Passive Digital Signature for Early Identification of Alzheimer's Disease and Related Dementia

Malaz Boustani, MD, MPH,[*†‡] Anthony J. Perkins, MS,[§] Rezaul Karim Khandker, PhD, MBA,[¶] Stephen Duong, MA, MS,[‖] Paul R. Dexter, MD,[‡] Richard Lipton, MD,[**] Christopher M. Black, PhD, MPH,[¶] Vasu Chandrasekaran, PhD,[††] Craig A. Solid, PhD,[‡‡] and Patrick Monahan, PhD[§§]*

OBJECTIVES: Developing scalable strategies for the early identification of Alzheimer's disease and related dementia (ADRD) is important. We aimed to develop a passive digital signature for early identification of ADRD using electronic medical record (EMR) data.

DESIGN: A case-control study.

SETTING: The Indiana Network for Patient Care (INPC), a regional health information exchange in Indiana.

PARTICIPANTS: Patients identified with ADRD and matched controls.

MEASUREMENTS: We used data from the INPC that includes structured and unstructured (visit notes, progress notes, medication notes) EMR 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. We constructed models to identify early 1- to 10-year, 3- to 10-year, and 5- to 10-year ADRD signatures. The analyses included 14 diagnostic risk variables and 10 drug classes in addition to new variables produced from unstructured data (eg, disorientation, confusion, wandering, apraxia, etc). The area under the receiver operating characteristics (AUROC) curve was used to determine the best models.

RESULTS: The AUROC curves for the validation samples for the 1- to 10-year, 3- to 10-year, and 5- to 10-year models that used only structured data were .689, .649, and .633, respectively. For the same samples and years, models that used both structured and unstructured data produced AUROC curves of .798, .748, and .704, respectively. Using a cutoff to maximize sensitivity and specificity, the 1- to 10-year, 3- to 10-year, and 5- to 10-year models had sensitivity that ranged from 51% to 62% and specificity that ranged from 80% to 89%.

CONCLUSION: EMR-based data provide a targeted and scalable process for early identification of risk of ADRD as an alternative to traditional population screening. **J Am Geriatr Soc 68:511-518, 2020.**

Key words: Alzheimer's disease; dementia; risk factors

The National Plan to Address Alzheimer's Disease of the US Department of Health and Human Services and the Affordable Care Act through the Medicare Annual Wellness visit identify earlier detection of Alzheimer's disease and related dementias (ADRD) as a core aim for improving care quality for older adults.[1,2] Furthermore, if the development of disease-modifying therapeutics for ADRD is successful, it may require the use of such therapeutics at a very early stage of the disease.

Previous approaches to early identification included the use of brief cognitive screening tests (eg, the Mini-Mental State Examination,[3] Telephone Interview for Cognitive Status,[4] Montreal Cognitive Assessment Scale,[5] the Memory Impairment Screen [MIS],[6] and MIS by telephone[7]) and the use of biological markers. However, screening every

From the *Indiana University Center for Health Innovation and Implementation Science, Indiana Clinical Translational Science Institute, Indianapolis, Indiana; †Sandra Eskenazi Center for Brain Care Innovation, Eskenazi Health, Indianapolis, Indiana; ‡Indiana University Center for Aging Research, Regenstrief Institute, Inc, Indianapolis, Indiana; §Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana; ¶Center for Observational and Real-World Evidence, Merck & Co., Inc., Kenilworth, Indiana; ‖The Gerontology Institute, Georgia State University, Atlanta, Georgia; **Department of Neurology, Albert Einstein College of Medicine, Bronx, New York; ††Center for Observational and Real-World Evidence, Merck & Co., Inc., Boston, Massachusetts; ‡‡Solid Research Group, LLC, Saint Paul, Minnesota; and the §§Department of Biostatistics, Indiana University School of Medicine and School of Public Health, Indianapolis, Indiana.

Address correspondence to Patrick Monahan, PhD, Indiana University School of Medicine and School of Public Health, 410 W. 10th Street, Suite 3000, Indianapolis, IN 46202. E-mail: pmonahan@iu.edu

DOI: 10.1111/jgs.16218

0002-8614/20/$15.00

American for ADRD would be expensive and may be unfeasible because significant numbers of older patients refuse brief cognitive screening tests,[8] especially if they do not feel they have memory problems and do not know anyone with ADRD.[8,9] Additionally, there are uncertainties in the area of early detection; in 2014 the US Preventive Services Task Force concluded that it was unclear whether the benefits of early detection outweighed potential harms of false positives and side effects of pharmacologic interventions.[10] However, previous research showed that patient characteristics, such as age, sex, race, education, cardiovascular risk factors, family history, comorbidity burden, and lifestyle factors, correlate with ADRD screening results.[9,11-13] The continuous growth in the amount of electronic medical record (EMR) data captured by health information exchanges across the United States offers an opportunity to develop a targeted and a scalable process that leverages these data to identify various early, passive, and digital ADRD signatures (prodromal ADRD) as an alternative to traditional population screening.

