UCSF
UC San Francisco Previously Published Works

Title

Dementia and electronic health record phenotypes: a scoping review of available phenotypes and opportunities for future research.

Permalink

https://escholarship.org/uc/item/8mv4q340

Journal

Journal of the American Medical Informatics Association : JAMIA, 30(7)

ISSN

1067-5027

Authors

Walling, Anne M
Pevnick, Joshua
Bennett, Antonia V
et al.

Publication Date

2023-06-01

DOI

10.1093/jamia/ocad086

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of the American Medical Informatics Association, 2023, **30(7)**, 1333–1348
https://doi.org/10.1093/jamia/ocad086
Advance access publication 30 May 2023
Review

Review

Dementia and electronic health record phenotypes: a scoping review of available phenotypes and opportunities for future research

Anne M. Walling[1,2], Joshua Pevnick[3], Antonia V. Bennett[4], V. G. Vinod Vydiswaran[5], Christine S. Ritchie[6]

[1]Department of Medicine, VA Greater Los Angeles Health System, Los Angeles, California, USA
[2]Department of Medicine, University of California, Los Angeles, Los Angeles, California, USA
[3]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA
[4]Department of Health Policy and Management, University of North Carolina, Chapel Hill, North Carolina, USA
[5]Department of Learning Health Sciences, University of Michigan, Ann Arbor, Michigan, USA
[6]Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
Corresponding Author: Anne M. Walling, MD, PhD, 1100 Glendon Ave, Los Angeles, CA 90095, USA; awalling@mednet.ucla.edu

ABSTRACT

Objective: We performed a scoping review of algorithms using electronic health record (EHR) data to identify patients with Alzheimer's disease and related dementias (ADRD), to advance their use in research and clinical care.

Materials and Methods: Starting with a previous scoping review of EHR phenotypes, we performed a cumulative update (April 2020 through March 1, 2023) using Pubmed, PheKB, and expert review with exclusive focus on ADRD identification. We included algorithms using EHR data alone or in combination with non-EHR data and characterized whether they identified patients at high risk of or with a current diagnosis of ADRD.

Results: For our cumulative focused update, we reviewed 271 titles meeting our search criteria, 49 abstracts, and 26 full text papers. We identified 8 articles from the original systematic review, 8 from our new search, and 4 recommended by an expert. We identified 20 papers describing 19 unique EHR phenotypes for ADRD: 7 algorithms identifying patients with diagnosed dementia and 12 algorithms identifying patients at high risk of dementia that prioritize sensitivity over specificity. Reference standards range from only using other EHR data to in-person cognitive screening.

Conclusion: A variety of EHR-based phenotypes are available for use in identifying populations with or at high-risk of developing ADRD. This review provides comparative detail to aid in choosing the best algorithm for research, clinical care, and population health projects based on the use case and available data. Future research may further improve the design and use of algorithms by considering EHR data provenance.

Key words: dementia, electronic health record phenotype

INTRODUCTION

There is tremendous unmet need for improving the quality of care for persons with Alzheimer's disease and related dementias (ADRD).[1] In 2023, ADRD affected 6.7 million Americans age 65 years and older and their caregivers[2]; it is the 6th leading cause of death among all adults in the United States. Furthermore, caregiver strain can be significant.[3–5] Patients with dementia have an average life expectancy of 4–8 years after diagnosis, and advanced dementia is characterized by prolonged disability.[6–9] A recent large systematic review of 627 dementia care intervention studies called for "larger, longer-term, and more rigorous" interventions to improve care delivery and there is increasing interest in how health systems can provide better care.[10,11]

Most information about dementia prevalence and disease trajectory comes from primary data collection, Medicare claims data, the Minimum Data Set (MDS)[12–15] and the Health Retirement Study.[16] However, these data sources have limited value for real-time patient identification. Electronic Health Record (EHR) phenotypes provide a significant opportunity to identify relevant patients for clinical care, quality improvement, and research.[17–20] EHR phenotypes refer to clinical conditions or characteristics that can be determined using a computerized EHR query with a defined set of data elements.[17]

Creating reliable EHR phenotypes requires an iterative approach, and this work supports the formation of patient registries for targeted interventions and evaluation. Engaging with health systems to improve data capture in parallel with improving quality and conducting research offers the potential to improve data quality, and hence the specificity and sensitivity of the patient registries, while improving care over time. To optimally harness information contained within EHRs, we need to understand how EHRs have been used to date to identify and characterize patients living with dementia (PLWD).

Received: 15 November 2022. **Revised:** 27 April 2023. **Editorial Decision:** 29 April 2023. **Accepted:** 16 May 2023
Published by Oxford University Press on behalf of the American Medical Informatics Association 2023.
This work is written by (a) US Government employee(s) and is in the public domain in the US.

We therefore reviewed the literature to identify algorithms that used EHR data alone or in combination with claims or other non-EHR data to identify PLWD, based on the idea that EHR data could both improve discrimination as compared to administrative data alone and be immediately available for use in patient care, cohort identification, or pragmatic clinical trials.

MATERIALS AND METHODS

We used the Arksey and O'Malley framework (5 stages) and PRISMA checklist for Scoping Reviews to guide our literature search.[21] Our research question was: What types of EHR phenotypes have been developed to screen for or identify PLWD in the United States? We used the Using Population/Concept/Context (PCC) Framework[22–24] to refine our overall objectives. Our target population was patients with dementia. Our target concept was electronic health record (EHR) phenotypes. Because of differences in health care practice and billing in the United States and other countries, we limited our context to United States health systems to inform use in future clinical quality improvement and research.

We modeled this research on work by Lee et al[25] entitled "Electronic Medical Record-Based Case Phenotyping for the Charlson Conditions: Scoping Review.", which spanned several clinical conditions, including dementia. We included papers referenced in the dementia section from Lee's review that met our inclusion criteria and then expanded on Lee's work by updating the review in Pubmed from the end date of the search in Lee's article (April 1, 2020 through March 1, 2023) using the same search criteria for dementia ((((Electronic medical record or EMR or electronic medical records or EMRs or hospital information systems or HIS or electronic health records or EHR) AND (case or identification or ascertainment or diagnosis or phenotype)) AND (dementia or Alzheimer's disease or senile degeneration))) AND (("2020/04/01"[Date—Publication]: "2023/03/01"[Date—Publication])) and United States. We developed our protocol for eligibility criteria prior to conducting our cumulative update and the protocol is available upon request.

For the cumulative update, we searched the US-based phenotype registry, PheKB, and asked key experts for any additional articles not covered in our search.

Our eligibility criteria included the following inclusion and exclusion criteria:

Inclusion criteria

- Articles in English
- Articles based in the United States
- Includes Electronic Health Record (EHR) data alone to identify patients with dementia
- Includes EHR data in combination with claims or other non-EHR data to identify patients with dementia
- Includes a reference standard to test the EHR phenotype
- Includes information on the performance of the EHR phenotype

Exclusion criteria

- Not in English
- Evaluated data from countries other than the United States
- Does not have a reference standard to test the EHR phenotype
- Does not report on the performance of the EHR phenotype
- Articles that do not include dementia and only cover cognitive impairment
- *Note: We did not exclude papers that also reported on EHR phenotypes for other conditions, however we only covered details of the dementia algorithms for our results.

