_Final.xlsx."

2 | METHODS

2.1 | Data sources

We retrieved individual patient-level data from the UFHealth Integrated Data Repository (IDR) after obtaining approval from the UF Institutional Review Board (IRB). The UFHealth IDR serves as an enterprise data warehouse consisting of data from across UFHealth's clinical and administrative information systems (e.g., Epic EHR system; Janesville, WI), covering a population of over 2 million patients.[26] We then used data from the UTHealth and the UMN Academic Health Center Information Exchange (AHC-IE) clinical data repository (CDR) for external validation of the CPs developed using UFHealth data. The UTHealth Physicians CDW encompasses all UTHealth Physicians outpatient EHR data, serving approximately 1.8 million patients. The UMN AHC-IE CDR comprises data from over 4.5 million patients who received care at eight hospitals and more than 40 clinics.

2.2 | Overall study design

We developed the CP for identifying AD patients using structured and unstructured EHR data. As shown in Figure 1, we adopted a two-step process to develop the AD CP: (1) we applied a baseline CP (i.e., "*patients with at least one AD-related diagnosis codes*") to identify a potential AD cohort via searching EHRs using International Classification of Diseases-Ninth/Tenth Revision-Clinical Modification (ICD-9/10-CM) codes as shown in Table 1. For all the patients within the cohort, we collected their EHR data, including structured data (e.g., demographics, diagnoses, procedures, medications, laboratory results, procedures) and unstructured clinical notes (e.g., progress notes, discharge summaries, pathology reports, identified via regular expressions); and (2) we iteratively derived the CP rules through manual chart reviews on selected samples from the potential AD cohort.

2.3 | Derive CP rules based on manual chart reviews on selected samples from the potential AD cohort

Based on insights from previous studies on case-finding algorithms for AD[17] and dementia,[11,12,27,28] as well as consultations with clinicians who specialize in AD patient care, we proposed seven initial base rules: (1) age equal to or greater than 65 years, (2) having at least two of the AD diagnosis codes, (3) having at least five of the AD diagnosis codes, (4) having at least one of the other ADRD diagnoses, (5) having at least one of the cognitive decline diagnoses, (6) use of at least one of the relevant medications, and (7) having at least one of the relevant keywords in notes from an AD encounter. The value sets, codes, and keywords used in the seven base rules are listed in Table 1.

We generated 69 distinct combinations by combining these seven base rules, as shown in Table 2. Subsequently, we randomly sampled 10% of the patients who met the rule combination for each of these 69 combinations. If the number of patients for the 10% sample was larger than 20, we employed random selection to pick 20 patients for this combination. Conversely, if the total patient count for a combination was less than 3, we manually reviewed all patients within that combination. In total, we selected 363 patients and split them into a training set and a testing set in an 8:2 ratio (i.e., 282 for training samples and 81 for testing samples). Only the training set was used to develop the CP rules.

We first developed an annotation guideline for the manual chart reviews to ensure consistent criteria were applied across all reviewers. Three annotators (T.L., P.Y., and S.B.) conducted the chart review iteratively. In each round, they independently reviewed the same 10 samples from the training set following the annotation guideline. After each round, if any disagreements arose among the three reviewers, the entire study team engaged in discussions to resolve these conflicts and reach a consensus, and the annotation guideline was iteratively revised accordingly to these discussions. After five rounds of chart

RESEARCH IN CONTEXT

1. **Systematic review**: The authors reviewed the literature using PubMed and identified that previous Alzheimer's disease (AD) related computable phenotypes (CPs) were often limited by misclassifications due to reliance on AD diagnosis codes alone, highlighted the need for incorporating diverse data elements (e.g., medications, procedures, and keywords) to enhance specificity and sensitivity of the CPs.
2. **Interpretation**: Our CP integrates structured and unstructured data, improving upon previous CPs by enhancing the accuracy of AD patient identification. The best-performing CP, which includes AD diagnoses and relevant keywords from clinical encounters, demonstrated high sensitivity and F1-scores.
3. **Future directions**: Future research should focus on refining these CPs to address variability in performance across different EHR systems. Further studies could also explore advanced natural language processing tools to better interpret the context of keywords within clinical narratives, improving the robustness and applicability of CPs for identifying AD patients in diverse healthcare settings.