Therefore, we aimed to develop a multivariable, validated, and generalizable model using structured and unstructured EMR data to identify early prodromal ADRD signatures up to 5 years in advance. These signatures may allow feasible and scalable segmentation of patient populations including a high-risk group that could be targeted for further cognitive or biological screening tests.

METHODS

Data Sources and Patient Identification

The data for this study were sourced from the Indiana Network for Patient Care (INPC). The INPC is the oldest and largest health information exchange in the United States, comprising 38 healthcare systems and serving more than 13 million patients across the entire state of Indiana. In addition to patient demographics (age, sex, race), the INPC database contains EMR-based information on comorbid diagnoses, medications, imaging, and healthcare utilization, as well as unstructured medical encounter notes.

This study used a case-control methodology nested within the INPC where cases and controls were matched on nonmodifiable risk factors (age, race, and sex). Cases included adult patients (aged ≥18 y) with at least two clinical visits with ADRD diagnostic codes (Table S1) between January 1, 2008, and December 31, 2016. The choice to include patients as young as 18 years old was intentional. We wanted to capture patients with early-onset ADRD (even though the prevalence is small) as well as an emerging group whose risk for long-term cognitive impairment increases as a result of a critical illness, as identified by the BRAIN-ICU study.[14] The first clinical visit was used as the index date, that is, the date at which a patient was first diagnosed with ADRD.

To be eligible, study cases were required to be active in the INPC by having at least one medical encounter every other year for 10 years leading up to their first ADRD-coded visit. Patients with any ADRD-coded clinical visit in their EMRs during the 10 years before the index date were excluded because these patients were considered to have prevalent ADRD. Controls were then matched 4:1 with the sample of incident ADRD cases based on sex, race, index year, and age within 5 years. Controls were required to be active in the INPC and to lack a previous ADRD-coded clinic visit in their EMR. After the full study sample was identified, it was randomly split into two data sets: a derivation sample (70%) and a validation sample (30%). Where possible, we attempted to meet the standards described by the Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.[15,16]

Variables of Interest

Using the structured EMR data available from the INPC database, we identified a priori a set of diagnostic and generic product identifier drug groups and classes as possibly being related to ADRD based on clinical expertise (M.B., P.R.D. and R.L.). Several of these variables were excluded from the models due to collinearity. We coded diagnoses as present if the diagnostic code occurred during the time frame for a specific patient. Medications were coded by the drug class (eg, anticholinergics, psychotropics, proton pump inhibitors [PPIs], etc) and identified as either being current (prescription during the index year), prior (history of prescription in the EMR but no prescription during the index year), or never (no current or prior history).

Unstructured EMR data from free text fields included patient visit notes, progress notes, medication notes, and so on. To utilize these data, we used 48 predefined terms (Table S2) from our team of clinical experts (M.B., P.R.D. and R.L.) and searched through all encounter notes in the EMR using software (nDepth) that incorporates natural language processing capabilities and advanced text mining. The nDepth algorithm includes negation for a specific keyword (negated or affirmed) and specifies the appropriate experiencer (patient or other) as well as temporality of occurrence (recent or historical). This allowed us to use the unstructured data to improve our definition of cases and controls through possible missed diagnoses, to refine accurate variable definitions using prescription and diagnoses information, and to create new variables of interest that are not easily coded with structured data. Results for nDepth included information on whether any of the prespecified search terms were found in any form of notes. Data were aggregated across all notes to create an indicator variable for each term of interest for the specified time period.

Analytic Methods

We constructed three sets of logistic regression models using data from three different time periods before the index date. These time periods were more than one but less than 10 years before the index date (1- to 10-y model), more than 3 but less than 10 years prior (3- to 10-y model), and more than 5 but less than 10 years prior (5- to 10-y model). Data for the year before the index date were not used in the modeling because many of the orders during that time could be directly related to establishing an ADRD diagnosis (eg, computed tomography [CT] scan or serum thyroid-stimulating hormone test to confirm suspicion of ADRD). For each time period, we constructed two sets of models: the first set used only structured EMR data (diagnostic and medication variables), and the second set of models used additional information obtained from the unstructured EMR data. The goal of leveraging the

unstructured data was to include patient information on disease burden and conditions not available in the structured data and to define more accurately the presence or absence of conditions that may be indicative of ADRD. The unstructured data were used in three ways.

First, unstructured data were used to refine definitions for chronic disease conditions. We searched for nine conditions (anxiety, cerebrovascular disease, depression, delirium, diabetes, ischemic heart disease, myocardial infarction [MI], stroke, and weight loss); in four the use of text data enhanced identification of these conditions (depression, delirium, diabetes, and MI), and new variables were created that were marked as present if the condition was found in the structured or unstructured data.

Second, a newly created variable was included that reflected the presence in the notes of specific terms not available in the structured data but that might commonly be made for patients with ADRD (disorientation, confusion, MMSE, the presence of a caregiver, memory or cognitive complaints, functional issues, social withdrawal, personality changes, wandering, repetitive behavior, agnosia, apraxia, cortical atrophy, white matter lesions, and family history of ADRD). We explored models using multiple variables for the individual terms but decided to use the "any note mention" variable because it allowed for consistency across years and models and provided a more reliable odds ratio (OR) estimate due to sparseness of some terms.

