Received: 31 August 2023 | Revised: 29 January 2024 | Accepted: 18 February 2024

DOI: 10.1002/dad2.12572

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Predicting 5-year dementia conversion in veterans with mild cognitive impairment

Chase Irwin[1,2] | **Donna Tjandra**[1,3] | **Chengcheng Hu**[2,4] | **Vinod Aggarwal**[5,6] | **Amanda Lienau**[6] | **Bruno Giordani**[3] | **Jenna Wiens**[3] | **Raymond Q. Migrino**[1,2]

[1]Phoenix Veterans Affairs Health Care System, Phoenix, Arizona, USA

[2]University of Arizona College of Medicine-Phoenix, Phoenix, Arizona, USA

[3]University of Michigan, Ann Arbor, Michigan, USA

[4]College of Public Health, University of Arizona, Tucson, Arizona, USA

[5]MDClone Limited, Be'er Sheva, Israel

[6]VHA Office of Healthcare Innovation and Learning, VA Central Office, Washington, District of Columbia, USA

Correspondence
Raymond Q. Migrino, Phoenix Veterans Affairs Health Care System, 650 E. Indian School Road, Phoenix, AZ 85012, USA.
Email: raymond.migrino@va.gov

Funding information
National Science Foundation, Grant/Award Number: IIS 2124127

Abstract

INTRODUCTION: Identifying mild cognitive impairment (MCI) patients at risk for dementia could facilitate early interventions. Using electronic health records (EHRs), we developed a model to predict MCI to all-cause dementia (ACD) conversion at 5 years.

METHODS: Cox proportional hazards model was used to identify predictors of ACD conversion from EHR data in veterans with MCI. Model performance (area under the receiver operating characteristic curve [AUC] and Brier score) was evaluated on a held-out data subset.

RESULTS: Of 59,782 MCI patients, 15,420 (25.8%) converted to ACD. The model had good discriminative performance (AUC 0.73 [95% confidence interval (CI) 0.72–0.74]), and calibration (Brier score 0.18 [95% CI 0.17–0.18]). Age, stroke, cerebrovascular disease, myocardial infarction, hypertension, and diabetes were risk factors, while body mass index, alcohol abuse, and sleep apnea were protective factors.

DISCUSSION: EHR-based prediction model had good performance in identifying 5-year MCI to ACD conversion and has potential to assist triaging of at-risk patients.

KEYWORDS
Alzheimer's disease, dementia, electronic health records, mild cognitive impairment, prediction modeling, synthetic data

Highlights

- Of 59,782 veterans with mild cognitive impairment (MCI), 15,420 (25.8%) converted to all-cause dementia within 5 years.
- Electronic health record prediction models demonstrated good performance (area under the receiver operating characteristic curve 0.73; Brier 0.18).
- Age and vascular-related morbidities were predictors of dementia conversion.
- Synthetic data was comparable to real data in modeling MCI to dementia conversion.

Key Points

- An electronic health record–based model using demographic and co-morbidity data had good performance in identifying veterans who convert from mild cognitive impairment (MCI) to all-cause dementia (ACD) within 5 years.
- Increased age, stroke, cerebrovascular disease, myocardial infarction, hypertension, and diabetes were risk factors for 5-year conversion from MCI to ACD.
- High body mass index, alcohol abuse, and sleep apnea were protective factors for 5-year conversion from MCI to ACD.
- Models using synthetic data, analogs of real patient data that retain the distribution, density, and covariance between variables of real patient data but are not attributable to any specific patient, performed just as well as models using real patient data. This could have significant implications in facilitating widely distributed computing of health-care data with minimized patient privacy concern that could accelerate scientific discoveries.

1 | BACKGROUND

Mild cognitive impairment (MCI) is a heterogenous syndrome characterized by cognitive impairment that is more than normal aging and could be an early manifestation of neurodegenerative diseases that later progress to dementia.[1] Prior autopsy studies show that the brain pathology in MCI is intermediate in severity between cognitively normal controls and patients with more advanced Alzheimer's disease (AD), the most common neurodegenerative condition.[2–4] However, MCI could also be a precursor to other non-AD dementing conditions such as cerebrovascular disease and Lewy body disease[5,6]. Identifying MCI patients at risk of developing dementia could be helpful for targeting candidates for early treatment especially as promising drugs that slow the cognitive and pathologic decline, such as lecanemab[7], become increasingly available. It will also allow selection of patients that are most at risk for participation in clinical trials of new candidate therapeutics, potentially requiring smaller sample sizes to show benefit, leading to reduced study cost and enhanced research participant safety.

Existing models to predict dementia, mainly AD, have focused on neuropsychologic test scores and biomarkers from cerebrospinal fluid and brain imaging[8–10]. The generalizability of these models is limited by the relatively small number of participants and the complex and sometimes invasive nature of the input variables that are not widely obtained in clinical practice. Electronic health record (EHR)–based prediction models of dementia potentially have an advantage in generalizability over existing models because of the large number of unique patients involved and access to high-dimensional data that are collected during routine clinical encounters[11,12]. The primary aim of the study is to develop a generalizable EHR-based model to predict MCI to all-cause dementia (ACD) conversion at 5 years using the large multicenter Veterans Affairs (VA) health-care database. While EHR-based prediction models have advantages due to access to a large dataset, creating and optimizing these models are constrained by limitation of access to patient medical records due to privacy concerns. This problem could be partially addressed by providing wide access to and using synthetic data to augment model building. Synthetic data are analogs of original patient data that aim to retain the distribution, density, and co-variance between variables within clusters of similar patients, but are not attributable to the original patients[13]. However, the validation of model performance based on EHR synthetic data on various disease models remains limited[14]. A secondary aim of the study is to compare the performance of MCI to ACD prediction models based on EHR-derived real patient versus synthetic data.

2 | METHODS

2.1 | Study population

We assembled a retrospective cohort of veterans who were seen January 1, 1999 to December 31, 2016 in the US VA Healthcare System using an internal cloud analytics environment that hosts a copy of the Corporate Data Warehouse (CDW), which is a consolidation of data from disparate sources within the VA into a single coherent data model. The study protocol was reviewed and approved by the Institutional Review Board of the Phoenix VA Health Care System with a waiver of informed consent (Protocol Migrino1593816).