For our title review, we aimed to use broad criteria to capture any title that might cover our topic and thus at this level included titles that covered EHR phenotypes and/or diagnosis for dementia/Alzheimer's disease, cognitive impairment or illnesses in general, titles that discussed using data and/or electronic health record to identify patients with or at high risk for dementia, and titles that cover machine learning if it relates to identifying patients with or at high risk for cognitive impairment or dementia. For abstract and full paper review, we applied our detailed inclusion and exclusion criteria listed above. We used dual review for all stages (AMW and JP). If either reviewer selected a title or abstract, it moved on to the next stage of review. For the full paper review, we met to resolve differences and ensured agreement prior to data abstraction for inclusion in our final table.

Beyond the work initiated by Lee et al, which included several clinical conditions, we considered the clinical and research impacts specific to identification of patients with dementia to inform the data we would abstract from articles to populate the main table. We used an iterative process using 2 reviewers (AMW and JP) with over-reading by other team members to finalize the most relevant columns in the table for the final abstraction tool. Specifically, we abstracted whether the algorithm was designed to identify at risk populations or diagnosed populations, the studied population, data elements from the EHR that were used to create the phenotype, what reference standard was used to test the phenotype, and the performance of the EHR-based phenotype compared to the reference standard. We also specifically noted limitations of the papers such as algorithm performance or narrow criteria for tested population (eg, sample derived only from the intensive care unit). All investigators reviewed the final articles and agreed with the content of the final table.

Finally, for algorithms that used International Classification of Disease (ICD) codes to identify patients with dementia, we compared which codes were used.

RESULTS

Figure 1 details the flow diagram for this scoping review of dementia EHR phenotypes. For our cumulative update of the literature, we reviewed 271 titles meeting our search criteria, 49 abstracts (73% [$n = 36$] agreement with 13 titles included only selected by one reviewer), and 26 full text papers (88% agreement [$n = 23$] with 3 abstracts only selected by one reviewer). We only differed on one article in our initial rating for full text papers which was easily resolved with discussion. Overall we identified 8 articles from the original systematic review, 8 from our new search, and 4 recommended by an expert[26–29] that met our inclusion criteria. PheKB search revealed one algorithm that was already captured in the literature review, thus the table was annotated that further details for that algorithm was available in the PheKB database.[30]

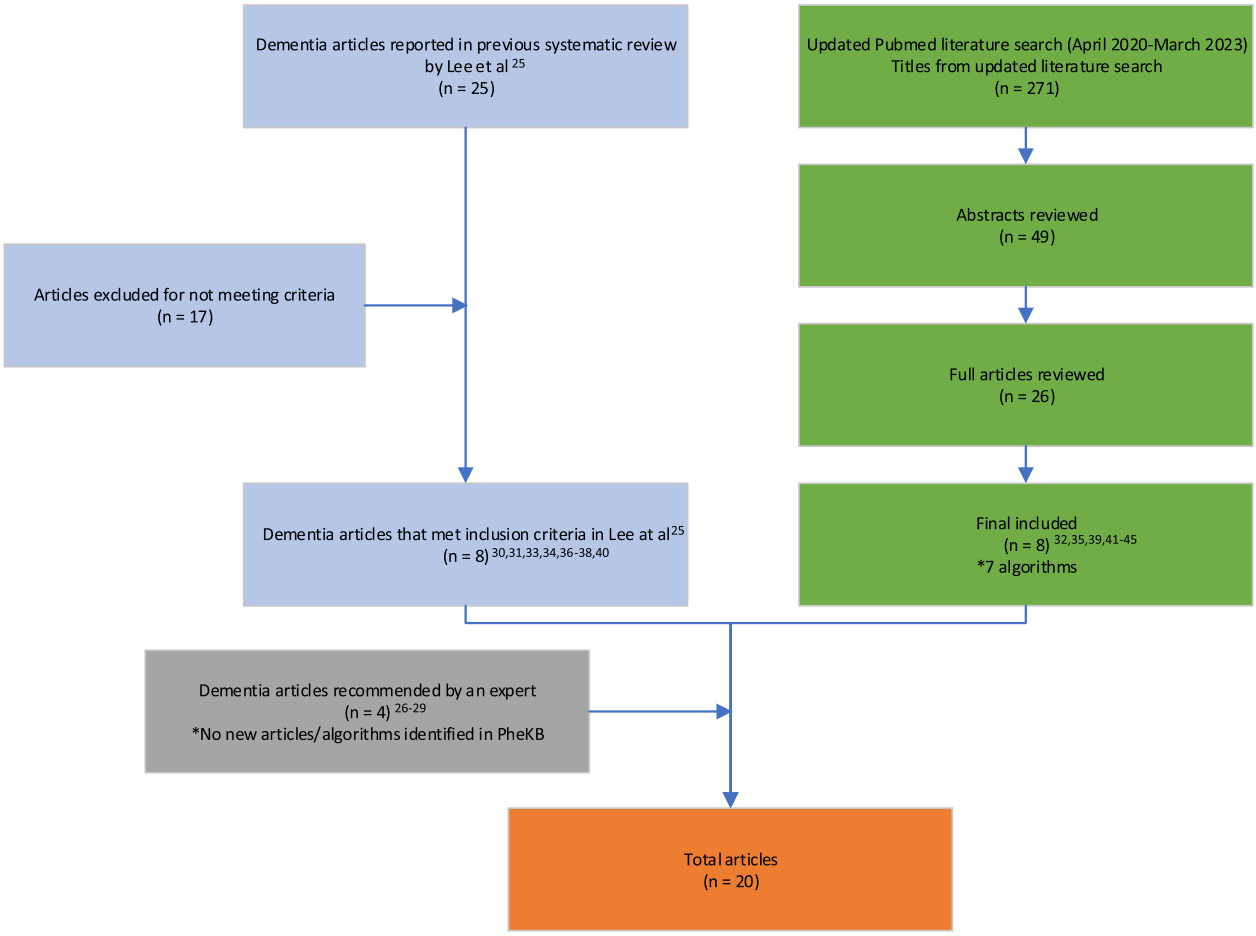

Figure 1. Flow diagram for scoping review of dementia electronic health record phenotypes.

Thus, the algorithms from 20 articles—19 unique algorithms—are the focus of this review. Tables 1 and 2 summarize the EHR-based dementia phenotyping algorithms we identified.[26–45] We noted 2 main categories of EHR phenotypes: algorithms to identify patients with a dementia diagnosis in the EHR and algorithms to identify patients at high risk of dementia. We abstracted the same data from both types of article, but present them separately in 2 tables for clarity.

Several studies in these tables used an iterative process and tested multiple strategies to identify the highest performing phenotype.[17] The phenotypes used both structured and unstructured data elements. Structured data elements include ICD codes, visits with specialists (eg, neurology consultations, enrollment in a dementia care program), medication use (eg, acetylcholinesterase inhibitors, N-methyl-D-aspartate receptor antagonists), and cognitive screening tests. Use of unstructured data elements, such as free text in clinical notes, relied on methods such as regular expression natural language processing (NLP) and machine learning.