Third, cases and controls were excluded if there was any mention in the unstructured data of dementia or any cholinesterase inhibitor. This was done to improve the case-control definition by accounting for possible missed diagnoses that could have arisen from using structured data only. All models were adjusted for age, race, and sex. Diagnostic variables that were originally included in the modeling process but were not included in the final model due to nonsignificance or collinearity were diagnoses of angina, anxiety, hypercholesterolemia, insomnia, and anxiety. Other candidate variables not included in the final model due to nonsignificance or multicollinearity were opioid analgesics, any inpatient admission, and any CT or magnetic resonance imaging.

All models were run two ways to compare the discrepancy in missing patient data. First, models were run using all patients, regardless of the completeness of patient data (referred to as "All Patients" models); then "Complete Patients" models were run using only patients who had at least one diagnostic record and at least one pharmacy record during

Table 1. Demographic and Clinical Characteristics for Cases and Controls

	Derivation sample (n = 50 014)		Validation sample (n = 21 452)	
	No dementia (n = 39 510) % (N)	Dementia (n = 10 504) % (N)	No dementia (n = 16 952) % (N)	Dementia (n = 4500) % (N)
Female	64.3 (25 400)	65.3 (6856)	65.1 (11 042)	66.1 (2976)
Race				
African American	20.2 (7972)	21.2 (2231)	19.6 (3323)	20.8 (934)
Hispanic	1.1 (439)	1.3 (136)	1.0 (178)	1.2 (54)
Other	2.5 (993)	2.8 (289)	2.4 (412)	2.6 (117)
White	76.2 (30 106)	74.7 (7848)	76.9 (13 039)	75.4 (3395)
Diabetes	34.6 (13 688)	43.4 (4,563)	34.7 (5877)	44.4 (2000)
Angina	15.1 (5975)	19.0 (2000)	15.1 (2560)	20.0 (898)
MI	5.1 (2003)	7.5 (784)	5.0 (855)	7.5 (339)
Chronic ischemic heart disease	30.5 (12 048)	39.2 (4112)	30.8 (5213)	39.6 (1781)
TIA	5.4 (2147)	10.0 (1045)	5.6 (952)	9.3 (418)
Stroke: cerebral infarction	4.4 (1724)	10.8 (1129)	4.8 (815)	10.7 (480)
Hemorrhagic stroke	.9 (363)	2.5 (264)	.8 (133)	2.3 (105)
Other CVD	15.9 (6284)	30.7 (3225)	16.1 (2724)	29.9 (1346)
Atherosclerosis	12.0 (4749)	20.3 (2130)	12.1 (2056)	20.4 (917)
Hypertensive diseases	71.8 (28 374)	80.0 (8400)	71.6 (12 131)	78.8 (3544)
Hypercholesterolemia	26.8 (10 573)	28.9 (3033)	27.1 (4589)	27.6 (1243)
Depression	24.6 (9725)	36.0 (3785)	24.9 (4222)	35.6 (1602)
Anxiety	6.6 (2597)	8.7 (915)	6.6 (1,126)	8.8 (398)
Bipolar disorder	5.6 (2221)	10.3 (1080)	5.7 (973)	10.0 (451)
Schizophrenia	5.8 (2290)	11.2 (1174)	5.7 (965)	11.1 (498)
Weight loss	8.6 (3399)	15.5 (1626)	8.6 (1458)	14.3 (642)
Insomnia	8.9 (3500)	12.0 (1263)	9.0 (1532)	12.2 (548)
Sleep apnea	11.6 (4603)	14.7 (1547)	12.0 (2035)	14.6 (657)
CT	23.7 (9357)	36.4 (3821)	24.0 (4064)	35.5 (1599)
MRI	11.8 (4662)	17.8 (1871)	12.2 (2064)	18.0 (809)
Inpatient admission	70.3 (27 781)	76.9 (8077)	70.7 (11 987)	78.2 (3517)
ED visit	67.9 (26 829)	78.6 (8256)	68.0 (11 519)	77.8 (3501)
Delirium	9.3 (3689)	23.4 (2455)	9.4 (1597)	23.8 (1071)

Abbreviations: CT, computed tomography; CVD, cardiovascular disease; ED, emergency department; MI, myocardial infarction; MRI, magnetic resonance imaging; TIA, transient ischemic attack.

the respective time frame. In the All Patients models, patients without any diagnostic record or without any pharmacy record variables were assumed to be "no" or "0" for the diagnoses and drug classes. Receiver operating characteristics analysis was used to evaluate model performance. In each case, we determined the cutpoints that achieved 80% sensitivity and 80% specificity, as well as the cutpoint that maximized those values. The area under the receiver operating characteristics (AUROC) was used to determine the best models.