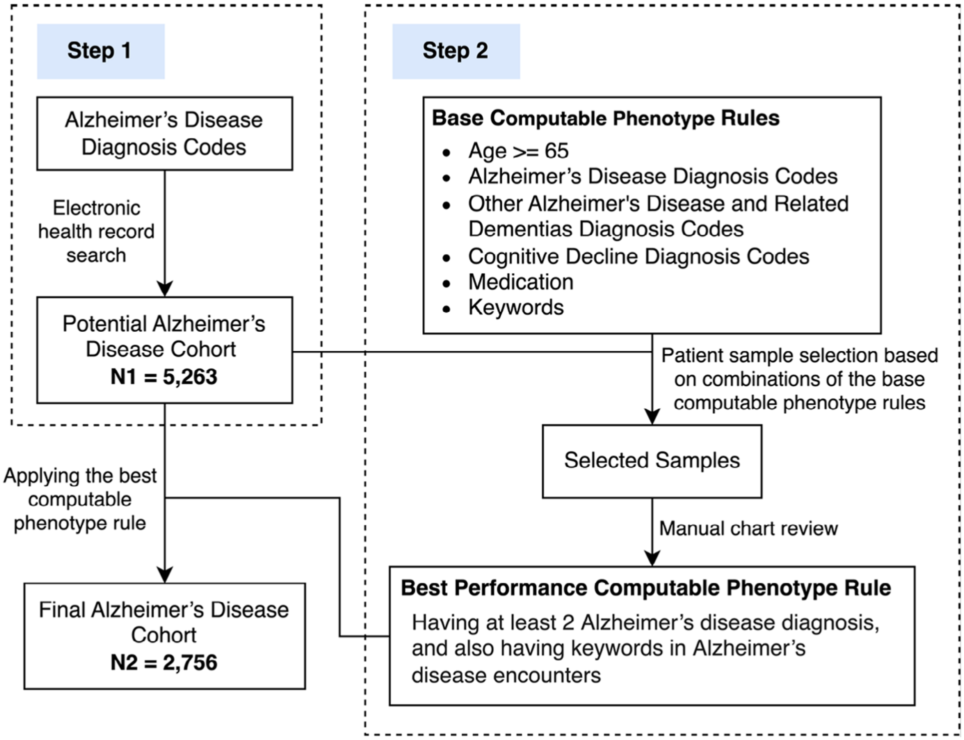
Step 1
Alzheimer's Disease Diagnosis Codes
Electronic health record search
Potential Alzheimer's Disease Cohort
N1 = 5,263
Applying the best computable phenotype rule
Final Alzheimer's Disease Cohort
N2 = 2,756
Step 2
Base Computable Phenotype Rules
• Age >= 65
• Alzheimer's Disease Diagnosis Codes
• Other Alzheimer's Disease and Related Dementias Diagnosis Codes
• Cognitive Decline Diagnosis Codes
• Medication
• Keywords
Patient sample selection based on combinations of the base computable phenotype rules
Selected Samples
Manual chart review
Best Performance Computable Phenotype Rule
Having at least 2 Alzheimer's disease diagnosis, and also having keywords in Alzheimer's disease encounters

FIGURE 1 Flow chart of the Alzheimer's' disease computable phenotype development process.

TABLE 1 Value sets, codes, and keywords used in the seven base rules.

Parameter	Value sets/codes/keywords
AD diagnosis codes	
ICD-9-CM	331.0 – "Alzheimer's disease"
ICD-10-CM	G30 – "Alzheimer's disease" G30.0 – "Alzheimer's disease with early onset" G30.1 – "Alzheimer's disease with late onset" G30.8 – "Other Alzheimer's disease" G30.9 – "Alzheimer's disease, unspecified"
Other ADRD diagnosis codes (selected examples)	
ICD-9-CM	*e.g., 290.0 – "Senile dementia, uncomplicated", 331.11 – "Pick's disease", 437.2 – "Hypertensive encephalopathy", etc.*
ICD-10-CM	*e.g., F01.50 – "Vascular dementia without behavioral disturbance", G91.0 – "Communicating hydrocephalus", G94 – "Other disorders of brain in diseases classified elsewhere", etc.*
Cognitive decline diagnosis codes (selected examples)	
ICD-9-CM	*e.g., 78093 – "Memory loss", 79952 – "Cognitive communication deficit", 331.83 – "Mild cognitive impairment, so stated", etc.*
ICD-10-CM	*e.g., G31.84 – "Mild cognitive impairment, so stated", G31.89 – "Other specified degenerative disease of nervous system", R41.81 – "Age-related cognitive decline", etc.*
Medication (selected examples)	
Names/RxNorm	*e.g., "Aricept", "Namzaric Oral Product", "24 HR galantamine hydrobromide 16 MG Extended Release Oral Capsule", "Namenda", etc.*
Keywords (selected examples)	
	e.g., "dementia", "Alzheimer", "memory loss", "cognitive deficits", "cognitive impairment", "cognitive disorders", "cognitive decline", "amnesia", etc.

TABLE 2 Summary of top 8 combinations of base rules, the number of patients identified by each rule combination, and the number of actual patients confirmed by manual chart reviews.