Patients were eligible to enter if they were ≥ 50 years old and were diagnosed with MCI (Figure 1). Diagnosis of MCI was based on the patient having International Classification of Diseases Ninth or Tenth revision (ICD-9 or -10) classification of MCI (Table S1 in supporting information) made on two or more separate clinic visits, an entry criterion based on MVP Cog Working Group validated to have 95%

specificity based on rigorous chart review[15]. The date of initial diagnosis of MCI was used as date of diagnosis. Patients with diagnosis of dementia (Table S1) prior to or up to 6 months after initial MCI diagnosis were excluded from analysis. Patients were classified into two groups based on whether they were (1) diagnosed to have ACD within 5 years after MCI diagnosis (ACD converters) or (2) did not have ACD diagnosis or were right censored (lost to follow-up or died) within 5 years after MCI diagnosis (ACD non-converters). The alive/dead status of those lost to follow-up was not determined using separate non-EHR datasets because the aim of the study was to evaluate the utility of data derived only from the VA EHR. ACD was defined using the ICD-9 or ICD-10 codes (Table S1) from the VA Centralized Interactive Phenomics Resource (CIPHER) Phenotype 00083 (https://www.research.va.gov/programs/cipher.cfm)[15] validated to have 82% specificity based on rigorous chart review.

2.2 | Demographic and co-morbid condition variables

Demographic (age, sex, race, ethnicity, and body mass index [BMI]) and selected co-morbid conditions were extracted from EHR at the time of MCI diagnosis and the dataset was locked prior to final analyses. All race data are self-reported and we used the last self-designation to group races into the following categories: White, Black or African American, Asian/Pacific Islander/Native Hawaiian, American Indian or Alaska Native, and Multiracial/Other (Declined to Answer/Unknown). If a patient did not have any recorded BMI measurement (2.43%), then we imputed the mean. Co-morbid conditions were selected a priori based on previous literature testing these conditions as potential risk factors for dementia[16–19] and identified using ICD-9 or ICD-10 codes (Table S1) using criteria for the Charlson Comorbidity Index[20]. For traumatic brain injury (TBI), we used ICD codes from a prior study on veterans showing association between TBI and later development of dementia[21]. If the condition is not included in the Charlson list, we used Elixhauser Comorbidity Index[22], CIPHER, or Saunders et al.'s[23] study (hearing loss).

2.3 | Statistical methods

2.3.1 | Descriptive statistics

We randomly partitioned our cohort into a training set (70%) and test set (30%) for prediction modeling. Descriptive statistics were stratified by conversion status and reported as frequencies and proportions or medians and interquartile ranges (IQRs). Chi-square tests and Wilcoxon rank-sum tests were used to evaluate differences between strata.

2.3.2 | Cox proportional hazards model

Patients were followed from MCI diagnosis (entry age) until they developed ACD, they were lost to follow-up, died, or 5 years after MCI diagnosis. We used the Kaplan–Meier estimator to estimate the conversion probability and corresponding 95% confidence interval (CI) to ACD at 5 years for our full, real cohort. We used Cox proportional hazards model to estimate the hazard ratio (HR), 95% CIs, and corresponding *P* value for the risk of developing ACD for each co-morbidity and demographic feature. We used backward stepwise selection on co-morbid predictors to identify a parsimonious model based on the Akaike information criterion (AIC)[24]. The proportional hazards assumption of the fitted Cox proportional hazard model was evaluated for each predictor by graphical methods and a formal score test of B(t).

2.3.3 | Model performance evaluation

We applied the trained Cox model to our held-out test set and estimated the linear predictor score and expected conversion probability for each observation in our test set. We reported the median and IQR of the expected conversion probabilities of our test set. Next, we evaluated our models' ability to predict patient ACD conversion at 5 years through non-parametric inverse probability of censoring weighting estimation of the time-dependent areas under the receiver operating characteristic curve (AUCs) and time-dependent Brier scores[25,26]. Time-dependent AUCs instead of C-index was used because previous simulation studies demonstrated that the C-index is not an appropriate

RESEARCH IN CONTEXT

1. **Systematic review**: We performed an extensive literature review to identify predictors of dementia conversion and current dementia prediction modeling approaches. We also identified previous work done to validate the use of synthetic data in statistical modeling.
2. **Interpretation**: Our findings show that routinely collected demographic and co-morbidity data can be used to predict 5-year conversion from mild cognitive impairment (MCI) to dementia. We also demonstrate that the predictive models using synthetic data derived from real patient data perform as well as predictive models from real patient data.
3. **Future directions**: The MCI to dementia predictive model derived from electronic health records could be used to identify high-risk patients for consideration of non-pharmacologic or new, expensive pharmacologic interventions. It could also be used to define an enriched at-risk patient group to target for clinical trials of new therapies. Importantly, validation of synthetically derived predictive models could allow widely distributed computing with minimal risk of privacy breach, reducing barriers to entry and facilitating scientific discovery.