Algorithms to identify patients diagnosed with dementia

Seven algorithms were developed to identify patients already diagnosed with dementia (Table 1). Thus, achieving high specificity was a primary goal. Of these 7, 4 were compared to a reference standard of manual chart abstraction, one was compared to a reference standard of a comprehensive in-person cognitive evaluation including neurological and neuropsychological testing, one was compared to a cognitive score that was available in the EHR and one was only based on SNOMED codes without further validation. All of these algorithms used ICD codes as predictors. The most commonly used data elements included ICD codes, medications, and key words using NLP. The approaches that showed the best performance for identifying patients diagnosed with dementia (usually measured by positive predictive value (PPV) or specificity) used either more than one ICD code, supplemental NLP methods, or keyword searches along with ICD codes. A noted limitation of all studies is that the provenance of ICD-codes was rarely indicated (for example problem list vs encounter diagnosis vs lab test). Further, many of the high performing algorithms were applied to limited environments, such as the intensive care unit. One study that evaluated EHR phenotypes for 10 different diseases concluded that dementia was one of the most challenging conditions to characterize using their algorithmic approach.[34]

Algorithms to identify patients at high risk for current or future diagnoses of dementia

The other 12 algorithms were designed to identify patients at high risk for current or future diagnoses of dementia (Table 2). These algorithms might be especially useful to identify patients for further testing, to address dementia prevention, or study prognosis, protective or risk factors. One used a detailed clinical assessment as a reference standard, 2 used detailed chart review and/or communication with a treating

Table 1. Characteristics of algorithms to identify patients with diagnosed dementia in the electronic health record

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
Amra S, et al.[31]	Goal to identify patients diagnosed with dementia for pragmatic research	Adult patients who were admitted to one of the ICUs at Mayo Clinic in Rochester, MN between Jan 2, 2006 and December 31, 2014 and were cognitively evaluated by the Mayo Clinic Study of Aging (n=993)	ICD codes and keywords Manual chart review conducted to identify keywords for an electronic search algorithm: dementia, cognitive impairment, cognitive deficit, cognitive decline, mild cognitive impairment, impaired memory, impaired judgment, impaired orientation, difficulty concentrating, patient is not independent in handling finances	Comprehensive in-person cognitive evaluation by the Mayo Clinic Study of Aging (including neurological and neuropsychological testing).	Dementia ICD-based: specificity 99% and sensitivity 79% With keyword using iterative approach: 97% sensitivity and 99% specificity	Sample was ICU only, it will likely perform better here than in other populations with less utilization.
Harding BN, et al.[32]	Goal to identify whether cognitive tests, as structured data elements in an electronic health record, can be used to optimize EHR algorithms for dementia	3,690 Kaiser patients and 2,981 VA Health System patients with cognitive testing results in EHR	Looked at number and timing of ICD codes, medications, and specialist visits	< 24/30 points for the MMSE or < 21/30 for SLUMS, among the small minority of patients who had received those tests	Algorithm with 2 ICD codes within 12 months had best performance: 65% met dementia threshold with MMSE and 77.4% met dementia threshold with SLUMS	
Kho AN, et al.[30]	Goal to identify patients diagnosed with dementia use in genome-wide association studies (GWAS)	(N=747 cases and N=2043 controls) from Group Health Seattle Biobank with comprehensive vendor-based EHR since 2004 with 20 plus years of pharmacy data and 15 plus years of ICD9 data	Diagnosis, medications using structured data, free text searches and manual chart review (Details on ICD codes and free text searches not provided)	Review of clinical notes and demographics (details not provided)	PPV 73%	*Note that this algorithm was identified in the PhekB database and additional details available.
Reuben DB, et al.[33]	Goal to identify patients diagnosed with dementia for use in clinical practice, quality improvement and research	Chart abstractions completed on 124 patients and analysis also completed on 989 patients enrolled in Dementia Care Program	ICD codes and NLP (medications tested but not included in final model because of prescribing for "off-label" conditions such as cholinesterase inhibitors for mild cognitive impairment, memantine for migraine headaches)	Chart reviews initially and then enrollment in dementia program (n=989)(internal reference standard)	For age 65 years and older: 91% or greater PPV and 62% or higher sensitivity for ICD and NLP approach	

(continued)

Table 1. (continued)

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
Singh B, et al.[26]	Goal to identify patients diagnosed with dementia for use in clinical practice, quality improvement, and research	1447 patients aged 18 years and older who gave research authorization and admitted to ICU during 2006 from Olmsted County, Minnesota	Looked at ICD's alone and also Automated algorithm available in supplemental appendix that used several other data elements in addition to ICD codes	Trained research fellows manually collected data according to definition published by Charlson. Only the medical and surgical history sections of the clinical notes were ascertained. N=240 charts were abstracted for validation cohort. Original cohort was from 1447 from an observational cohort study	For dementia, ICD codes alone had 100% specificity and 8% sensitivity with PPV=50% and NPV=95%. With the tailored algorithm the sensitivity was 92% and specificity was 99% (PPV was 92% and NPV was 100%).	Sample was ICU only, it will likely perform better here than in other populations with less utilization.
Wei WQ, et al.[34]	Goal to identify patients diagnosed with dementia for use in research	Vanderbilt University Medical Center de-identified data from 10 pre-selected diseases (one of which was Alzheimer's disease). The study included 175 patients for each disease.	ICD codes, Clinical Notes (using keywords), Medications	A group of 25 patients (per disease category) were randomly selected and reviewed, negative and uncertain patients classified as untrue. Inter-rater reliability was calculated: 0.68-0.9	One ICD-9 PPV =.0.28 ≥2 ICD9s PPV=0.74 ≥2 components=0.88	This paper looks at phenotypes for 10 different conditions. They point out in the discussion that dementia (and breast cancer) were the two most challenging phenotypes. Using more than 1 ICD code improved specificity (but decreased sensitivity). NLP improved sensitivity but lowered PPV.
Xu J, et al.[35]	Goal to identify patients with probable Alzheimer's disease and related dementia phenotypes	7,587 patients over 65 years old seen at a large, multispecialty urban academic medical center in New York (792 cases and 6795 controls)	EHR demographics, comorbidities (32 diagnoses), medications (171 medication classes)	Reference standard based on one or more AD-related SNOMED-CT concepts coming from diagnosis data (Alzheimer's disease; Primary degenerative dementia of the Alzheimer type, presenile onset; Primary degenerative dementia of the Alzheimer type, senile onset; Mild cognitive disorder; Minimal cognitive impairment)	AUC=0.764	Imperfect reference standard does not include any external validation

Table 2. Characteristics of algorithms to identify patients at high risk of dementia