Age (years)	AD diagnosis codes		Other ADRD diagnosis codes	Cognitive decline diagnosis codes	Medication	Keyword	Total # of patients[a]	# of patients selected	# of AD patients[b]
≥65	≥2	≥5	≥1	≥1	≥1	≥1			
+	–	–	+	–	+	+	773	20	2
+	–	–	+	–	–	+	616	20	1
+	+	+	+	+	+	+	565	20	19
+	+	–	+	–	+	+	554	20	11
+	+	–	+	+	+	+	552	20	17
+	–	–	+	+	+	+	430	20	0
+	+	+	+	–	+	+	283	20	17
+	+	–	+	–	–	+	235	20	12

Note: "+" indicates that the patient must meet this rule.

[a] Overall, there were 69 different rule combinations that identified patients. Only the top 8 rule combinations are displayed here. All the 69 rule combinations are reported in Supplement Table 1.

[b] Number of AD patients identified by chart review from the selected patient sample.

reviews (i.e., after assessing 50 samples), the inter-rater agreements between any two of the three annotators achieved a Cohen's kappa of 1, indicating perfect agreement. Subsequently, the three annotators began annotating the remaining training samples independently, with the explicit instruction to be cautious when a case was deemed ambiguous. A fourth reviewer (J.B.) helped resolve the discrepancies when inconsistent annotations were encountered. We assessed the performance of each rule combination on both the training and testing sets, which were derived from the chart-reviewed cohort, using multiple metrics, including sensitivity, PPV, and F1-score, with a satisfactory cutoff point of 0.8 (80%) for all these metrics. To well explain the metrics, we introduced four basic concepts for calculating sensitivity, PPV, and F1-score: P is the number of positive instances, N is the number of negative samples, PP is the number of samples that are predicted as positive, PN is the count of instance that predicted as negative, TP (true positive) denotes the number of samples predicted as positive correctly, FP (false positive) is the count of instances incorrectly predicted as positive, FN (false negative) indicates the number of samples that are misclassified into negative, and TN (true negatives) is the count of instances correctly predicted as negative.

Sensitivity measures the proportion of actual positive cases that are correctly identified as such (true positives) by the CP rule:

$$\text{Sensitivity} = \frac{TP}{P}$$

PPV, or precision, is the proportion of predicted positive cases that are truly positive. It reflects the probability that a predicted positive case accurately indicates the presence of the condition:

$$\text{PPV} = \frac{TP}{P}$$

F1-score is a balanced measure, encapsulating both sensitivity and PPV. It is the harmonic mean of the two, offering a single, consolidated score that mitigates the impact of extreme values in either sensitivity or PPV, thus providing a more balanced assessment of our CP rules:

$$\text{F1} - \text{score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

The rule with the highest F1 score was selected as the best-performing CP, considering two distinct scenarios: (1) only considering structured data, and (2) considering both structured and unstructured data.

2.4 | External validation

To further validate the best-performing CPs and evaluate their generalizability, we distributed the annotation guidelines, the best-performing CP rules, and the corresponding codesets to UTHealth and UMN. Both institutions independently performed manual chart reviews on a randomly selected sample of 50 patients from their EHR data, adhering to the same annotation guidelines previously utilized at UFHealth. The performance of the CP algorithm was subsequently evaluated based on these annotated patient samples.

2.5 | Statistical analysis

All three sites independently applied the baseline CP and the overall best-performing CP to their EHR data to identify a potential AD cohort and a definitive AD cohort. Subsequently, we employed chi-squared tests to statistically verify the significance of observed differences across demographic categories—such as age, gender, and race/ethnicity—between the definitive and potential AD cohorts at each site. Moreover, we performed proportional Z-tests to compare the prevalence rates of certain chronic conditions within the two identified AD groups at each site.

TABLE 3 Performance of the baseline CP rule and the best-performing CP rules in terms of F1-score on the training data and the testing data.

Parameter	Rule[a]	Sensitivity	PPV	F1-score
Training data				
Baseline	With at least 1 AD diagnoses	1.000	0.415	0.586
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.949	0.673	0.787
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.932	0.747	0.829
Testing data				
Baseline	With at least 1 AD diagnoses	1.000	0.370	0.541
Structured data only (best F1-score)	With at least 2 AD diagnoses	1.000	0.652	0.789
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.967	0.707	0.817

2.6 | Performance evaluation on demographic subgroups

To evaluate the performance variability of the developed CPs across different demographic subgroups, we assessed the performance of the baseline CP and the best-performing CPs within the UF-site chart-reviewed cohort (i.e., the combination of the training and testing data) on subgroups of gender (i.e., "Female," "Male"), age (i.e., "<= 64" and ">64"), and race/ethnicity (i.e., "Hispanic," "Non-Hispanic White," and "Non-Hispanic Black"). The performance metrics reported include sensitivity, PPV, and F1-score.