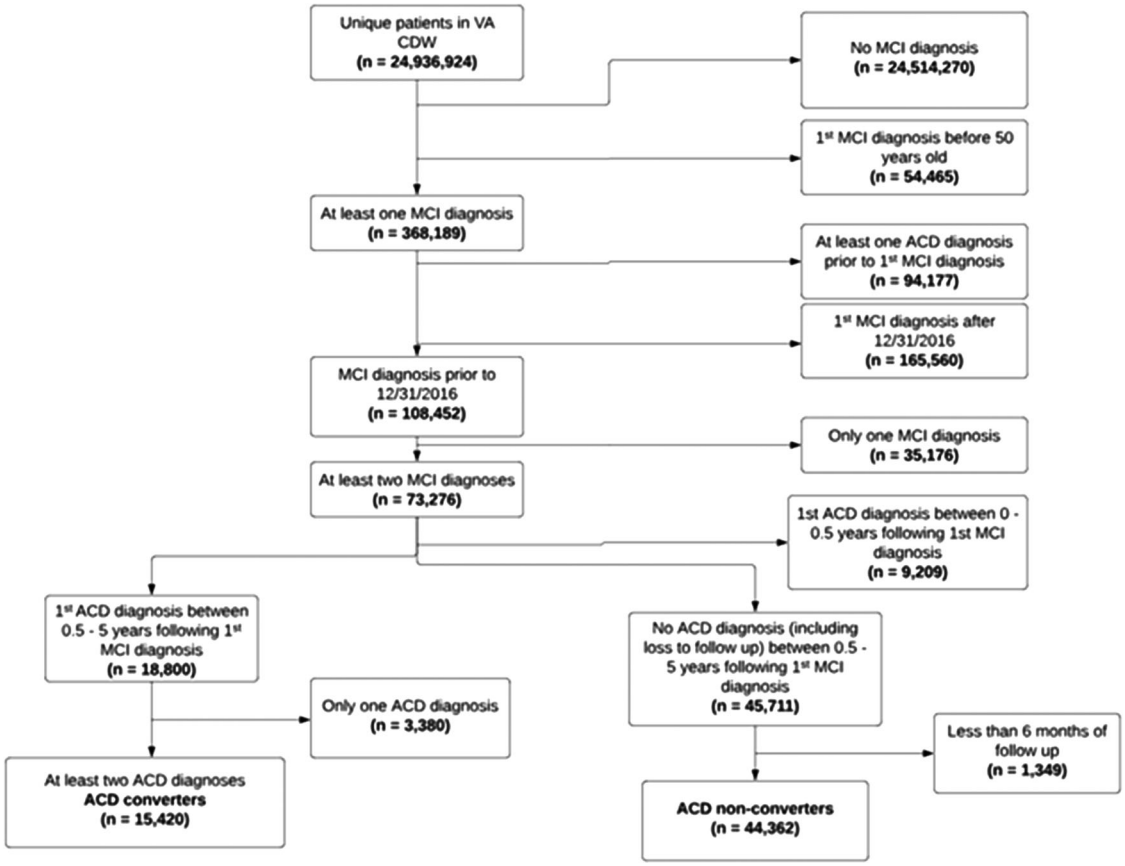

FIGURE 1 Flow chart of inclusion and exclusion criteria. ACD, all-cause dementia; CDW, Corporate Data Warehouse; MCI, mild cognitive impairment; VA, Veterans Affairs

discriminatory measure for evaluating *t*-year predicted risks due to biased estimates of mis-specified models[26].

2.4 | Synthetic data generation and model performance

We evaluated the utility of synthetically derived data for training models to predict MCI to ACD conversion by repeating our previously outlined methodology on three synthetic training sets and comparing the results to our real training set.

2.4.1 | Synthetic patient data generation

We used commercial software (MDClone ADAMS Platform, MDClone Ltd.) to derive synthetic patient data from CDW. The software is designed to compute and preserve the original cohort's statistical properties and higher-order relationships and to create a synthetic analog cohort without any one-to-one correspondence between the original and synthetic patients[13,14,27]. To best mimic the real cohort, the ADAMS Platform provides the ability to select only those variables that are relevant to the research question. These comprise input to the synthetic data generator. The generator first derives a statistical model of the real patient cohort. The generator then creates new fictitious (i.e., synthetic) records to fit that model while maintaining the distribution, density, covariance, and other statistical measures between similar patients. The outcome is a similar number of synthetic patient records based on the variables of interest that maintain the relationship between variables.

2.4.2 | Synthetic versus real data comparisons

We reported the descriptive statistics, estimated Cox proportional hazard model parameters, and prediction results of our three synthetically derived training sets. The time-dependent AUCs and time-dependent Brier scores for prediction of ACD conversion at 5 years for synthetic training sets were compared to the prediction results of our real training set. Tests of comparisons and estimated pointwise 95% CIs were derived from the limiting Gaussian processes and estimated asymptotic variances[25,28]. We evaluated the correlation between real and synthetic expected conversion probabilities through R-squared statistics.

All analyses were performed using R statistical software version 4.2.2 (https://www.R-project.org) with the gmodels, survival, stats, StepReg, timeROC, and riskregression extension packages. All statistical tests were two-sided; alpha level of 0.05 was used to determine statistical significance.

3 | RESULTS

Out of 24,936,924 unique patients from 1999 to 2016, 59,782 patients met inclusion criteria (Figure 1). Fifteen thousand four hundred twenty (25.8%) converted to ACD within 5 years, while the rest either did not have ACD diagnosis, died, or were lost to follow-up within 5 years. The Kaplan–Meier estimate of 5-year conversion from MCI to ACD was 28.4% (95% CI 28.0%–28.8%). Median time to conversion was 1.94 (IQR 1.09–3.10) years. Excluding patients with MCI diagnosis before age 50 years, the median age of MCI diagnosis in the VA cohort was 71.0 (IQR 63.7–80.7) years. MCI patients who converted to ACD were older than those who did not (Table 1A-B, Table S2 in supporting information). The overall cohort was predominantly male and White; male and White participants had greater representation in ACD converters than non-converters. There were fewer obese MCI patients who converted to ACD. On univariate analyses, all co-morbid conditions were significantly different between ACD converters versus non-converters in the full cohort (Table S3 in supporting information), but in the training set, co-morbid diabetes showed no significant difference between the groups ($P = 0.09$, Table 2A). Cox proportional hazards showed that increasing age is the strongest independent risk factor for ACD conversion, with HR of 1.53 (95% CI 1.26–1.85) in those 55 to 60 years old (compared to 50–55 years old), going up to HR of 8.94 (95% CI 7.60–10.53) in those > 85 years old (Table 3A; a comparison of Full versus Reduced Model is shown in Table S4 in supporting information). Other associated independent risk factors include cerebrovascular disease, stroke, myocardial infarction, hypertension, and diabetes, with HRs ranging from 1.06 to 1.09, which are less than that for age. Associated protective factors included high BMI, alcohol abuse, and sleep apnea. When the model was applied to the test set, the time-dependent AUC was 0.73 (95% CI 0.72–0.74) and Brier score was 0.18 (95% CI 0.17–0.18) suggesting good discriminative performance and calibration by the model (Table 4A, Figure 2).