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
Barnes DE, et al.[36]	The goal of this study was to develop and validate an EHR-based prediction tool to identify patients at high risk of unrecognized dementia.	Setting: Kaiser Permanente Washington (KPWA), an integrated healthcare delivery system 498 participants in the Adult Changes in Thought (ACT) study, who were found to have dementia based on a comprehensive process to detect and diagnose dementia conducted every 2 years and who had linked EHR data. Patients with existing dementia or memory loss diagnosis codes or dementia medication fills were excluded.	EHR predictors included demographics, medical diagnoses, vital signs, healthcare utilization and medications within the previous 2 years.	Cognitive Abilities Screening Instrument (CASI) and if abnormal in-depth evaluation including a neuropsychological test battery, physical examination including neurologic assessment, and detailed review of medical records. Imaging ordered if needed. Results confirmed by multidisciplinary consensus committee based on standard research criteria.	The final 31-predictor model included markers of dementia related symptoms, healthcare utilization pattern and dementia risk factors. Discrimination was good (c statistic, 0.78, 95% CI : 0.76,0.81) for development and validation (0.81; 0.78,0.84).	This paper describes an EHR-based tool to help detect patients with unrecognized dementia. Patients scoring in the top 5% were estimated to have a 16% chance of having unrecognized dementia vs 3% in the rest of the sample.
Ben Miled Z, et al.[29]	The purpose of the model is to automate the cost-effective, non-invasive, digital pre-screening of patients at risk for dementia.	2159 cases and 11,558 controls from 15 and 25 different institutions, respectively.	Potential predictors included age, gender, race, institution, medications, diagnoses, and note text.	Cases were identified via diagnosis codes.	The final model was generalizable across multiple institutions and predicted dementia within 1 year of its onset with sensitivity, specificity, and accuracy of 74.0%, 72.6%, and 72.6%, respectively.	
Boustani M, et al.[37]	Goal to develop a multivariable, validated, and generalizable model using structured and unstructured EHR data to identify early prodromal ADRD signatures up to 5 years before diagnosis.	Data from the Indiana Network for Patient Care (INPC) health information exchange that includes structured and unstructured (visit notes, progress notes, medication notes) EHR data. Cases and controls were matched on age, race, and sex. The derivation sample consisted of 10 504 cases and 39 510 controls; the validation sample included 4500 cases and 16 952 controls. Cases were required to be active in the INPC by having at least 1 medical encounter every other year for 10 years leading up to their first ADRD-coded visit.	Conceptual model-based algorithm—includes structured and unstructured data based on expert input. 14 diagnostic risk variables and 10 drug classes in addition to new variables produced from unstructured data (eg, disorientation, confusion, wandering, apraxia).	Adult patients with at least 2 clinical visits with ADRD diagnostic codes between Jan 1, 2008, and December 31, 2016.	Structured data plus unstructured data performed better than structured data alone based on AUROC of predicting diagnosis during the coming 1-10 years (.798 vs .689). Using a cutoff to maximize both sensitivity and specificity of 1-10 year prediction yielded 62% sensitivity and 88% specificity.	This paper aimed to identify future diagnosis of ADRD. We report data here for models predicting development of ADRD in the coming 1-10 years.

(continued)

Table 2. (continued)

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
Ernecoff NC, et al.[38]	To develop an electronic health record (EHR) phenotype to identify inpatients with late-stage dementia for a clinical trial of palliative care consultation.	Out of 1615 adult patients admitted to a tertiary academic hospital, this algorithm identified 371 patients with EHR data suggesting possible late-stage dementia. After manual chart review by research staff, 174 were referred to a palliative care physician, who served as a reference standard. 91 of these patients were confirmed to have advanced dementia after chart review.	The final algorithm used age and 34 ICD codes associated with dementia or cognitive impairment.	Palliative care physician determination based on review of the EHR and conversation with patient's attending physician to confirm dementia diagnosis and Global Deterioration Scale Stage 5–7	PPV of 76.3% for dementia and 24.5% for late stage dementia patients: a false discovery rate of 23.7% for dementia and 75.5%.Sensitivity 59.7% to identify hospitalized patients with dementia.	
Hane CA, et al.[39]	To facilitate identification of large potential patient populations for clinical trial recruitment.	Nori V, et al. Machine learning models to predict onset of dementia: a label learning approach. Alzheimer's Dement 2019; 5:918-925 Same sample as Nori 2019 study except that it is limited to patients that have 2 unique dates with a clinical note at least 31 days apart in a 2-year data collection period defined for each patient.	Adds proprietary NLP review of clinical notes to Nori 2019 study	Same as per Nori 2019	When using clinical notes, the area under the curve (AUC) improved from 0.85 to 0.94.	
Li Q, et al.[43]	To explore machine learning methods for early prediction of Alzheimer's disease and related dementias using real-world electronic health record (EHR) data.	23,835 ADRD and 1,038,643 control patients were identified from the OneFlorida+ Research Consortium, a clinical research network using data from EHRs, government claims, vital statistics, tumor registries, and other sources.	EHR data, including age, gender, race, ethnicity, marital status, smoking status, medications, vital signs, lab test results, procedural and diagnostic codes.	ADRD was defined using a combination of diagnostic codes and anti-dementia medications (ie, donepezil, galantamine, memantine, and rivastigmine) based on existing relevant computable phenotypes in the literature.	The gradient boosting tree models trained with the data-driven approach achieved the best area under the curve scores of 0.939, 0.906, 0.884, and 0.854 for early prediction of ADRD 0, 1, 3, or 5 years before diagnosis, respectively.	
McCoy TH, et al.[40]	To examine the association of cognitive symptoms documented in hospital discharge notes with incident	535,814 hospitalized adults from 2 large academic medical centers, of whom	Sociodemographic features (age at admission, sex, and race/ethnicity), administrative diagnostic	Diagnosis was ascertained using ICD 9 and 10 codes and problem list codes for dementia.	The C statistics for a diagnosis of dementia during up to 8 years of follow-up in the development cohort,	This paper was aimed at developing a tool to identify patients at risk for dementia.

(continued)

Table 2. (continued)

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
	dementia diagnosis during up to 8 years of follow-up.	10,418 (1.9%) received a new diagnosis of dementia	codes, and narrative hospital discharge summary notes. Applied NLP processing tool that identifies cognitive symptoms described in prior paper* to narrative hospital discharge summary notes. *McCoy TH et al. "High throughput phenotyping for dimensional psychopathology in EHRs"		the validation cohort, and a combined cohort were 0.61 (95% CI 0.60–0.61), 0.65 (95% CI 0.64–0.66), and 0.62 (95% CI 0.62–0.63), respectively.	They used EHR from 2 large academic medical centers and an NLP processing tool to estimate cognitive symptomatology. They showed that patients with cognitive symptoms are at higher risk for dementia up to 8 years in the future.
Nori VS, et al.[41]	To improve the accuracy of a predictive model for dementia that could be deployed as a first round screening tool for clinical follow-up, including neurological examination, neuropsychological testing, imaging, and recruitment to clinical trials.	De-identified administrative claims and EHR data from Optum Labs Data Warehouse, Patients are 45 years or older, study dates 1/1/2007-12/31/2017. N=>12 million	Age, gender, utilization (number of encounters), diagnoses (ICD-9 & 10), procedures (CPT-4), medication (from pharmacy fills and prescriptions written) in the sequence in which they occurred, and Episode Treatment Groups.	Algorithm derived from prior paper focused on administrative claims* Variables in algorithm include ICD codes (inpatient or outpatient), pharmacy claims, diagnosis or procedure codes for scans, failed cognitive test *Nori VS, et al. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLOS ONE. 2019; July 5:14(7):e020346.	AUC of 94.4% and F1 score of 54.1%	
Nori VS, et al.[27]	To find patients at high risk of developing ADRD for trial recruitment, clinical care, and identification of risk and protective factors.	De-identified administrative claims and EHR data from Optum Labs Data Warehouse, Patients are 45 years or older, study dates 1/1/2007-12/31/2017 Cohort (n=121,907 and controls n=5,307,045)	Various ICD, CPT, and Episode Treatment Group codes	5 rules to decide which patients are cases (1) Confirmed diagnosis (based on presence of ICD codes), (2) a single prescription from a list of cholinergic agents, (3) a combination of memantine prescription and a subsequent ICD code, (4) a brain scan with a compatible diagnosis on one claim or same date in EHR data followed by an ICD code (5) MMSE<=23 or Minicog<=4 and MOCA<=25 as identified by NLP	sensitivity of 47% and area-under-the-curve of 87%	
Shao Y, et al.[28]	This study sought to identify VA patients over age 65 years with undiagnosed	1861 cases (based on dementia ICD codes assigned by a specialty clinic to VA	ICD or CPT codes, medications, clinical document types, clinical document	Two dementia specialists reviewed 22,980 clinic notes from 120 controls to	AUC 0.91. Choosing a binary cutoff to optimize	

(continued)

Table 2. (continued)