RESULTS

Study Sample

The final sample consisted of 15 004 cases and 56 462 controls. Our study population had a mean age of 68.2 years (standard deviation = 16.1); 1.7% were aged 18 to 30 years; 8.7%, 31 to 54 years; 13.1%, 55 to 64 years; and the remaining 76.5% were 65 years or older. The population was 64.7% female and 20.2% African American. The most common diagnoses carried by the population included hypertension (73.4%), diabetes (36.6%), chronic ischemic heart disease (32.4%), and depression (27.1%). Prescriptions most commonly observed in patient histories included those for analgesics-opioids (36.0%), nonsteroidal anti-inflammatory drugs (NSAIDs; 32.6%), and antidepressants (23.2%). The derivation sample consisted of 10 504 cases and 39 510 controls; the validation sample consisted of 4500 cases and 16 952 controls. A descriptive comparison of the derivation and validation samples is shown in Table 1, and medication usage for these samples is shown in Table S3. The risk factor distributions were very similar in the derivation and validation data sets for both controls and cases. For each of the case selection periods (1-10 y, 3-10 y, and 5-10 y), approximately 20% to 40% of patients included in the All Patients models had no pharmacy records or no diagnostic records, largely driven by incomplete pharmacy data, and therefore they were excluded from the Complete Patients models.

Models Using Structured EMR Data Only

Table 2 presents the Discrimination results for the 1- to 10-year, 3- to 10-year, and 5- to 10-year models. Within the models using only structured data, those that included data closer to a patient's index date (1- to 10-y models) performed slightly to moderately better than those that included models using data further from the index date (3- to 10-y and 5- to 10-y models), as reflected by higher AUROC curve values.

Models Using Both Structured and Unstructured EMR Data

As demonstrated by the AUROC curves (Table 2), the models that used unstructured EMR data performed better than the models that used only the structured data. Like the structured-data-only models, the structured and unstructured data models that included data closer to a patient's index date (1- to 10-y models) had higher AUROC curve values and therefore performed better than models using data further from the index date (3- to 10-y and 5- to 10-y models).

Table 3 displays estimated ORs for model variables for the model using structured data only ("structured-data-only model") and the model using both structured and unstructured data ("structured and unstructured data model") for 1 to 10 years prior to allow for examination of how the addition of unstructured EMR data affected estimated

Table 2. Summary Statistics for Models

		Structured data only		Structured and unstructured data	
Sample		**n**	**AUROC curve**	**n**	**AUROC curve**
1- to 10-y model					
Derive	*ALL*	*24 239*	.703	23 101	.814
	COMPLETE	*15 921*	.727	15 055	.834
Validate	*ALL*	*10 349*	.689	9848	.798
	COMPLETE	*6819*	.716	6446	.826
3- to 10-y model					
Derive	*ALL*	*24 239*	.664	23 682	.763
	COMPLETE	*11 238*	.701	10 876	.797
Validate	*ALL*	*10 349*	.649	10 117	.748
	COMPLETE	*4809*	.688	4665	.792
5- to 10-y model					
Derive	*ALL*	*24 239*	.642	23 969	.717
	COMPLETE	*8428*	.695	8266	.769
Validate	*ALL*	*10 349*	.633	10 243	.704
	COMPLETE	*3614*	.686	3545	.764

Abbreviation: AUROC, area under the receiver operating characteristics.

Note: The 1- to 10-year model used data from 1 to 10 years before index date; 3- to 10-year model used data from 3 to 10 years before index date; 5- to 10-year model used data from 5 to 10 years before index date. ALL = all patients, regardless of completeness of data. COMPLETE = only those patients with least one diagnostic record and at least one pharmacy record during the respective time frame.

The structured-data-only models use only structured data. The structured and unstructured data models use text data to refine selected diagnoses, incorporate additional text-based predictor variables, and improve case-control selection by excluding cases (and some controls) with mention of dementia-related diagnoses not detected by the structured data.