3 | RESULTS

3.1 | Development of the CP for the identification of AD patients

Using the AD diagnosis codes (i.e., ICD codes in Table 1), we identified a potential AD cohort of 5,263 patients from the UFHealth IDR. Our final CP identified 2,756 AD patients among this cohort, as shown in Figure 1. A final set of CP algorithms was selected based on the best F1-score of the various base CP rule combinations listed in Table 3, using manual chart review results as the gold standard. The best-performing CP with structured data only was "*the patient has at least 2 AD diagnoses*," having an F1-score of 0.787 on the training set. When considering both structured and unstructured data, the best-performing CP was "*the patient has at least 2 AD diagnoses and has keywords in AD encounters*," with an F1-score of 0.829, outperforming the one considering structured data only. The performance of these CP algorithms was further assessed using the independent testing set (i.e., the test sample with 81 patients). When applying the final CP algorithms to the testing set, the CP with structured data received an F1-score of 0.789, while the CP using both structured and unstructured data achieved a better F1-score of 0.817. Both CP rules exhibited significant improvements in the F1-score compared to the baseline CP rule. Table 3 shows the different performance metrics (i.e., sensitivity, PPV, and F1-score) of these CP algorithms under different settings.

3.2 | External validation

There were 6,821 patients and 10,387 patients with at least one AD diagnosis code identified at UTHealth and UMN sites, respectively. For the validation process, a random sample of 50 patients from each site was selected for manual chart reviews. Table 4 shows the performance of the final CP rules on the two validation sites. Our CPs showed different performances across the two sites. For the structured data-only CP, the F1-score was 0.871 and 0.667, respectively, for the two validation sites. On the other hand, the CP using both structured data and unstructured data had an F1-score of 0.961 and 0.623 for the respective sites.

3.3 | Definitive AD cohort characteristics

We applied the best-performing CP using both structured and unstructured data (i.e., "*the patient has at least 2 AD diagnoses and has keywords in AD encounters*") to identify the definitive AD cohorts in each site. Table 5 describes the patient characteristics for both the definitive AD cohort and the potential AD cohort across the three sites. In our site-specific comparisons, significant demographic differences were identified between the definitive and potential AD cohorts. Notably, across each site, a comparison of age group and race/ethnicity distribution between the two cohorts revealed a statistically significant difference ($p < 0.05$). Conversely, when examining sex distribution, the analysis across all sites indicated no statistical significance ($p > 0.05$). Further examination focused on the prevalence of several chronic conditions within each cohort. Across all sites, the definitive AD cohort showed a significantly higher prevalence of depression and cancer compared to the potential AD cohort ($p < 0.05$).

3.4 | Performance on different demographic subgroups

Table 6 shows the performance of the baseline CP rule and the best-performing CP rules on different demographic subgroups in the UFHealth chart-reviewed cohort. Our overall best-performing

TABLE 4 Performance of best-performing CP rules on external validation sites.

Parameter	Rule	Sensitivity	PPV	F1-score
UTHealth validation data				
Structured data only	With at least 2 AD diagnoses	0.974	0.787	0.871
Structured and unstructured data	With at least 2 AD diagnoses, and with keywords in AD encounters	0.974	0.949	0.961
UMN validation data				
Structured data only	With at least 2 AD diagnoses	0.550	0.846	0.667
Structured and unstructured data	With at least 2 AD diagnoses, and with keywords in AD encounters	0.475	0.905	0.623

TABLE 5 Characteristics of the definitive AD cohort and the potential AD cohort within UFHealth IDR, UTHealth physicians CDW, and UMN AHC-IE CDR.