Article	Use of algorithm	Studied population (inclusion/exclusion, setting, and sample size)	Data elements from EHR to create phenotype (predictors)	Reference standard	Performance	Notes
	dementia by analyzing both structured and unstructured EHR data.	patients $\geq$ 65 years old) and 9305 controls (same age, but without dementia diagnoses or medications) matched 5:1 based on gender, age, and Charlson comorbidity index.	text from the VA EHR during the 3-year period immediately preceding the first ICD diagnosis of dementia (or a random visit date for controls).	determine whether each control demonstrated signs and symptoms that were consistent with "Dementia" or "Non-Dementia."	sensitivity and specificity yields 0.83 for both.	
Tjandra D, et al.[44]	The authors sought to develop and validate a model to show that blood pressure trajectories over time could be used to improve prediction of Alzheimer's disease.	Training data came from 6860 patients in an EHR instance serving 5 hospitals in one Veterans Integrated Service Network. Validation data came from 1201 patients in an EHR instance serving an academic medical center. Patients were required to have at least 35 blood pressure measurements.	Predictors included patient age, sex, race, weight, demographics, and longitudinal vital sign measurements.	Diagnosis based on ICD codes.	Area under the receiver operating characteristic curve was 0.64 [95% confidence interval (CI) 0.54–0.73] in the development dataset vs 0.66 (95% CI 0.55–0.76) in the validation dataset.	
Zheng NS, et al.[42] Wan NC, et al.[45]	Zheng et al describes automated development of phenotyping algorithms. The authors reported on an example algorithm targeting identification of patients with dementia. Wan et al improved on the performance of the initial algorithm.	Electronic health record data from 84,821 adult patients in a deidentified DNA biobank.	Diagnosis and procedure codes, SNOMED CT concepts, laboratory and medication data	Clinician-validated algorithms from the Electronic Medical Records and Genomics (eMERGE) network, which used ICD-9 codes, medications, and visit history	Zheng et al reported area under the curve 0.867, accuracy 0.983, recall 0.552, positive predictive value 1.000, negative predictive value 0.983. Wan et al reported improvements across several algorithms, but did not report these metrics.	

clinician, one compared performance to another clinician validated algorithm, and 8 were only tested using other data elements in the EHR without any external validation. Approaches to finding undiagnosed ADRD in these algorithms included identifying cognitive symptomology using NLP,[28,32] using a broad scope of ICD's that include cognitive impairment,[28,31] and calculating a risk model using other variables available in the EHR, such as using demographics, medical diagnoses, vital signs, health care utilization, and medications within the previous 2 years.[29,43,44]

Variation in ICD code use among EHR phenotypes for dementia

Nine out of 20 of the articles overall included ICD codes and provided details of which ICD codes they used in the manuscript or publicly available supplementary material. Table 3 summarizes the ICD codes used in each of the algorithms and compares them to those in the Chronic Conditions Warehouse (CCW) for ADRD. The CCW serves as an appropriate comparison since it is a source for national CMS Medicare and Medicaid research data.[46] Surprisingly, there was substantial variation across algorithms in terms of which ICDs were used.

DISCUSSION

Our review of the literature identified 19 algorithms that use EHR data to identify dementia-related phenotypes; 7 for identifying diagnosed dementia and 12 for identifying individuals with increased risk for dementia. We expect this review to be useful for clinicians and health systems interested in identifying patients with ADRD for research (including pragmatic clinical trials), quality improvement, or enhanced clinical care.

Investigators may favor certain algorithms over others based on the goals of their project and available data. Seven algorithms favored specificity over sensitivity and would be more appropriate in a pragmatic study or quality improvement effort focused on dementia treatment. For these algorithms, we concluded that ICD-based algorithms tended to have the highest specificity. However, given known challenges in timely diagnosis of ADRD, ICD documentation may lag diagnosable disease.[47] Future research should explore how refining these algorithms can improve their sensitivity and specificity and how these vary according to patient population, setting, and data availability.

The other 12 algorithms are suited for identifying patients at high risk for dementia since the goal was to create a sensitive, rather than specific, algorithm. These studies may be relevant for improving early diagnosis of high risk populations, investigating risk factors for ADRD, or implementation of early interventions to improve care.

Compared to the initial review that considered this topic for a range of clinical conditions,[25] we also found several types of variation across phenotypes driven by differences in patient populations, provider organizations, EHRs, and culture, incentives, and regulations that affect what data is entered into the EHR. As with nearly all EHR data, it is biased in that it only consists of patients with utilization. To the extent that ADRD algorithms require specific ICDs or other data that is common only in highly specialized care settings that may be difficult to access, the extent of bias is likely increased.[48]

We found a wide range of reference standards, ranging from clinical assessment, manual chart review, and comparisons to internal reference standards within the EHR among the manuscripts included in the final tables. Many of the algorithms were validated using other data within the EHR, however, this approach has limitations. Diagnosis codes such as ICD codes are often used in these studies as a definitive diagnosis, but this assumes billing practices accurately reflect diagnostic rigor or care provided, which is often not the case.[49] Further, accuracy of billing codes may vary depending on payment structure or care continuity. For example, a health system that provides fee-for-service billing may have billing codes that differently reflect care received as compared to a capitated payment system.[50] Incentives to use dementia billing codes to increase reimbursement may also vary across health systems. Further, a patient that is not continuously followed within a health system will be less likely to be identified, even if they are known to have dementia, simply because there are less data points available.

Another key point is that not all types of billing codes have similar accuracy. For example, a billing code for dementia may be required to order an imaging study to conduct an evaluation for the disease, even though the study results do not indicate dementia. In contrast, an ICD code that is reflected on a patient's problem list is likely to be more accurate. To improve the discrimination of EHR-based algorithms that utilize ICD data, future research should consider ICD provenance (eg, problem list, physician billing, or facility billing) and insurance type.

As with nearly any diagnosis, we found that, in general, requiring more ICD codes improved specificity, but reduced sensitivity. We also conclude that although NLP approaches may improve sensitivity, patients identified by algorithms using this approach may require further screening depending on the organizational context and clinical care delivery culture and these algorithms may require adaptation and further validation when applied to different healthcare settings.

Our review identifies the array of studied approaches for defining an EHR phenotype for ADRD and considers which approaches may work best in different scenarios. EHR infrastructure and culture of care is different across health systems. It may be necessary to adapt and validate a common approach at each site to take into account varying documentation practices, data structures, patient populations, and local practices. Future research should describe how various project scenarios led to the choice of a certain algorithm, and to what extent these algorithms aided in the achievement of project goals. Future work should expand on work done to date for multisite validation of AD/ADRD EHR phenotypes to improve accuracy and generalizability.[51,52] Understanding how the EHR could be leveraged to further classify the severity of dementia is an additional important focus area.[53–55]

EHRs are not static and active engagement with health system can lead to more innovative research and better EHR phenotypes that are more generalizable.[42,56] Health systems may benefit from EHR modifications that allow consistent documentation of cognitive screening results or dementia severity so that this information is readily available as structured data and thereby available to inform opportunities for clinical improvements and research opportunities for patients

Table 3. ICD dementia codes used in EHR phenotype algorithms that are available in the manuscripts of publically available supplements compared to ICD dementia codes used in the CMS chronic condition warehouse for ADRD