Table 3. Logistic Regression Results

	Structured data model		Structured and unstructured data model	
	OR (95% CI)	*P*	OR (95% CI)	*P*
Age	**1.01 (1.00-1.01)**	**<.001**	**1.01 (1.00-1.01)**	**.001**
Female vs male	.99 (.92-1.06)	.779	.98 (.90-1.07)	.667
African American vs white	.97 (.89-1.05)	.411	.90 (.81-1.00)	.043
Hispanic vs white	1.11 (.64-1.94)	.711	1.09 (.57-2.07)	.805
Other race vs white	**1.28 (1.09-1.51)**	**.004**	1.43 (1.17-1.75)	.001
Diabetes	**1.20 (1.11-1.29)**	**<.001**	**1.53 (1.39-1.68)**	**<.001**
MI	1.16 (1.00-1.36)	.056	**4.55 (3.94-5.26)**	**<.001**
Acute ischemic heart disease	**1.60 (1.11-2.30)**	**.011**	.90 (.54-1.49)	.677
Chronic ischemic heart disease	**1.18 (1.09-1.28)**	**<.001**	**.84 (.76-.94)**	**.002**
TIA	**1.30 (1.12-1.51)**	**.001**	1.07 (.87-1.31)	.552
Stroke (cerebral infarction)	**1.41 (1.21-1.64)**	**<.001**	1.07 (.86-1.33)	.545
Hemorrhagic stroke	**1.50 (1.12-2.00)**	**.007**	1.14 (.75-1.74)	.548
Other CVD	**1.61 (1.46-1.77)**	**<.001**	**1.59 (1.40-1.80)**	**<.001**
Atherosclerosis	1.01 (.90-1.12)	.922	1.06 (.92-1.22)	.448
Hypertensive diseases	**1.27 (1.16-1.39)**	**<.001**	1.07 (.96-1.20)	.223
Depression	**1.61 (1.45-1.79)**	**<.001**	**6.19 (5.55-6.91)**	**<.001**
Bipolar disorder	1.42 (1.09-1.85)	.009	.95 (.69-1.30)	.744
Schizophrenia	**2.61 (2.05-3.32)**	**<.001**	**3.13 (2.35-4.18)**	**<.001**
Weight loss	**1.67 (1.48-1.88)**	**<.001**	**1.50 (1.28-1.75)**	**<.001**
Benzodiazepines				
Current vs never	**1.37 (1.18-1.58)**	**<.001**	**1.65 (1.38-1.98)**	**<.001**
Former vs never	1.08 (.94-1.24)	.306	1.06 (.89-1.28)	.505
NSAIDs				
Current vs never	**.82 (.72-.92)**	**.001**	.94 (.80-1.10)	.439
Former vs never	.96 (.87-1.07)	.495	.97 (.85-1.12)	.703
Proton pump inhibitor				
Current vs never	**.78 (.70-.87)**	**<.001**	**.76 (.66-.87)**	**.000**
Former vs never	1.00 (.88-1.12)	.930	.97 (.83-1.14)	.698
H_2-blockers				
Current vs never	1.03 (.86-1.23)	.753	.94 (.75-1.19)	.619
Former vs never	**1.22 (1.06-1.40)**	**.005**	1.12 (.94-1.34)	.215
Psychotherapeutic agents				
Current vs never	**12.04 (9.35-15.50)**	**<.001**	**2.92 (1.89-4.50)**	**<.001**
Former vs never	**2.11 (1.68-2.64)**	**<.001**	**1.47 (1.08-2.01)**	**.015**
Antidepressants				
Current vs never	**1.33 (1.19-1.47)**	**<.001**	**.67 (.58-.77)**	**<.001**
Former vs never	**1.13 (1.00-1.29)**	**.050**	**.69 (.59-.82)**	**<.001**
Anticonvulsants				
Current vs never	**1.45 (1.29-1.64)**	**<.001**	**1.47 (1.25-1.71)**	**<.001**
Former vs never	**1.29 (1.12-1.48)**	**.000**	**1.35 (1.13-1.62)**	**.001**
Minerals and electrolytes				
Current vs never	**1.23 (1.08-1.40)**	**.002**	**1.38 (1.17-1.63)**	**.000**
Former vs never	1.11 (.96-1.28)	.152	1.06 (.88-1.28)	.558
Anti-lipidemics				
Current vs never	**.77 (.71-.84)**	**<.001**	**.72 (.65-.81)**	**<.001**
Former vs never	.93 (.82-1.04)	.205	.96 (.83-1.13)	.643
Analgesics, non-narcotic				
Current vs Never	**5.09 (3.97-6.53)**	**<.001**	**5.84 (4.34-7.86)**	**<.001**
Former vs never	.96 (.83-1.12)	.601	**.75 (.61-.92)**	**.005**
Any ED visit	**.92 (.86-.99)**	**.032**	**.73 (.67-.81)**	**<.001**
Delirium	**2.33 (2.07-2.63)**	**<.001**	**1.32 (1.11-1.56)**	**.002**
Any note mention	NA		**75.95 (55.83-103.32)**	**<.001**

Abbreviations: CI, confidence interval; CVD, cardiovascular disease; ED, emergency department; MI, myocardial infarction; NSAIDs, nonsteroidal anti-inflammatory drugs; OR, odds ratio; TIA, transient ischemic attack.

Note: Any note mention, coded as 1 if (vs 0) if any mention in note of any of the following terms: disorientation, confusion, Mini-Mental State Examination (MMSE), the presence of a caregiver, memory/cognitive complaints, wandering, apraxia/functional issues, atrophy/white matter lesions, social withdrawal, personality change, agnosia, repetitive behavior, and family history of Alzheimer's disease or related dementia (ADRD); NA = not applicable because notes from unstructured data were not used in these models. Values in boldface represent those with a *P* value <.05.