	UFHealth IDR			UTHealth Physicians CDW			UMN AHC-IE CDR		
	Definitive AD cohort[a]	Potential AD cohort[b]		Definitive AD cohort	Potential AD cohort		Definitive AD cohort	Potential AD cohort	
Parameter	$N = 2{,}756$	$N = 5{,}263$	p-Value	$N = 1{,}823$	$N = 6{,}821$	p-Value	$N = 5{,}311$	$N = 10{,}387$	p-Value
Age									
<55	53 (1.9%)	111 (2.1%)	<0.05	124 (6.80%)	433 (6.35%)	<0.05	52 (1.0%)	105 (1.0%)	<0.05
55-64	194 (7.0%)	364 (6.9%)		327 (17.94%)	981 (14.38%)		170 (3.2%)	354 (3.4%)	
65-74	665 (24.1%)	1,165 (22.1%)		607 (33.30%)	2,011 (29.48%)		778 (14.6%)	1,347 (13.0%)	
75-84	1,178 (42.7%)	2,183 (41.5%)		620 (34.01%)	2,472 (36.24%)		2,136 (40.2%)	3,939 (37.9%)	
>=85	666 (24.2%)	1,440 (27.4%)		144 (7.90%)	911 (13.36%)		2,175 (41.0%)	4,642 (44.7)	
Unknown	0 (0%)	0 (0%)		1 (0.05%)	13 (0.19%)		0 (0%)	0 (0%)	
Sex									
Male	1,028 (37.3%)	2,023 (38.4%)	0.331	669 (36.70%)	2,490 (36.50%)	0.696	1,783 (33.6%)	3,620 (34.9%)	0.114
Female	1,728 (62.7%)	3,240 (61.6%)		1,152 (63.19%)	4,317 (63.29%)		3,528 (66.4%)	6,766 (65.1%)	
Unknown	0 (0%)	0 (0%)		2 (0.11%)	14 (0.21%)		0 (0%)	0 (0%)	
Race/ethnicity									
Hispanics	148 (5.4%)	239 (4.5%)	<0.05	59 (3.24%)	492 (7.21%)	<0.05	50 (0.9%)	98 (0.9%)	<0.05
NHW	1,828 (66.3%)	3,648 (69.3%)		1,011 (55.46%)	2,883 (42.27%)		4,681 (88.1%)	9,253 (89.1%)	
NHB	643 (23.3%)	1,094 (20.8%)		323 (17.72%)	1,548 (22.69%)		142 (2.7%)	275 (26.5%)	
Other	90 (3.3%)	136 (2.6%)		291 (15.96%)	1,295 (18.99%)		438 (8.2%)	84 (0.8%)	
Unknown	47 (1.7%)	146 (2.8%)		139 (7.62%)	603 (8.84%)		396 (7.5%)	677 (6.5%)	
Chronic conditions									
Depression	1,162 (42.2%)	1,972 (37.5%)	<0.05	710 (38.95%)	1,857 (27.22%)	<0.05	2,530 (47.6%)	4,531 (43.6%)	<0.05
Diabetes	876 (31.8%)	1,611 (30.6%)	0.270	412 (22.60%)	1,815 (26.61%)	<0.05	1,271 (23.9%)	2,549 (24.5%)	0.407
Hypertension	2,108 (76.5%)	3,921 (74.5%)	<0.05	1,226 (67.25%)	4,489 (65.81%)	0.249	4,154 (78.2%)	7,951 (76.6%)	<0.05
Cancer	852 (30.9%)	1,414 (26.9%)	<0.05	455 (24.96%)	1,219 (17.87%)	< 0.05	1,668 (25.2%)	3,028 (29.1%)	<0.05

Abbreviations: NHB = Non-Hispanic Black; NHW = Non-Hispanic White; UFHealth IDR = University of Florida Health Integrated Data Repository; UTHealth Physicians CDW = University of Texas Health Science Center at Houston (UTHealth) Physicians Clinical Data Warehouse (CDW); UMN AHC-IE CDR = University of Minnesota (UMN) Academic Health Center Information Exchange (AHC-IE) clinical data repository (CDR).

[a] The definitive AD cohort identified through the developed best performance CP rules.

[b] The potential AD cohort identified through the "having at least 1 AD-related diagnosis code" criterion.

TABLE 6 Performance of the baseline CP rule and the best-performing CP rules on different demographic subgroups in the UF-site chart-reviewed cohort.

Parameter	Rule	Sensitivity	PPV	F1-score
Sex: Female (*N* = 229)				
Baseline	With at least 1 AD diagnoses	1.000	0.402	0.573
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.946	0.664	0.780
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.913	0.757	0.828
Sex: Male (*N* = 134)				
Baseline	With at least 1 AD diagnoses	1.000	0.410	0.582
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.982	0.675	0.800
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.982	0.711	0.824
Age: < = 64 (*N* = 72)				
Baseline	With at least 1 AD diagnoses	1.000	0.361	0.531
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.962	0.641	0.769
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.962	0.694	0.806
Age: > 64 (*N* = 291)				
Baseline	With at least 1 AD diagnoses	1.000	0.416	0.587
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.959	0.674	0.792
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.934	0.748	0.831
Race: Non-Hispanic White (*N* = 252)				
Baseline	With at least 1 AD diagnoses	1.000	0.381	0.552
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.958	0.630	0.760
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.938	0.714	0.811
Race: Non-Hispanic Black (*N* = 67)				
Baseline	With at least 1 AD diagnoses	1.000	0.448	0.619
Structured data only (best F1-score)	With at least 2 AD diagnoses	0.933	0.718	0.812
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	0.933	0.718	0.812
Race: Hispanic (*N* = 24)				
Baseline	With at least 1 AD diagnoses	1.000	0.583	0.737
Structured data only (best F1-score)	With at least 2 AD diagnoses	1.000	0.875	0.933
Structured and unstructured data (best F1-score)	With at least 2 AD diagnoses, and with keywords in AD encounters	1.000	0.933	0.966

CP (i.e., "*with at least 2 AD diagnoses, and with keywords in AD encounters*") demonstrates consistently satisfactory performance (i.e., F1-score > 0.8) across all demographic subgroups. Meanwhile, the best-performing CP rule, considering both structured and unstructured data, consistently overperformed the best-performing CP, which only considers structured data.