Univariate analysis showed less TBI co-morbidity in ACD converters versus non-converters (6.02 vs. 8.89%, $P < 0.001$; Table S5 in supporting information, Table 2A), but multivariable analysis did not reveal TBI to be an independent risk factor (Table 4A). To explore this further, we compared the age of MCI patients with TBI versus those without TBI in the full cohort and found MCI patients with co-morbid TBI were younger (63.36 [IQR 56.43—70.86] versus 71.81 [IQR 64.53—81.11] years, $P < 0.001$). We next performed a comparison between ACD converters and age-matched non-converters in our full cohort and showed no significant difference in TBI co-morbidity (6.02 vs. 5.68%, $P = 0.22$; Table S5).

Synthetic data performance

The demographic profiles of MCI ACD converters versus non-converters were similar when each of the synthetic datasets was compared to the real dataset (Table 1 and Table S6 in supporting information). In similar fashion, the co-morbidity profiles of MCI ACD converters versus non-converters were similar between each synthetic dataset compared to the real dataset (Table 2 and Table S7 in supporting information). Cox proportional hazards models of the synthetic datasets showed similar risk and protective factors for ACD conversion between real and synthetic patient data, with magnitude of HRs in close approximation (Table 3).

Expected conversion probabilities of our real model were similar and highly correlated ($R^2 = 0.99$) to all synthetic model values (Table 4). The predictive models' time-dependent AUCs (all 0.73) and time-dependent Brier scores (all 0.18) of synthetic data were also similar to real data (Table 4, Figure 2).

4 | DISCUSSION

MCI represents the clinical and neuropathologic transition between the cognitive changes in normal aging and early AD[1,2] and non-AD causes of dementia, such as cerebral infarction and neocortical Lewy bodies[6]. A meta-analysis of cohort studies shows that ≈ 39% of MCI patients convert to dementia with 34% and 6% converting to AD and vascular dementia, respectively, with annual conversion rate of 9.6%[29]. This compares to our 5-year conversion rate to ACD estimate of 28.4%, representing an important subset of MCI patients. Early identification of MCI patients at risk for developing dementia could be useful for closer disease surveillance and early initiation of non-pharmacologic interventions, pharmacologic treatments for symptomatic relief,[30] or newer disease-modifying agents, such as the recently US Food and Drug Administration–approved agent lecanemab[7].

There is consensus that for meaningful disease modification in AD, treatment should be initiated very early in the preclinical stage, requiring future clinical trials to have trial-ready cohorts enriched with identified high-risk participants[31]. This could be enhanced by exploiting the EHR with its expansive data obtained during routine clinical care. We previously demonstrated the utility of an EHR-based machine learning model to predict AD onset from demographic, diagnostic, and medication information from patient encounters collected from > 4 million patients, with the model achieving good accuracy (AUC 0.70)[11]. In the current study, we focus on creating a model to predict ACD conversion within 5 years of MCI diagnosis derived from VA EHR of close to 25 million patients. Results show that age is the overwhelming risk factor for MCI to ACD conversion, with HRs of 1.53 from age 55 to 60 to 8.94 in those > 85 years, consistent with prior studies showing that the greatest risk factor for AD is advanced age[32], including data from three large longitudinal studies[33]. Vascular disease–related co-morbidities such as stroke, cerebrovascular disease, myocardial infarction, hypertension, and diabetes are, comparatively, more modest risk factors (HRs 1.06–1.09). Prior epidemiologic, preclinical, and clinical data also show that vascular disease is strongly associated with AD[34,35]. Unbiased data-driven analyses showed that vascular dysfunction is the earliest brain pathology in AD[36] and regional blood flow differences were shown to discriminate between MCI

TABLE 1 Demographic data of real and synthetic training set #1.

	A. Training set real ($n = 41,817$)			B. Test set real ($n = 17,965$)			C. Training set synthetic #1 ($n = 41,709$)		
Demographics	**ACD ($n = 10,784$)**	**No ACD ($n = 31,033$)**	***P* value**	**ACD ($n = 4,636$)**	**No ACD ($n = 13,329$)**	***P* value**	**ACD ($n = 10,729$)**	**No ACD ($n = 30,980$)**	***P* value**
Age at MCI DX, median (IQR), years	77.01 (69.17–83.34)	69.01 (62.11–79.18)	<0.001	76.72 (69.08–83.16)	68.91 (62.09–79.07)	<0.001	77.02 (69.20–83.35)	69.02 (62.12-79.20)	<0.001
Age at MCI DX, no. (%), years			<0.001			<0.001			<0.001
50–55	162 (1.50)	2,611 (8.41)		67 (1.44)	1,105 (8.29)		157 (1.46)	2,599 (8.39)	
55–60	311 (2.88)	3,220 (10.38)		136 (2.93)	1,405 (10.54)		301 (2.81)	3,209 (10.36)	
60–65	890 (8.25)	4,916 (15.84)		398 (8.59)	2,103 (15.78)		884 (8.24)	4,914 (15.86)	
65–70	1,673 (15.51)	5,809 (18.72)		710 (15.32)	2,560 (19.21)		1,663 (15.50)	5,802 (18.73)	
70–75	1,662 (15.41)	3,835 (12.36)		743 (16.03)	1,655 (12.42)		1,659 (15.46)	3,819 (12.33)	
75–80	1,924 (17.84)	3,445 (11.10)		831 (17.93)	1,469 (11.02)		1,913 (17.83)	3,439 (11.10)	
80–85	2,135 (19.80)	3,542 (11.41)		928 (20.02)	1,531 (11.49)		2,135 (19.90)	3,531 (11.40)	
> 85	2,027 (18.80)	3,655 (11.78)		823 (17.75)	1,501 (11.26)		2,017 (18.80)	3,667 (11.84)	
Race, no. (%)			<0.001			0.002			<0.001
Asian/Pacific[a]	146 (1.35)	426 (1.37)		60 (1.29)	178 (1.34)		144 (1.34)	423 (1.37)	
Black	1,349 (12.51)	4,431 (14.28)		590 (12.73)	1,998 (14.99)		1,342 (12.51)	4,427 (14.29)	
Native American	56 (0.52)	210 (0.68)		20 (0.43)	78 (0.59)		55 (0.51)	209 (0.68)	
Other[b]	1,006 (9.33)	2,886 (9.30)		423 (9.12)	1,234 (9.26)		999 (9.31)	2,879 (9.29)	
White	8,227 (76.29)	23,080 (74.37)		3,543 (76.42)	9,841 (73.83)		8,189 (76.33)	23,042 (74.38)	
Ethnicity, no. (%)			0.03			0.03			0.04
Not Hispanic or Latino	9,590 (88.93)	27,436 (88.41)		4,069 (87.77)	11,802 (88.54)		9,541 (88.93)	27,389 (88.41)	
Hispanic or Latino	650 (6.03)	1,828 (5.89)		322 (6.95)	783 (5.87)		647 (6.03)	1,826 (5.89)	
Other[c]	544 (5.05)	1,769 (5.70)		245 (5.29)	744 (5.58)		541 (5.04)	1,765 (5.70)	
Sex, no. (%)			<0.001			<0.001			<0.001
Female	354 (3.28)	1,472 (4.74)		156 (3.37)	667 (5.00)		353 (3.29)	1,471 (4.75)	
Male	10,430 (96.72)	29,561 (95.26)		4,480 (96.64)	12,662 (95.00)		10,376 (96.71)	29,509 (95.25)	
BMI, no. (%)			<0.001			<0.001			<0.001
Underweight	105 (0.97)	372 (1.20)		60 (1.29)	166 (1.25)		104 (0.97)	371 (1.20)	
Normal	2,952 (27.37)	6,717 (21.65)		1,273 (27.46)	2,909 (21.83)		2,926 (27.27)	6,709 (21.66)	
Overweight	4,725 (43.82)	12,584 (40.55)		1,958 (42.24)	5,312 (39.85)		4,714 (43.94)	12,561 (40.55)	
Obese	3,002 (27.84)	11,360 (36.61)		1,345 (29.01)	4,942 (37.08)		2,985 (27.82)	11,339 (36.60)	
Follow-up time, median (IQR), years	6.08 (4.55–7.89)	6.45 (4.86–9.10)	<0.001	6.12 (4.60–8.10)	6.44 (4.89–9.04)	<0.001	6.08 (4.55-7.88)	6.45 (4.86–9.09)	<0.001