Table 4. Sensitivity and Specificity of Structured and Unstructured Data Models Using Data of All Patients

	Cutpoint	Sensitivity, %	Specificity, %
1- to 10-y model			
Default .5 cutoff	.500	37.4	98.8
Sensitivity 80%	.084	80.0	53.9
Specificity 80%	.126	68.4	80.0
Maximize sensitivity/specificity	.189	61.9	88.0
3- to 10-y model			
Default .5 cutoff	.500	30.8	98.0
Sensitivity 80%	.112	80.0	44.6
Specificity 80%	.162	59.4	80.0
Maximize sensitivity/specificity	.219	52.0	89.2
5- to 10-y model			
Default .5 cutoff	.500	22.1	98.0
Sensitivity 80%	.138	80.0	38.7
Specificity 80%	.186	50.5	80.0
Maximize sensitivity/specificity	.184	51.1	79.7

associations with ADRD. Many of the same demographic characteristics and comorbid conditions that were significant in the structured-data-only model are also significant in the structured and unstructured data model.

Older age and the presence of certain comorbidities (eg, diabetes, stroke, heart disease, depression, bipolar disorder, schizophrenia) were associated with an increased risk of ADRD. In the structured and unstructured data model, MI is associated with a significantly higher risk (in the structured-data-only model, MI was not statistically significant), and diabetes is estimated to increase odds of ADRD by more than 50% (vs 20% in the structured-data-only model). Notably, these comorbidities were two that were refined using unstructured data in the structured and unstructured data model. In addition, the use of benzodiazepines, psychotherapeutic agents, anticonvulsants, non-narcotic analgesics, and minerals and electrolytes were associated with increased ADRD risk, whereas NSAIDs, PPIs, and anti-lipidemics were associated with a lower risk of ADRD in one or both models. A prior emergency department visit was associated with a lower risk of ADRD, whereas delirium and any note mention of ADRD-related terms (in the structured and unstructured data model) were associated with increased risk of ADRD.

The sensitivity and specificity for the structured and unstructured data model are presented at different cutpoints (Table 4). For the 1- to 10-year model, using All Patients data, 80% sensitivity could be achieved with 53.9% specificity, and using a different cutpoint, 80% specificity could be achieved with 68.4% sensitivity. Using a cutpoint that maximizes the combination of these two operating characteristics provides 88.8% specificity and 61.9% sensitivity. Operating characteristics diminish gradually when using 3- to 10- or 5- to 10-year periods (Table 4). When models were run using only those cases aged 65 and older at the index diagnosis and their matched controls, results for the coefficients and the AUROC (Table S4) were similar.

DISCUSSION

Our EMR-based models successfully identified patients with a high risk of developing ADRD up to 5 years in advance, with the 1- to 10-year model providing the most predictive power. Variables that significantly increased likelihood of later presence of ADRD in our models included older age and the presence of several comorbidities including diabetes, ischemic heart disease, and stroke/transient ischemic attack (TIA). The fact that all of these were previously reported as risk factors for either the incidence or progression of cognitive impairment and ADRD[17-19] provides support that the model is clinically sensible and reasonable. Multiple medications that were also previously reported to increase the risk of cognitive impairment, such as anticholinergics,[20] antidepressants,[21] and psychotropic medications,[22] also coincide with several medications our model identified as significantly associated with the later presence of ADRD. The presence of the predefined ADRD-related terms in EMR free text was associated with a substantially increased risk of future ADRD, suggesting that these data are critical for early identification of ADRD. Results were similar in sensitivity analyses limited to those aged 65 years and older. In practice, the group identified by these models could be targeted for subsequent cognitive or biological screening tests. Additionally, as discussed later, these models may allow for more efficient enrollment efforts when identifying potential participants for future prospective studies to better understand the development and progression of ADRD.

Our study also demonstrates that the addition of unstructured EMR data in the form of encounter notes increases the accuracy of the model. This is likely because of the additional detail often provided in these notes, such as diagnostic certainty as well as information not available from current *International Classification of Diseases*-9/10 codes. Specifically, the unstructured data allowed for additional terms that reflect mention in EMR notes of patient characteristics or behaviors such as disorientation, confusion, the presence of a caregiver, memory/cognitive complaints, and occurrences of social withdrawal and personality changes, to name a few. Additionally, the unstructured data allowed for a more accurate identification of cases and controls.

Others have explored the use of natural language processing or machine learning to discriminate between patients with and without dementia.[23-25] For example, Amra

et al[25] developed electronic search algorithms with high sensitivity and specificity for identifying ADRD from EMR notes. Reuben et al[23] examined the impact of combining natural language processing with the presence of dementia-related diagnosis codes and medications (cholinesterase inhibitors and memantine). Notably, the authors observed a significantly improved positive predictive value of ADRD when removing medications from the search algorithm. Although these studies provide promise for the usefulness of these types of techniques, it should be noted that many clinical settings may not have the resources or technical capabilities to perform these kinds of exercises. Amra et al, for example, note that their study leveraged data and software that was specific to the Mayo Clinic. Also, often these approaches may identify patients who are already known by clinicians to potentially have ADRD and are focused on accurately diagnosing those with ADRD. In contrast, the current study attempts to identify a passive digital signal for prodromal ADRD as a way to screen patients long before they exhibit specific symptoms of ADRD.