4 | DISCUSSION

In this study, we successfully developed and validated CP algorithms for identifying AD patients in EHRs, leveraging information from multiple EHR domains. Our study extended previous work on AD patient identification in several significant ways. First, in the development of the CP algorithm, we introduced more flexibility in the inclusion and exclusion criteria by considering and testing multiple EHR domains (i.e., medications, procedures, and keywords) in addition to AD/ADRD diagnosis codes. Second, in addition to considering information directly related to AD, we also considered diagnosis codes for cognitive decline. These codes may be recorded for patients with known AD status, and including them in the CP algorithms further improved its coverage and robustness. Third, our final CP algorithm is simple (i.e., "*patient has at least 2 AD diagnoses and AD-related keywords in AD encounters*"), making it readily applicable to other EHR systems. Despite the use of varied data models by institutions in our validation study, the algorithm's focus on diagnosis codes and clinical notes allowed for easy adaptation without significant effort. We employed code matching in diagnosis/condition tables and regular expressions in clinical notes, streamlining the adaptation process across diverse data architectures. Further, compared with previous studies,[12,14–17] our algorithm demonstrated superior performance, achieving higher sensitivity while maintaining comparable PPV. Our final algorithm achieved a perfect sensitivity on the testing dataset, indicating that it can correctly identify all patients who truly have AD.

Nevertheless, there were a few false positives because: (1) in the unstructured data, the patient is recorded as "*suspicious of two or more subtypes of ADRD*" but was either diagnosed as "*having subtypes of ADRD other than AD*" or there is no conclusion yet at the time of our chart review; (2) in the unstructured data, the patient was recorded as "*has dementia, possibly Alzheimer's*," but whether the patient truly had AD was not confirmed; and (3) potential document errors, where the patient had not been diagnosed with AD, but the condition was listed incorrectly in patient's chart because of suspicion of AD. Our high-performing EHR-based CPs provide the opportunity for fast and accurate identification of AD patients from EHR systems, which can be used to build patient cohorts for research, clinical care, and public health initiatives. Our CP algorithms successfully identified a total of 2,756, 1,823, and 5,311 AD patients across the three sites, respectively, with most of them older than 65. The statistical analysis of both definitive and potential AD cohorts reveals significant differences in demographic attributes, notably age and race/ethnicity, as well as their clinical characteristics. The baseline CP rule incorporates a large number of non-AD patients to form a broad AD cohort with low PPV, which significantly deviates from the precise AD cohort identified by our designed best-performing CP rule, thereby impacting the cohort's characteristics. Such disparities between the broad and precise cohorts could potentially skew findings in downstream analyses. This underscores the importance and benefits of our validated and well-performed CP rules in EHR-data-based AD research.

In addition, our overall best-performing CP demonstrated consistently superior performance across a wide range of demographic subgroups, showcasing its robustness and ensuring fairness in its application. The analysis revealed that the CP, which incorporates both structured and unstructured data, significantly exceeded the performance of the CP which utilizes only structured data. This pattern holds true across all demographic groups, thereby emphasizing the substantial value of integrating unstructured data into CP development.

Our study has several limitations. One concerns the generalizability of our best-performing CP rules. While validating the best-performing CPs at different sites (i.e., UTHealth and UMN), we observed substantial variability in their performance, particularly in sensitivity metrics. The low sensitivity at UMN for both CP rules—the one relying solely on structured data and the one incorporating both structured and unstructured data—suggests a generalizability issue. This problem may stem from differences in data distribution across sites, such as variations in demographics, clinical practices, or documentation styles. Consequently, these differences hinder the application of our CP rules in settings that are markedly different from the UFHealth population. To address this issue, one potential strategy is to fine-tune the CP rules at each site, particularly if initial evaluations show subpar performance. Fine-tuning involves enhancing the precision and breadth of diagnosis codes and keywords, which includes eliminating those with poor PPV in identifying AD patients and incorporating additional relevant codes and keywords tailored to the specific site. These findings underscore the challenges in developing robust CP algorithms across diverse settings, emphasizing the necessity of incorporating data from a variety of sources throughout the CP development process, not merely during the validation phase. Another potential strategy is using federated learning[29] to improve CP generalizability and applicability. This approach allows for collaborative development across multiple sites without direct data sharing, thus addressing privacy and data exchange concerns while aiming for a universally effective CP algorithm.