		Chronic condition warehouse for ADRD	Amra	Boustani	Ernecoff	Hane	Harding	Nori 2019	Nori 2020	Reuben	Wei
ICD-9											
V40.31	Wandering in diseases classified elsewhere					X					
46.19	Creutzfeldt-Jakob disease, unspecified				X						
290	Dementias				X		X	X	X		
290.0	Senile dementia uncomplicated	X	X	X			X	X	X	X	
290.1	Presenile dementia				X		X	X	X	X	
290.10	Presenile dementia uncomplicated	X	X	X			X	X	X	X	
290.11	Presenile dementia with delirium	X	X				X	X	X	X	
290.12	Presenile dementia with delusional features	X	X	X			X	X	X	X	
290.13	Presenile dementia with depressive features	X	X	X			X	X	X	X	
290.2	Senile dementia with delusional or depressive features				X		X	X	X	X	
290.20	Senile dementia with delusional features	X	X	X			X	X	X	X	
290.21	Senile dementia with depressive features	X	X	X			X	X	X	X	
290.3	Senile dementia with delirium	X	X	X	X		X	X	X	X	
290.4	Vascular dementia				X		X	X	X	X	
290.40	Vascular dementia, uncomplicated	X	X	X			X	X	X	X	
290.41	Vascular dementia, with delirium	X	X	X	X		X	X	X	X	
290.42	Vascular dementia, with delusions	X	X	X	X		X	X	X	X	
290.43	Vascular dementia, with depressed mood	X	X	X			X	X	X	X	
290.8	Other unspecified senile psychotic conditions				X		X			X	
290.9	Unspecified senile psychotic condition			X	X	X	X			X	
291.1	Alcohol-induced persisting amnestic disorder				X	X				X	
291.2	Alcohol-induced persisting dementia		X	X	X					X	
292.82	Drug-induced persisting dementia		X							X	
293.1	Subacute delirium					X					
293.81	Psychotic disorder with delusions in conditions classified elsewhere					X					
294.0	Amnestic disorder in conditions classified elsewhere	X		X				X	X	X	
294.1	Dementia in conditions classified elsewhere			X	X		X	X	X		
294.10	Dementia in conditions classified elsewhere without behavioral disturbance	X	X	X			X	X	X	X	
294.11	Dementia in conditions classified elsewhere with behavioral disturbance	X	X	X	X		X	X	X	X	
294.2	Dementia, unspecified				X		X	X	X		
294.20	Dementia, unspecified, without behavioral disturbance	X	X				X	X	X	X	
294.21	Dementia, unspecified, with behavioral disturbance	X	X	X	X		X	X	X		
294.8	Other persistent mental disorders due to conditions classified elsewhere	X		X	X			X	X		
294.80	Atypical or mixed organic brain syndrome							X	X		
294.9	Unspecified persistent mental disorders due to conditions classified elsewhere			X		X					
298.2	Reactive confusion					X					
300.12	Dissociative amnesia					X					
300.16	Factitious disorder with predominantly psychological signs and symptoms					X					
300.9	Unspecified nonpsychotic mental disorder				X						
310.0	Frontal lobe syndrome			X	X	X					

(continued)

Table 3. (continued)

		Chronic condition warehouse for ADRD	Amra	Boustani	Ernecoff	Hane	Harding	Nori 2019	Nori 2020	Reuben	Wei
310.1	Personality change due to conditions classified elsewhere					X					
310.8	Other specified nonpsychotic mental disorders following organic brain damage					X					
310.89	Other specified nonpsychotic mental disorders following organic brain damage					X					
310.9	Unspecified nonpsychotic mental disorder following organic brain damage					X					
327.41	Confusional arousals					X					
331	Other cerebral degenerations				X						
331.0	Alzheimer's disease	X		X			X	X	X	X	X
331.1	Frontotemporal dementia		X	X	X		X	X	X		
331.11	Pick's disease	X		X			X	X	X	X	
331.19	Other frontotemporal dementia	X	X	X			X	X	X	X	
331.2	Senile degeneration of brain	X		X	X		X	X	X		
331.6	Corticobasal degeneration			X							
331.7	Cerebral degeneration in diseases classified elsewhere	X		X							
331.82	Dementia with Lewy bodies		X	X	X		X	X	X	X	
331.83	Mild cognitive impairment, so stated		X		X		X	X	X		
331.89	Other cerebral degeneration							X			
331.9	Cerebral degeneration unspecified			X							
333.0	Other degenerative diseases of the basal ganglia			X							
333.4	Huntington's chorea				X						
434.91	Cerebral artery occlusion unspecified without cerebral infarction				X						
436	Acute but ill-defined cerebrovascular disease				X						
437.1	Other generalized ischemic cerebrovascular disease				X						
437.8	Other ill-defined cerebrovascular disease				X						
437.9	Unspecified cerebrovascular disease				X						
438	Late effects of cerebrovascular disease				X						
438.0	Cognitive deficits							X			
438.89	Other late effects of cerebrovascular disease				X						
780.9	Other general symptoms					X					
780.93	Memory loss				X			X	X		
797	Senility without psychosis	X		X		X		X	X		
799.52	Cognitive communication deficit							X			
799.53	Visuospatial deficit					X					
799.55	Frontal lobe and executive function deficit					X					
799.59	Other signs and symptoms involving cognition				X	X		X			
ICD-10											
A81.00	Creutzfeldt-Jakob disease, unspecified				X						
F01.50	Vascular dementia without behavioral disturbance	X	X	X	X		X	X	X		
F01.51	Vascular dementia with behavioral disturbance	X	X	X	X		X	X	X		
F02.80	Dementia in other diseases classified elsewhere without behavioral disturbance	X	X	X	X		X	X	X		
F02.81	Dementia in other diseases classified elsewhere with behavioral disturbance	X	X	X	X		X	X	X		
F03.90	Unspecified dementia without behavioral disturbance	X	X	X	X		X	X	X		
F03.91	Unspecified dementia with behavioral disturbance	X	X	X	X		X	X	X		

(continued)

Table 3. (continued)

		Chronic condition warehouse for ADRD	Amra	Boustani	Ernecoff	Hane	Harding	Nori 2019	Nori 2020	Reuben	Wei
F04	Amnestic disorder due to known physiological condition	X		X				X	X		
F05	Delirium due to known physiological condition	X		X							
F06.0	Psychotic disorder with hallucinations due to known physiological condition			X							
F06.1	Catatonic disorder due to known physiological condition	X						X	X		
F06.8	Other specified mental disorders due to known physiological condition	X		X							
F07.0	Personality change due to known physiological condition			X							
F10.27	Alcohol dependence with alcohol-induced persisting dementia			X	X		X				
F10.97	Alcohol use, unspecified with alcohol-induced persisting dementia		X		X		X				
F13.27	Sedative, hypnotic, or anxiolytic dependence with sedative, hypnotic, or anxiolytic-induced persisting dementia						X				
F13.97	Sedative, hypnotic, or anxiolytic use, unspecified with sedative, hypnotic, or anxiolytic-induced persisting dementia						X				
F18.17	Inhalant abuse with inhalant-induced dementia						X				
F18.27	Inhalant dependence with unspecified inhalant-induced dementia						X				
F18.97	Inhalant use, unspecified with inhalant-induced persisting dementia						X				
F19.17	Other psychoactive substance abuse with psychoactive substance-induced persisting dementia						X				
F19.27	Other psychoactive substance dependence with psychoactive substance-induced persisting dementia						X				
F19.97	Other psychoactive substance use, unspecified with psychoactive substance-induced persisting dementia						X				
F30.8	Other manic episodes							X			
F99	Mental disorder, not otherwise specified				X						
G10	Huntington's disease				X						
G13.8	Systemic atrophy primarily affecting central nervous system in other diseases classified elsewhere	X									
G23.0	Hallervorden-Spatz disease			X							
G23.1	Progressive supranuclear ophthalmoplegia Steele-Richardson-Olszweski			X			X				
G23.2	Striatonigral degeneration			X							
G23.8	Other specified degenerative diseases of basal ganglia			X							
G30.0	Alzheimer's disease with early onset	X		X	X		X	X	X		
G30.1	Alzheimer's disease with late onset	X		X	X		X	X	X		
G30.8	Other Ahlzeimer's disease	X			X		X		X		
G30.9	Alzheimer's disease, unspecified	X		X	X		X	X	X		
G31.01	Pick's disease	X		X	X		X	X	X		
G31.09	Other frontotemporal dementia	X	X	X	X		X	X	X		
G31.1	Senile degeneration of brain, not elsewhere classified	X		X	X			X	X		
G31.2	Degeneration of nervous system due to alcohol	X									
G31.83	Dementia with Lewy bodies		X	X	X		X	X	X		
G31.84	Mild cognitive impairment, so stated		X		X		X	X	X		
G31.85	Corticobasal degeneration			X							
G31.89	Other specified degenerative diseases of nervous system							X			