Abbreviations: ACD, all-cause dementia; BMI, body mass index; DX, diagnosis; IQR, interquartile range; MCI, mild cognitive impairment.

[a] Patients who self-identified as Asian or Native Hawaiian or other Pacific Islander

[b] Patients who self-identified as multiracial, unknown, declined to answer, or missing.

[c] Patients who self-identified as declined to answer, unknown by patient, or missing.

TABLE 2 Co-morbidity data of real and synthetic training set #1.

	A. Training set real ($n = 41,817$)			B. Test set real ($n = 17,965$)			C. Training set synthetic #1 ($n = 41,709$)		
	ACD ($n = 10,784$)	No ACD ($n = 31,033$)	*P* value	ACD ($n = 4,636$)	No ACD ($n = 13,329$)	*P* value	ACD ($n = 10,729$)	No ACD ($n = 30,980$)	*P* value
Co-morbidities, no. (%)									
Heart failure	1,689 (15.66)	4,636 (14.94)	0.07	750 (16.18)	1,967 (14.76)	0.02	1,678 (15.64)	4,628 (14.94)	0.08
Renal disease	1,860 (17.25)	4,863 (15.67)	<0.001	865 (18.66)	2,031 (15.24)	<0.001	1,849 (17.23)	4,852 (15.66)	<0.001
Rheumatic disease	468 (4.34)	1,189 (3.83)	0.02	218 (4.70)	494 (3.71)	0.003	470 (4.38)	1,175 (3.79)	0.008
Hyperlipidemia	8,704 (80.71)	23,984 (77.29)	<0.001	3,769 (81.30)	10,323 (77.45)	<0.001	8,642 (80.54)	23,908 (77.17)	<0.001
Sleep apnea	2,300 (21.33)	8,261 (26.62)	<0.001	1,001 (21.59)	3,636 (27.28)	<0.001	2,291 (21.35)	8,247 (26.62)	<0.001
Peripheral vascular disease	2,661 (24.68)	6,490 (20.91)	<0.001	1,139 (24.57)	2,729 (20.47)	<0.001	2,644 (24.64)	6,483 (20.93)	<0.001
Peptic ulcer disease	733 (6.80)	1,854 (5.97)	<0.001	346 (7.46)	788 (5.91)	<0.001	720 (6.71)	1,856 (5.99)	0.008
Atrial fibrillation	1,652 (15.32)	4,104 (13.23)	<0.001	782 (16.87)	1,725 (12.94)	<0.001	1,642 (15.30)	4,094 (13.22)	<0.001
Myocardial infarction	1,283 (11.90)	3,175 (10.23)	<0.001	558 (12.04)	1,343 (10.08)	<0.001	1,267 (11.81)	3,169 (10.23)	<0.001
Hypertension	9,071 (84.12)	24,763 (79.80)	<0.001	3,938 (84.94)	10,598 (79.51)	<0.001	9,003 (83.91)	24,694 (79.71)	<0.001
Cerebrovascular disease no stroke	2,072 (19.21)	5,314 (17.12)	<0.001	917 (19.78)	2,192 (16.45)	<0.001	2,052 (19.13)	5,299 (17.12)	<0.001
Stroke	1,057 (9.80)	2,810 (9.06)	0.02	454 (9.79)	1,172 (8.79)	0.04	1,049 (9.78)	2,807 (9.06)	0.03
Depression	5,702 (52.88)	19,260 (62.06)	<0.001	2,470 (53.28)	8,324 (62.45)	<0.001	5,630 (52.48)	19,194 (61.96)	<0.001
Alcohol abuse	1,675 (15.53)	7,035 (22.67)	<0.001	689 (14.86)	3,068 (23.02)	<0.001	1,656 (15.44)	7,016 (22.65)	<0.001
Liver disease	933 (8.65)	3,429 (11.05)	<0.001	392 (8.46)	1,541 (11.56)	<0.001	927 (8.64)	3,418 (11.03)	<0.001
Diabetes	4,296 (39.84)	12,073 (38.90)	0.09	1,967 (42.43)	5,125 (38.45)	<0.001	4,255 (39.66)	12,048 (38.89)	0.16
Hearing loss	6,160 (57.12)	16,071 (51.79)	<0.001	2,639 (56.92)	6,864 (51.50)	<0.001	6,112 (56.97)	16,035 (51.76)	<0.001
Traumatic brain injury	636 (5.90)	2,833 (9.13)	<0.001	292 (6.30)	1,109 (8.32)	<0.001	614 (5.72)	2,807 (9.06)	<0.001