Further inspection of our CP revealed two additional limitations regarding AD-related keywords. In the current developed CP algorithm, we focused solely on AD-relevant keyword matchings without considering their context within unstructured clinical notes. For instance, we overlooked negations (e.g., "*the patient does not have cognitive impairment*") and references to third parties (e.g., "*he lived with a relative who has cognitive impairment*"). Moreover, the keyword list from UFHealth did not work consistently well on other sites. This was evident in UMN's validation, where the CP rule that included unstructured data underperformed compared to the structured data-only rule. This discrepancy could be due to lexical variations (i.e., differences in word usage for the same medical concept) in clinical documentation across institutions,[30] rendering the UFHealth-developed keywords less effective at UMN. However, a previous study[30] indicated that while lexical variations are notable, semantic-level information (i.e., the meaning and context of the medical concept) might remain relatively consistent, suggesting that advanced natural language processing methods that could understand documentation from the semantic level could enhance CP algorithm accuracy in the future. For instance, leveraging large medical language models,[31,32] known for their proficiency in understanding and answering questions over free-text clinical notes, could transform the AD-related keyword-matching approach into a more accurate and robust question-answering approach.

In sum, we have successfully developed and rigorously validated CP algorithms for accurately identifying AD patients in large medical databases. The final CP can be effectively applied in structured data alone or in combination with unstructured clinical notes. The CPs we developed achieved good overall performance. The AD patient cohort identified through our CP can be used in downstream analysis to provide real-world evidence in understanding the disease burdens, social and behavioral determinants of health, patterns in utilization of services, and health outcomes in AD patients.

AUTHOR CONTRIBUTION

X.H., Z.C., Q.W., T.L., S.B., and J.B. were responsible for the overall design, development, and evaluation of this study. X.H., Z.C., Q.W., and J.B. did the initial drafts and revisions of the manuscript. J.B. was responsible for conceptualizing the research. J.T., L.L., Y.Z., F.W., H.X., Y.C., R.Z., Z.L. helped with the external validation of the algorithm. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

ACKNOWLEDGMENTS

This study was partially supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), grants from National Institute on Aging R01AG080991, R01AG080624, R01AG076234, R01AG078154 and R56AG069880, and an Ed and Ethel Moore Alzheimer's Disease Research Program from the Florida Department of Health (FL DOH #23A09). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

CONFLICT OF INTEREST STATEMENT

Dr. Steven T. DeKosky reports royalties or licenses from UpToDate (point of care electronic medical text), editor for dementia; consulting fees from Boxer Capital, Brainstorm Cell Therapeutics, Lundbeck Pharmaceuticals, Amylyx Pharmaceuticals, Reata Pharmaceuticals, Biogen; payment or honoraria from NYU Grant Rounds, SUNY Downstate; participation on a Data Safety Monitoring Board or Advisory Board for Acumen Pharmaceuticals, Biogen DSMB, Cognition Therapeutics, Prevail Pharmaceuticals, Vaccinex; leadership role as Associate Editor, Neurotherapeutics. Dr. Jingchuan Guo reports grants or contracts from NIH; consulting fees from Pfizer. Huilin Tang reports grants or contracts from AFPE Predoctoral Fellowship, PhRMA Foundation Predoctoral Fellowship. Yonghui Wu reports support for the present manuscript from PCORI ME-2018C3-14754, NIA R56AG069880; payment or honoraria from IEEE BHI. Other authors have no conflicts of interest to disclose. Author disclosures are available in the supporting information.

CONSENT STATEMENT

This is a secondary analysis of electronic health records, where a waiver of consent was approved with an approved protocol by the University of Florida IRB.

ORCID

Xing He https://orcid.org/0000-0003-0290-8058
Jiang Bian https://orcid.org/0000-0002-2238-5429