The models explored in the current study could have application in clinical trials designed to study treatments that delay or prevent the onset of ADRD. Current approaches to primary prevention identify individuals free of cognitive impairment and screen them with biomarkers, usually based on neuroimaging or cerebrospinal fluid examination that can be invasive, expensive, or involve exposure to radiation. A low-cost methodology for identifying high-risk individuals using EMR data for subsequent screening could greatly improve the efficiency of trials. Our models provide flexibility to optimize operating characteristics based on trial design and subsequent follow-up. Once effective preventive interventions are developed, EMR-based screening could provide an efficient and cost-effective method for case-finding persons in need of treatment. Although the current study used EMR data, similar procedures could pursue other types of data, such as administrative data.

This study had several strengths. First, the data are from a large database that reflects 38 healthcare systems and serves more than 13 million patients, providing a large and diverse study population. Second, we explored a variety of models, using various follow-up periods and separately analyzing models with and without the addition of unstructured data. Additionally, we used derivation and validation samples to assess model adequacy.

However, the results should be viewed in light of several limitations. First, several potentially important variables were not available including socioeconomic status and lifestyle variables (eg, smoking, alcohol use, exercise level, and diet). Additionally, there were missing data, and the prevalence of missing data may correlate with institutions and/or individual physicians. However, we attempted to account for missing data by running models only using subjects with "complete" data (ie, those with least one diagnostic and one pharmacy record during the respective time frame) in addition to models using all subjects. Although the Complete Patients models use more information and therefore potentially suffer from less misspecification bias, the All Patients models use all available patients and therefore potentially suffer from less selection bias. Furthermore, the All Patients models that assume absence (ie, score of 0) if drug class information is not available are perhaps more realistic for future use because they can be applied to all patients in the EMR even for patients missing pharmacy data. Therefore, we chose to display results for the risk factors for All Patients models in Table 3.

It should also be noted that EMR data is collected for clinical purposes, not research, and health encounters that occur outside of this health system may not be captured by these data. Another limitation is that these analyses used only data on diagnoses, medications, and notes, and they did not leverage laboratory information or procedures. Further, case identification was based on the criterion of at least two clinical visits with ADRD diagnostic codes that could have misclassified subjects as demonstrated by studies of the accuracy of claims-based dementia definitions.[26,27] However, the unstructured data was used to improve case identification and hopefully mitigated any bias associated with misclassification from diagnosis codes alone. Additionally, medication use was classified as "current," "prior," or "never" and ignored any dose changes or drug interactions. Also, although the INPC includes EMR data from most of the hospital-based healthcare systems in the state and reflects a diverse and representative sample from this type of health system, it is not representative of small practices or independent ambulatory care centers, and therefore our results may not be generalizable to these populations.

However, in addition to the split-sample training and validation procedure performed in the current study, in future research we plan to validate our results within these other populations and within a different database to assess the greater generalizability of our findings. Finally, although the data were from a large database, the results may not be generalizable to the entire US population.

We anticipate several next steps. To further refine these models, future iterations could examine subsets of these populations by age group or other relevant demographic characteristics. Other types of models, such as time-series and/or growth-mixed models, would allow for more complex classifications of comorbidities and medication use. Combining these data with other routinely collected data may also serve to extend our understanding of the early predictors of ADRD. In summary, EMR-based data allow for a targeted and scalable process for early identification of the risk of ADRD as an alternative to traditional cognitive and biological population screening.

ACKNOWLEDGMENTS

Financial Disclosure: Malaz Boustani is the founding director of the Sandra Eskenazi Center for Brain Care Innovation. Richard Lipton is the Edwin S. Lowe Professor of Neurology at the Albert Einstein College of Medicine in New York. He receives research support from the National Institutes of Health (NIH): 2PO1 AG003949 (multiple Principal Investigator), 5U10 NS077308 (Principal Investigator), R21 AG056920 (Investigator), 1RF1 AG057531 (Site PI), RF1 AG054548 (Investigator), 1RO1 AG048642 (Investigator), R56 AG057548 (Investigator), U01062370 (Investigator), RO1 AG060933 (Investigator), K23 NS09610 (Mentor), K23AG049466 (Mentor), K23 NS107643 (Mentor). He also receives support from the Migraine Research Foundation and the National Headache Foundation. He serves on the editorial board of *Neurology*, is a senior advisor to

Headache, and associate editor at *Cephalalgia*. He has reviewed for the National Institute on Aging and the National Institute for Neurological Disorders, holds stock options in eNeura Therapeutics and Biohaven Holdings; serves as consultant, advisory board member, or has received honoraria from: American Academy of Neurology, Alder, Allergan, American Headache Society, Amgen, Avanir, Biohaven, Biovision, Boston Scientific, Dr. Reddy's, Electrocore, Eli Lilly, eNeura Therapeutics, GlaxoSmithKline, Merck, Pernix, Pfizer, Supernus, Teva, Trigemina, Vector, and Vedanta. He receives royalties from Wolff's *Headache*, 7th and 8th eds. (Oxford University Press, 2009), Wiley, and Informa. Rezaul Karim Khandker, Christopher Black, and Vasu Chandrasekaran are employees of Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. Anthony Perkins, Stephen Duong, Paul R. Dexter, Craig Solid, and Patrick Monahan have no interests to disclose.