Abbreviation: ACD, all-cause dementia.

converters to AD versus non-converters.[37,38] The modest contribution of vascular-related co-morbidities vis-a-vis age highlights the need to identify non-traditional novel mechanistic determinants by which aging induces pathology[39]. On the other hand, high BMI was protective of ACD conversion. This is consistent with prior studies that in late life, elevated BMI was found to be associated with lower AD risk[40] and slower disease progression in MCI[41]. The biological mechanisms underlying this observation remain unknown with some proposing changes in behaviors such as eating, decreased energy metabolism leading to decline in BMI and cognition, and changes in adipose tissue hormone levels[41]. Our data show that alcohol abuse is associated with ≈ 6% lower ACD conversion risk. Prior epidemiologic data do not provide strong evidence that alcohol use affects AD development[42] but interestingly, consumption of wine, but not liquor, beer, or total alcohol, was associated with lower risk of dementia, although this was confined to those without the apolipoprotein E ε4 allele[43]. The mechanistic basis of our observation on alcohol abuse and ACD conversion should be investigated further. Additionally, sleep apnea was found to be protective against ACD conversion. In contrast, a meta-analysis of 14 studies showed that sleep-disordered breathing was associated with increased risk of cognitive impairment[44] although it did not address the role of sleep-disordered breathing in MCI to ACD conversion. The underlying bases of these discrepant observations need to be explored further.

TBI is a known dementia risk factor to which veterans are disproportionately exposed[45]. A prior study by Barnes et al.[21] of US veterans aged ≥ 55 years seen from 2000 to 2003 and followed until 2012 showed that TBI was associated with a 60% higher risk of developing dementia during the follow-up period compared to those without TBI. On multivariable analyses, our data did not show that TBI was an independent positive or negative predictor of MCI to ACD conversion, although univariate analysis showed fewer TBI in non-converters versus ACD converters. This discrepancy is likely explained by the younger age of MCI patients with TBI versus those without, as age is the dominant risk factor for MCI to ACD conversion. Indeed, ACD converters age-matched with non-converters show no significant difference in proportion of TBI co-morbidity. It is possible that temporal changes in intensity of TBI screening and reporting within the VA health care system (that may lead to underestimation of TBI diagnosis frequency for older patients) could explain the difference in mean age of MCI patients with and without TBI co-morbidity and should be explored further when evaluating the modulating role of TBI in dementia. When our

TABLE 3 Cox proportional hazard model for real and synthetic data training sets (backward stepwise selection).

	A. Training Set real		B. Training set synthetic #1		C. Training set synthetic #2		D. Training set synthetic #3	
	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value
Age MCI DX, years		<0.001		<0.001		<0.001		<0.001
50–55	Ref		Ref		Ref		Ref	
55–60	1.53 (1.26–1.85)		1.53 (1.26–1.85		1.51 (1.24–1.82)		1.53 (1.26–1.86)	
60–65	2.77 (2.34–3.27)[a]		2.84 (2.39–3.36)[a]		2.82 (2.38–3.34)[a]		2.86 (2.41–3.39)[a]	
65–70	4.16 (3.54–4.90)		4.28 (3.63–5.05)		4.21 (3.58–4.96)		4.31 (3.65–5.08)	
70–75	5.99 (5.09– 7.05)		6.19 (5.25–7.31)		6.07 (5.15–7.15)		6.18 (5.23–7.29)	
75–80	7.54 (6.41–8.87)		7.79 (6.61–9.19)		7.62 (6.47–8.98)		7.79 (6.60–9.19)	
80–85	8.36 (7.11–9.84)		8.71 (7.39–10.27)		8.48 (7.20–9.99)		8.67 (7.35–10.23)	
> 85	8.94 (7.60–10.53)		9.24 (7.83–10.90)		9.19 (7.78–10.86)		9.22 (7.80–10.88)	
Race		0.28		0.24		0.27		0.27
Asian/Pacific[a]	0.86 (0.73–1.01)		0.86 (0.73–1.01)		0.86 (0.73–1.01)		0.86 (0.73–1.01)	
Black	1.02 (0.96–1.08)		1.02 (0.96–1.08)		1.02 (0.96– 1.08)		1.02 (0.96– 1.08)	
Native American	0.91 (0.70–1.18)		0.90 (0.69– 1.17)		0.90 (0.69–1.17)		0.90 (0.69– 1.18)	
Other[b]	1.03 (0.96–1.11)		1.03 (0.96–1.08)		1.03 (0.96–1.10)		1.03 (0.96–1.10)	
White	Ref		Ref		Ref		Ref	
Ethnicity		0.11		0.11		0.13		0.13
Not Hispanic or Latino	Ref		Ref		Ref		Ref	
Hispanic or Latino	1.00 (0.93–1.09)		1.01 (0.93–1.09)		1.01 (0.93–1.09)		1.00 (0.93–1.09)	
Other[c]	0.91 (0.83–1.00)		0.91 (0.83–1.00)		0.91 (0.83–1.00)		0.91 (0.83–1.00)	
Sex		0.85		0.95		0.90		0.90
Female	0.99 (0.89–1.10)		1.00 (0.90–1.11)		0.99 (0.89–1.11)		1.00 (0.89–1.11)	
Male	Ref		Ref		Ref		Ref	
BMI		<0.001		<0.001		<0.001		<0.001
Underweight	0.87 (0.72–1.06)		0.87 (0.72–1.06)		0.87 (0.71–1.05)		0.86 (0.71–1.05)	
Normal	Ref		Ref		Ref		Ref	
Overweight	0.87 (0.83–0.91)		0.87 (0.83–0.91)		0.87 (0.83–0.91)		0.87 (0.83–0.91)	
Obese	0.75 (0.71–0.79)		0.74 (0.70–0.78)		0.75 (0.71–0.80)		0.75 (0.71–0.79)	
Co-morbidities								
Cerebrovascular disease (no stroke)	1.06 (1.01–1.12)	0.03	1.06 (1.00–1.11)	0.04	1.06 (1.00–1.12)	0.03	1.06 (1.00–1.12)	0.03
Stroke	1.07 (1.01–1.15)	0.05	1.07 (1.00–1.11)	0.05	1.08 (1.00–1.15)	0.04	1.07 (1.00–1.15)	0.04
Myocardial infarction	1.09 (1.03–1.16)	0.003	1.09 (1.02–1.15)	0.006	1.10 (1.03–1.17)	0.002	1.09 (1.03–1.16)	0.004
Hypertension	1.08 (1.02–1.14)	0.005	1.07 (1.02–1.13)	0.01	1.07 (1.02–1.13)	0.01	1.08 (1.02–1.14)	0.008
Diabetes	1.06 (1.02–1.10)	0.005	1.05 (1.01–1.10)	0.02	1.06 (1.01–1.10)	0.009	1.05 (1.01–1.10)	0.01
Alcohol abuse	0.94 (0.89–0.99)	0.02	0.93 (0.88–0.99)	0.01	0.93 (0.88–0.98)	0.007	0.94 (0.88–0.98)	0.01
Sleep apnea	0.95 (0.91–1.00)	0.06	[d]		0.95 (0.91–1.00)	0.05	0.95 (0.91–1.00)	0.06
Liver disease	[d]		[d]		[d]		[d]	
Peripheral vascular disease	[d]		[d]		[d]		[d]	
Heart failure	[d]		[d]		[d]		[d]	
Renal disease	[d]		[d]		[d]		[d]	
Rheumatic disease	[d]		[d]		[d]		[d]	
Hyperlipidemia	[d]		[d]		[d]		[d]	