REFERENCES

1. Rajan KB, Weuve J, Barnes LL, McAninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers Dement*. 2021;17(12):1966-1975. doi:10.1002/alz.12362
2. Nichols E, Vos T. Estimating the global mortality from Alzheimer's disease and other dementias: a new method and results from the Global Burden of Disease study 2019. *Alzheimers Dement*. 2020;16(S10):e042236. doi:10.1002/alz.042236
3. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*. Published online March 14, 2023. doi:10.1002/alz.13016
4. NAPA—National Alzheimer's Project Act. ASPE. Accessed March 25, 2022. https://aspe.hhs.gov/collaborations-committees-advisory-groups/napa
5. of the Commissioner O. Real-World Evidence. Published April 7, 2020. Accessed May 7, 2020. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
6. Office of the Commissioner. Real-World Evidence. Published 2021. Accessed August 7, 2021. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
7. Miller AH, Marra DE, Wu Y, et al. Characterizing dementia prevalence in the State of Florida: an electronic health record study. *Alzheimers Dement*. 2021;17(S10):e052364. doi:10.1002/alz.052364
8. Marra DE, Miller AH, Li Q, et al. Utilizing electronic medical record data to predict onset of Alzheimer's disease and related dementias. *Alzheimers Dement*. 2020;16(S10):e041233. doi:10.1002/alz.041233
9. Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. *PLoS One*. 2019;14(7):e0203246. doi:10.1371/journal.pone.0203246
10. Barnes DE, Zhou J, Walker RL, et al. Development and validation of eRADAR: a tool using EHR data to detect unrecognized dementia. *J Am Geriatr Soc*. 2020;68(1):103-111. doi:10.1111/jgs.16182
11. Nori VS, Hane CA, Crown WH, et al. Machine learning models to predict onset of dementia: a label learning approach. *Alzheimers Dement*. 2019;5:918-925. doi:10.1016/j.trci.2019.10.006
12. Wilkinson T, Schnier C, Bush K, et al. Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data. *Eur J Epidemiol*. 2019;34(6):557-565. doi:10.1007/s10654-019-00499-1
13. Rizzuto D, Feldman AL, Karlsson IK, Dahl Aslan AK, Gatz M, Pedersen NL. Detection of dementia cases in two swedish health registers: a validation study. *J Alzheimers Dis*. 2018;61(4):1301-1310. doi:10.3233/JAD-170572
14. Phung TKT, Andersen BB, Høgh P, Kessing LV, Mortensen PB, Waldemar G. Validity of dementia diagnoses in the Danish hospital registers. *Dement Geriatr Cogn Disord*. 2007;24(3):220-228. doi:10.1159/000107084
15. Salem LC, Andersen BB, Nielsen TR, et al. Overdiagnosis of dementia in young patients—a nationwide register-based study. *Dement Geriatr Cogn Disord*. 2012;34(5-6):292-299. doi:10.1159/000345485
16. Taylor DH Jr, Østbye T, Langa KM, Weir D, Plassman BL. The accuracy of Medicare claims as an epidemiological tool: the case of dementia revisited. *J Alzheimers Dis*. 2009;17(4):807-815. doi:10.3233/JAD-2009-1099
17. Tjandra D, Migrino RQ, Giordani B, Wiens J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimers Dement*. 2020;6(1):e12035. doi:10.1002/trc2.12035
18. Electronic Health Records-Based Phenotyping. Published June 27, 2014. Accessed May 15, 2022. https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/
19. Liu Y, Siddiqi KA, Cook RL, et al. Optimizing identification of people living with HIV from electronic medical records: computable phenotype development and validation. *Methods Inf Med*. 2021;60(3-04):84-94. doi:10.1055/s-0041-1735619
20. McDonough CW, Babcock K, Chucri K, et al. Optimizing identification of resistant hypertension: computable phenotype development and validation. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1393-1401. doi:10.1002/pds.5095
21. Guo Y, He X, Lyu T, et al. Developing and validating a computable phenotype for the identification of transgender and gender nonconforming individuals and subgroups. *AMIA Annu Symp Proc*. 2020;2020:514-523. https://www.ncbi.nlm.nih.gov/pubmed/33936425
22. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;17(1):128-144. https://www.ncbi.nlm.nih.gov/pubmed/18660887. Published online.

23. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230. doi:10.1136/amiajnl-2013-001935
24. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J Am Med Inform Assoc*. 2013;20(e2):e341-e348. doi:10.1136/amiajnl-2013-001939
25. He X. *Ad_ehr_computable_phenotype*. Github Accessed April 19, 2024. https://github.com/YeechingTiger/ad_ehr_computable_phenotype
26. About Us » Integrated Data Repository Research Services » Clinical and Translational Science Institute » University of Florida. Accessed April 6, 2022. https://idr.ufhealth.org/about-us/
27. Carlson C, Group Health Cooperative. Dementia. PheKB. Accessed April 19, 2022. https://phekb.org/phenotype/dementia
28. 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. doi:10.1371/journal.pone.0199026
29. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37(3):50-60. doi:10.1109/MSP.2020.2975749
30. Sohn S, Wang Y, Wi CI, et al. Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. *J Am Med Inform Assoc*. 2018;25(3):353-359. doi:10.1093/jamia/ocx138
31. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194. doi:10.1038/s41746-022-00742-2
32. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *NPJ Digit Med*. 2023;6(1):210. doi:10.1038/s41746-023-00958-w

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: He X, Wei R, Huang Y, et al. Develop and validate a computable phenotype for the identification of Alzheimer's disease patients using electronic health record data. *Alzheimer's Dement*. 2024;16:e12613. https://doi.org/10.1002/dad2.12613