Conflicts of Interest: The authors have declared no conflicts of interest for this article.

Author Contributions: *Study concept and design, acquisition of data/information*: Boustani, Perkins, and Monahan. *Data analysis and interpretation*: Perkins and Monahan. *Critical revision of the manuscript for important intellectual content and final approval of manuscript for submission*: All authors.

Sponsor's Role: This study was supported by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. It was also supported in part by the Einstein Aging Study: NIH/NIA 2PO1 AG 003949.

REFERENCES

1. US Department of Health and Human Services. *National plan to address Alzheimer's disease: 2018 update*. 2018. https://aspe.hhs.gov/system/files/pdf/259581/NatPlan2018.pdf. Accessed March 22, 2019.
2. Medicare Interactive. *Annual wellness visit*. Medicare Rights Center. https://www.medicareinteractive.org/get-answers/medicare-covered-services/preventive-services/annual-wellness-visit. Accessed March 22, 2019.
3. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189-198.
4. Brandt J, Spencer M, Folstein M. The telephone interview for cognitive status. Neuropsychiatry Neuropsychol Behav Neurol. 1988;1:111-117.
5. Nasreddine ZS, Phillips NA, Bedirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005;53(4):695-699.
6. Buschke H, Kuslansky G, Katz M, et al. Screening for dementia with the memory impairment screen. Neurology. 1999;52(2):231-238.
7. Lipton RB, Katz MJ, Kuslansky G, et al. Screening for dementia by telephone using the memory impairment screen. J Am Geriatr Soc. 2003;51(10):1382-1390.
8. Fowler NR, Perkins AJ, Turchan HA, et al. Older primary care patients' attitudes and willingness to screen for dementia. J Aging Res. 2015;2015:423265.
9. Harrawood A, Fowler NR, Perkins AJ, LaMantia MA, Boustani MA. Acceptability and results of dementia screening among older adults in the United States. Curr Alzheimer Res. 2018;15(1):51-55.
10. US Preventive Service Task Force. Final recommendation statement on cognitive impairment in older adults: screening. 2014. https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/cognitive-impairment-in-older-adults-screening. Accessed August 18, 2019.
11. Alzheimer's Association. 2018 Alzheimer's disease facts and figures. 2018. https://www.alz.org/media/HomeOffice/Facts%20and%20Figures/facts-and-figures.pdf. Accessed January 11, 2019.
12. Steenland K, Goldstein FC, Levey A, Wharton W. A meta-analysis of Alzheimer's disease incidence and prevalence comparing African-Americans and Caucasians. J Alzheimers Dis. 2016;50(1):71-76.
13. Fowler NR, Perkins AJ, Gao S, Sachs GA, Uebelhor AK, Boustani MA. Patient characteristics associated with screening positive for Alzheimer's disease and related dementia. Clin Interv Aging. 2018;13:1779-1785.
14. Pandharipande PP, Girard TD, Jackson JC, et al. Long-term cognitive impairment after critical illness. N Engl J Med. 2013;369(14):1306-1316.
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med. 2015;162(10):735-736.
16. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73.
17. Bleckwenn M, Kleineidam L, Wagner M, et al. Impact of coronary heart disease on cognitive decline in Alzheimer's disease: a prospective longitudinal cohort study in primary care. Br J Gen Pract. 2017;67(655):e111-e117.
18. Cheng G, Huang C, Deng H, Wang H. Diabetes as a risk factor for dementia and mild cognitive impairment: a meta-analysis of longitudinal studies. Intern Med J. 2012;42(5):484-491.
19. Kalaria RN, Akinyemi R, Ihara M. Stroke injury, cognitive impairment and vascular dementia. Biochim Biophys Acta. 2016;1862(5):915-925.
20. Richardson K, Fox C, Maidment I, et al. Anticholinergic drugs and risk of dementia: case-control study. BMJ. 2018;361:k1315.
21. Lee CW, Lin CL, Sung FC, Liang JA, Kao CH. Antidepressant treatment and risk of dementia: a population-based, retrospective case-control study. J Clin Psychiatry. 2016;77(1):117-122. quiz 122.
22. Shash D, Kurth T, Bertrand M, et al. Benzodiazepine, psychotropic medication, and dementia: a population-based cohort study. Alzheimers Dement. 2016;12(5):604-613.
23. 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):658-659.
24. Jammeh EA, Carroll CB, Pearson SW, et al. Machine-learning based identification of undiagnosed dementia in primary care: a feasibility study. BJGP Open. 2018;2(2): bjgpopen18X101589. https://bjgpopen.org
25. 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;37:202-205.
26. Lee E, Gatz M, Tseng C, et al. Evaluation of Medicare claims data as a tool to identify dementia. J Alzheimers Dis. 2019;67(2):769-778.
27. Zhu CW, Ornstein KA, Cosentino S, Gu Y, Andrews H, Stern Y. Misidentification of dementia in Medicare claims and related costs. J Am Geriatr Soc. 2019;67(2):269-276.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Appendix S1: Supplementary material.