(Continues)

TABLE 3 (Continued)

	A. Training Set real		B. Training set synthetic #1		C. Training set synthetic #2		D. Training set synthetic #3	
	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value	Hazard ratio (95% CI)	*P* value
Peptic ulcer disease	[d]		[d]		[d]		[d]	
Atrial fibrillation	[d]		[d]		[d]		[d]	
Depression	[d]		[d]		[d]		[d]	
Hearing loss	[d]		[d]		[d]		[d]	
Traumatic brain injury	[d]		[d]		[d]		[d]	

Abbreviations: BMI, body mass index; CI, confidence interval; DX, diagnosis; MCI, mild cognitive impairment.
[a]Patients who self-identified as Asian or Native Hawaiian or other Pacific Islander.
[b]Patients who self-identified as multiracial, declined to answer, unknown by patient, or missing.
[c]Patients who self-identified as declined to answer, unknown by patient, or missing.
[d]Not applicable due to variable being removed from final Cox proportional hazards model by selection procedure.

TABLE 4 Performance in ACD prediction at 5 years on real data test set.

	A. Training set real	B. Training set synthetic #1	C. Training set synthetic #2	D. Training set synthetic #3
Time-dependent AUC (95% CI)	0.73 (0.72–0.74)	0.73 (0.72–0.74)	0.73 (0.72–0.74)	0.73 (0.72–0.74)
Time-dependent AUC comparisons[a], (difference) [*P*-value]	Ref	(< 0.001) [*P* = 0.79]	(< 0.001) [*P* = 0.88]	(< 0.001) [*P* = 0.83]
Time-dependent brier (95% CI)	0.18 (0.17–0.18)	0.18 (0.17–0.18)	0.18 (0.17–0.18)	0.18 (0.17–0.18)
Brier score comparisons[a], (difference) [*P*-value]	Ref	(< 0.001) [*P* = 0.68]	(< 0.001) [*P* = 0.92]	(< 0.001) [*P* = 0.73]
Prediction expected conversion probability, median (IQR)	22.52% (27.61)	22.45% (27.74)	22.45% (27.86)	22.46% (27.72)
Correlation of expected conversion probability	Ref	0.99	0.99	0.99

Abbreviations: ACD, all-cause dementia; AUC, area under the receiving operator characteristic; CI, confidence interval; IQR, interquartile range.
[a]Absolute value of real minus synthetic.

findings are put in the context of the findings of Barnes et al.,[21] our data suggest that once a patient has MCI, TBI status is no longer a modulator of conversion to ACD within 5 years.

We previously showed that EHR-derived diagnosis of AD performed well against rigorously adjudicated AD diagnosis from the Michigan Alzheimer's Disease Research Center[11] and that EHR blood pressure trajectory records from two large health-care systems could be used to predict AD[12]. Using only demographic and co-morbidity conditions based on ICD-9/10 codes, our model showed good predictive performance for MCI to ACD conversion, demonstrating the feasibility of computational analyses on large-scale datasets without need for labor-intensive chart review. However, the utility of EHR datasets in disease modeling remains limited as data access is restricted to local investigators authorized by institutional regulatory bodies to ensure patient privacy. This restricts access to the dataset by outside data scientists or computational resources that could handle the complex analyses using increasingly sophisticated machine learning approaches. Experience in genomics research and large-scale clinical trials demonstrates the advantages of sharing raw data for widely distributed analyses to develop new models and statistical methods, test reproducibility, and enhance rigor of scientific discoveries[46]. An "honest broker" system[47] whereby protected health information and clinical data are stored in separate storage systems to protect patient privacy is a potential solution to this issue, but this does not eliminate privacy risk and is associated with great logistical cost. Our results show for the first time that the MCI to ACD predictive model using a synthetic dataset derived from real patient data but not attributable to any specific patient (hence removing data privacy concerns), performed just as well as the model from real patient data. The implication of this finding is that EHR-based synthetic datasets can potentially be made available for widely distributed computing to the scientific community, which could accelerate scientific discoveries. Models from synthetic data derived by outside scientists must then be validated using real patient data by investigators with access to identified patient data to maintain information security and, importantly, verify clinical validity. Using synthetic data for model building could lower cost, reduce barriers to entry, ease external validation using datasets from multiple health-care systems, and facilitate hypotheses generation of disease mechanisms.