(continued)

Table 3. (continued)

		Chronic condition warehouse for ADRD	Amra	Boustani	Ernecoff	Hane	Harding	Nori 2019	Nori 2020	Reuben	Wei
G31.9	Degenerative disease of nervous system, unspecified			X							
G46.8	Other vascular syndromes of brain in cerebrovascular diseases				X						
G93.1	Anoxic brain damage, not elsewhere classified				X						
G94	Other disorders of brain in diseases classified elsewhere	X		X							
I60.XX	Nontraumatic subarachnoid hemorrhage						X				
I61.X	Nontraumatic intracerebral hemorrhage						X				
I61.9	Nontraumatic intracerebral hemorrhage, unspecified				X		X				
I62.XX	Other and unspecified nontraumatic intracranial hemorrhage						X				
I63.XX	Cerebral infarction						X				
I63.511	Cerebral infarction due to unspecified occlusion or stenosis of right middle cerebral artery				X		X				
I63.521	Cerebral infarction due to unspecified occlusion or stenosis of right anterior cerebral artery				X		X				
I63.9	Cerebral infarction, unspecified				X		X				
I65.XX	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction						X				
I65.29	Occlusion and stenosis of other precerebral arteries				X		X				
I66.XX	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction						X				
I67.XX	Other cerebrovascular diseases						X				
I67.89	Other cerebrovascular disease				X		X				
I67.9	Cerebrovascular disease, unspecified				X		X				
I68.X	Cerebrovascular disorders in diseases classified elsewhere						X				
I68.8	Other cerebrovascular disorders in diseases classified elsewhere				X		X				
I69.XX	Sequelae of cerebrovascular disease						X				
I69.80	Unspecified sequelae of other cerebrovascular disease				X		X				
I69.81	Cognitive deficits following other cerebrovascular disease				X		X				
I69.91	Cognitive deficits following unspecified cerebrovascular disease				X		X	X			
R41.1	Anterograde amnesia							X	X		
R41.2	Retrograde amnesia							X	X		
R41.3	Other amnesia							X	X		
R41.81	Age-related cognitive decline	X		X				X			
R41.841	Cognitive communication deficit							X			
R41.844	Frontal lobe an executive function deficit				X						
R41.89	Other symptoms and signs involving cognitive functions and awareness							X			
R54	Age-related physical debility	X									
S06.9X0A	Unspecified intracranial injury without loss of consciousness, initial encounter				X						
S06.9X0D	Unspecified intracranial injury without loss of consciousness, subsequent encounter				X						
S06.9X0S	Unspecified intracranial injury without loss of consciousness, sequela				X						
S09.90XA	Unspecified injury of head, initial encounter				X						
S09.90XD	Unspecified injury of head, subsequent encounter				X						
S09.90XS	Unspecified injury of head, sequela				X						

with dementia. Closer relationships between payers and providers, such as in accountable care organizations, may also make real-time linkages to other data sources such as Medicare, MDS, or OASIS data more feasible which could also lead to more robust phenotypes for dementia.

FUNDING

This work was funded by the National Institute on Aging (NIA) of the National Institutes of Health under Award 5U54AG063546, which funds NIA Embedded Pragmatic Alzheimer's and AD-Related Dementias Clinical Trials Collaboratory (NIA IMPACT Collaboratory). This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Drs Ritchie, Bennett, and Vydiswaran are supported by the IMPACT Collaboratory 5U54AG063546.

AUTHOR CONTRIBUTIONS

CSR, JP, and AMW conceived of the study and all authors made substantial contributions to the acquisition, analysis, and interpretation of the data for the work. AMW and JP drafted the work and all authors participated in revising it critically for important intellectual content and all authors approved of the final submitted version. All authors agree to be accountable for all aspects of the work. CSR supervised the project.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Katherine Santos for her administrative work on this paper.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

DATA AVAILABILITY

Our study relied exclusively on analysis of publically available data.

REFERENCES

1. CDC. CDC's Alzheimer's Disease and Healthy Aging Program. Centers for Disease Control and Prevention. https://www.cdc.gov/aging/index.html. Published April 25, 2022. Accessed July 21, 2022.
2. Alzheimer's Association. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement J Alzheimers Assoc*. 2023; 19 (4): 1598–695.
3. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* 2013; 80 (19): 1778–83.
4. Murphy SL, Xu J, Kochanek KD, Arias E. Mortality in the United States, 2017. NCHS Data Brief 2018; 328: 1–8.
5. James BD, Leurgans SE, Hebert LE, Scherr PA, Yaffe K, Bennett DA. Contribution of Alzheimer disease to mortality in the United States. *Neurology* 2014; 82 (12): 1045–50.
6. Helzner EP, Scarmeas N, Cosentino S, Tang MX, Schupf N, Stern Y. Survival in Alzheimer disease: a multiethnic, population-based study of incident cases. *Neurology* 2008; 71 (19): 1489–95.
7. Larson EB, Shadlen MF, Wang L, *et al.* Survival after initial diagnosis of Alzheimer disease. *Ann Intern Med* 2004; 140 (7): 501–9.
8. Wolfson C, Wolfson DB, Asgharian M, *et al.*; Clinical Progression of Dementia Study Group. A reevaluation of the duration of survival after the onset of dementia. *N Engl J Med* 2001; 344 (15): 1111–6.
9. Rountree SD, Chan W, Pavlik VN, Darby EJ, Doody RS. Factors that influence survival in a probable Alzheimer disease cohort. *Alzheimers Res Ther* 2012; 4 (3): 16.
10. Hollister BA, Yeh J, Ross L, Schlesinger J, Cherry D. Building an advocacy model to improve the dementia-capability of health plans in California. *J Am Geriatr Soc* 2021; 69 (12): 3641–9.
11. Butler M, Gaugler JE, Talley KMC, *et al. Care Interventions for People Living With Dementia and Their Caregivers*. Comparative Effectiveness Review No. 231. AHRQ Publication No. 20-EHC023. Rockville, MD: Agency for Healthcare Research and Quality; 2020. doi:10.23970/AHRQEPCCER231.
12. Teno JM, Keohane LM, Mitchell SL, *et al.* Dying with dementia in medicare advantage, accountable care organizations, or traditional medicare. *J Am Geriatr Soc* 2021; 69 (10): 2802–10.
13. Hua CL, Thomas KS, Bunker JN, *et al.* Dementia diagnosis in the hospital and outcomes among patients with advanced dementia documented in the Minimum Data Set. *J Am Geriatr Soc* 2022; 70 (3): 846–53.
14. Mitchell SL, Teno JM, Kiely DK, *et al.* The clinical course of advanced dementia. *N Engl J Med* 2009; 361 (16): 1529–38.
15. Kelley AS, McGarry K, Gorges R, Skinner JS. The burden of health care costs for patients with dementia in the last 5 years of life. *Ann Intern Med* 2015; 163 (10): 729–36.
16. McCarthy EP, Chang CH, Tilton N, Kabeto MU, Langa KM, Bynum JPW. Validation of claims algorithms to identify Alzheimer's disease and related dementias. *J Gerontol A Biol Sci Med Sci* 2022; 77 (6): 1261–71.
17. Electronic Health Records-Based Phenotyping | Rethinking Clinical Trials®. https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/. Accessed July 21, 2022.
18. Richesson RL, Green BB, Laws R, *et al.* Pragmatic (trial) informatics: a perspective from the NIH Health Care Systems Research Collaboratory. *J Am Med Inform Assoc* 2017; 24 (5): 996–1001.
19. Richesson RL, Hammond WE, Nahm M, *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* 2013; 20 (e2): e226–31.
20. Bynum JPW, Dorr DA, Lima J, *et al.* Using healthcare data in embedded pragmatic clinical trials among people living with dementia and their caregivers: state of the art. *J Am Geriatr Soc* 2020; 68 (Suppl 2): S49–54.
21. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005; 8 (1): 19–32.
22. PRISMA-ScR Item 4: Objectives. Knowledge Translation Program. Toronto: St. Michael's Hospital, Unity Health Toronto. https://knowledgetranslation.net/wp-content/uploads/2019/05/PRISMA-ScR_TipSheet_Item4.pdf.
23. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: Scoping reviews (2020 version). In: Aromataris E, Munn Z, eds. *JBI Manual for Evidence Synthesis*. JBI; 2020. doi: 10.46658/JBIMES-20-12.
24. Thomas D. BeckerGuides: Scoping Review Guide: PCC Question Outline. https://beckerguides.wustl.edu/scopingreviews/PCC. Accessed April 26, 2023.
25. Lee S, Doktorchik C, Martin EA, *et al.* Electronic medical record-based case phenotyping for the Charlson conditions: scoping review. *JMIR Med Inform* 2021; 9 (2): e23934.
26. Singh B, Singh A, Ahmed A, *et al.* Derivation and validation of automated electronic search strategies to extract Charlson comorbidities from electronic medical records. *Mayo Clin Proc* 2012; 87 (9): 817–24.
27. Nori VS, Hane CA, Crown WH, *et al.* Machine learning models to predict onset of dementia: a label learning approach. *Alzheimers Dement (N Y)* 2019; 5: 918–25.

28. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med Inform Decis Mak* 2019; 19 (1): 128.
29. Ben Miled Z, Haas K, Black CM, *et al.* Predicting dementia with routine care EMR data. *Artif Intell Med* 2020; 102: 101771.
30. Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011; 3 (79): 79re1.
31. Amra S, O'Horo JC, Singh TD, *et al.* Derivation and validation of the automated search algorithms to identify cognitive impairment and dementia in electronic health records. *J Crit Care* 2017; 37: 202–5.
32. Harding BN, Floyd JS, Scherrer JF, *et al.* Methods to identify dementia in the electronic health record: comparing cognitive test scores with dementia algorithms. *Healthc (Amst)* 2020; 8 (2): 100430.
33. Reuben DB, Hackbarth AS, Wenger NS, Tan ZS, Jennings LA. An automated approach to identifying patients with dementia using electronic medical records. *J Am Geriatr Soc* 2017; 65 (3): 658–9.
34. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016; 23 (e1): e20–7.
35. 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 Syst* 2020; 4 (4): e10246.
36. 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): 103–11.
37. Boustani M, Perkins AJ, Khandker RK, *et al.* Passive digital signature for early identification of Alzheimer's disease and related dementia. *J Am Geriatr Soc* 2020; 68 (3): 511–8.
38. Ernecoff NC, Wessell KL, Gabriel S, Carey TS, Hanson LC. A novel screening method to identify late-stage dementia patients for palliative care research and practice. *J Pain Symptom Manage* 2018; 55 (4): 1152–8.e1.
39. Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P. Predicting onset of dementia using clinical notes and machine learning: case-control study. *JMIR Med Inform* 2020; 8 (6): e17819.
40. McCoy TH, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. *Alzheimers Dement* 2020; 16 (3): 531–40.
41. Nori VS, Hane CA, Sun Y, Crown WH, Bleicher PA. Deep neural network models for identifying incident dementia using claims and EHR datasets. *PLoS One* 2020; 15 (9): e0236400.
42. Zheng NS, Feng Q, Kerchberger VE, *et al.* PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J Am Med Inform Assoc* 2020; 27 (11): 1675–87.
43. 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 J Alzheimers Assoc* Published online February 23, 2023; doi:10.1002/alz.12967
44. Tjandra D, Migrino RQ, Giordani B, Wiens J. Use of blood pressure measurements extracted from the electronic health record in predicting Alzheimer's disease: a retrospective cohort study at two medical centers. *Alzheimers Dement* 2022; 18 (11): 2368–72.
45. Wan NC, Yaqoob AA, Ong HH, Zhao J, Wei WQ. Evaluating resources composing the PheMAP knowledge base to enhance high-throughput phenotyping. *J Am Med Inform Assoc* 2023; 30 (3): 456–65.
46. Centers for Medicare & Medicaid Services. Chronic Conditions Data Warehouse. Chronic Conditions Data Warehouse. https://www2.ccwdata.org. Accessed February 13, 2023.
47. Dubois B, Padovani A, Scheltens P, Rossi A, Dell'Agnello G. Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges. *J Alzheimers Dis* 2016; 49 (3): 617–31.
48. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. *J Am Med Inform Assoc* 2023; 30 (5): 971–7.
49. Haut ER, Pronovost PJ, Schneider EB. Limitations of administrative databases. *JAMA* 2012; 307 (24): 2589–90.
50. Behrendt CA, Debus ES, Mani K, Sedrakyan A. The strengths and limitations of claims based research in countries with fee for service reimbursement. *Eur J Vasc Endovasc Surg* 2018; 56 (5): 615–6.
51. Coley RY, Smith JJ, Karliner L, *et al.* External validation of the eRADAR risk score for detecting undiagnosed dementia in two real-world healthcare systems. *J Gen Intern Med* 2023; 38 (2): 351–60.
52. John LH, Kors JA, Fridgeirsson EA, Reps JM, Rijnbeek PR. External validation of existing dementia prediction models on observational health data. *BMC Med Res Methodol* 2022; 22 (1): 311.
53. Wang L, Sha L, Lakin JR, *et al.* Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. *JAMA Netw Open* 2019; 2 (7): e196972.
54. Shankle WR, Mania S, Dick MB, Pazzani MJ. Simple models for estimating dementia severity using machine learning. *Stud Health Technol Inform* 1998; 52(Pt 1): 472–6.
55. Kharrazi H, Anzaldi LJ, Hernandez L, *et al.* The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc* 2018; 66 (8): 1499–507.
56. Greiver M, Wintemute K, Aliarzadeh B, *et al.* Implementation of data management and effect on chronic disease coding in a primary care organisation: a parallel cohort observational study. *J Innov Health Inform* 2016; 23 (3): 843.