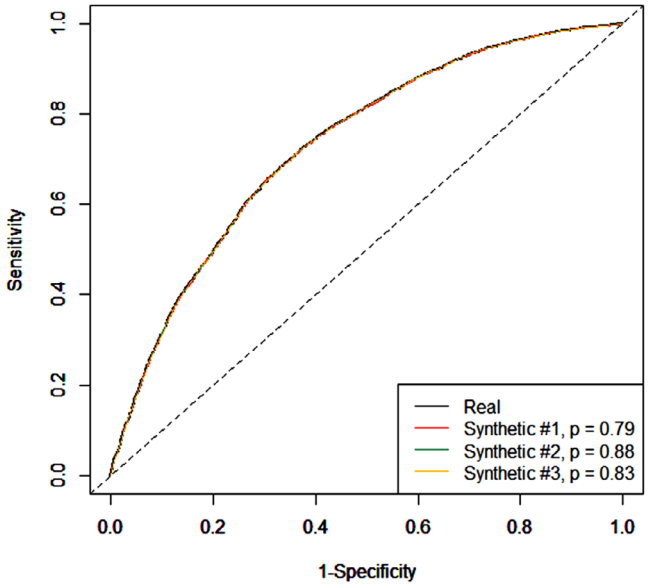

FIGURE 2 Real and synthetic data model performance. Area under the receiver operating characteristic curve (AUC) for prediction of ACD conversion within 5 years of MCI diagnosis using models trained on real and synthetic data. *P* value is comparing to AUC of real model. ACD, all-cause dementia; MCI, mild cognitive impairment

Various health systems are already using synthetic datasets for quality improvement and medical research.[13,14,27,48]

A study limitation is the predominance of male and White subjects in this cohort with potentially greater exposure to traumatic brain injury and post-traumatic stress disorder in combat veterans and applicability to a more diverse patient population or other health-care systems needs empirical testing. Our prior study showed that the performance of a machine learning model to predict AD onset using blood pressure trajectories trained using VA EHR data was similar when applied to University of Michigan EHR data even though the demographic compositions are different[12]. Although we used at least two encounters with ICD codes for MCI that the MVP Cog Working Group validated to have 95% specificity based on rigorous chart review[15], a recent study on VA patients showed that deriving MCI and AD diagnosis using clinical notes captured more MCI and AD cases versus ICD-based codes alone[49] so the model should be validated in the future using clinics' note-based diagnostic classification. The study is restricted to demographic and co-morbid conditions and adding data elements easily extracted from EHR such as vital signs, medication history, procedures codes, and others could further improve the model. The risk/protective factors identified represent associational and not necessarily causal relationships with ACD conversion. Associational relationships may contain spurious correlations, such as those from collider bias. Investigating whether EHR data have the potential to provide evidence for causal relationships between features of interest and ACD conversion is a topic for future work. We used a linear model and whether synthetic datasets perform as well as real patient datasets in non-linear models remains to be determined. In similar fashion, the associational nature of our findings does not imply causation and whether synthetic data can replicate real data in establishing causal relationships requires future empiric testing and validation. The decision to model ACD instead of specific type (such as AD) was made in light of the known difficulty in distinguishing among various dementia syndromes given the overlap of many common clinical features[39,50,51], the heterogeneity of expertise among clinical providers in a large health-care system making the ICD diagnosis decision, and the heterogeneity of intensity of diagnostic workup leading to dementia diagnosis. As such, identified at-risk individuals using the model will require further clinical and laboratory phenotyping to assess candidacy for clinical trials or specific interventions.

In conclusion, an EHR-derived model predicts MCI to ACD conversion at 5 years with good discriminative performance and calibration. The predictive model performance is similar when using real patient data versus synthetic data derived from real patient data. EHR-based prediction models could be used to identify high-risk MCI patients for early treatment interventions or clinical trial participation.

ACKNOWLEDGMENTS

Funding was provided by the Phoenix VA Office of Research and National Science Foundation (NSF award no. IIS 2124127). We would like to thank Gail Farrell for administrative help. The content and views do not represent the views of the VA, NSF, or the United States government.

CONFLICT OF INTEREST STATEMENT

VA is an employee of MDClone; there are no additional conflicts to declare from the other authors. Author disclosures are available in the supporting information.

CONSENT STATEMENT

Waiver of informed consent from human subjects were requested and granted by the Phoenix Veterans Affairs Institutional Review Board.

REFERENCES

1. Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol*. 1999;56(3):303-308.
2. Petersen RC, Parisi JE, Dickson DW, et al. Neuropathologic features of amnestic mild cognitive impairment. *Arch Neurol*. 2006;63(5):665-672.
3. Sabbagh MN, Shah F, Reid RT, et al. Pathologic and nicotinic receptor binding differences between mild cognitive impairment, Alzheimer disease, and normal aging. *Arch Neurol*. 2006;63(12):1771-1776.
4. Markesbery WR, Schmitt FA, Kryscio RJ, Davis DG, Smith CD, Wekstein DR. Neuropathologic substrate of mild cognitive impairment. *Arch Neurol*. 2006;63(1):38-46.
5. Schneider JA, Arvanitakis Z, Leurgans SE, Bennett DA. The neuropathology of probable Alzheimer disease and mild cognitive impairment. *Ann Neurol*. 2009;66(2):200-208.
6. Wilson RS, Leurgans SE, Boyle PA, Schneider JA, Bennett DA. Neurodegenerative basis of age-related cognitive decline. *Neurology*. 2010;75(12):1070-1078.
7. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. *N Engl J Med*. 2023;388(1):9-21.
8. Dukart J, Sambataro F, Bertolino A. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. *J Alzheimers Dis*. 2016;49(4):1143-1159.

9. Llano DA, Bundela S, Mudar RA, Devanarayan V. Alzheimer's disease neuroimaging I. A multivariate predictive modeling approach reveals a novel CSF peptide signature for both Alzheimer's disease state classification and for predicting future disease progression. *PLoS One*. 2017;12(8):e0182098.
10. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF. Alzheimer's disease neuroimaging i. multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci Rep*. 2018;8(1):5697.
11. Tjandra D, Migrino RQ, Giordani B, Wiens J. Cohort discovery and risk stratification for Alzheimer's disease: an electronic health record-based approach. *Alzheimers Dement (N Y)*. 2020;6(1):e12035.
12. 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):2368-2372.
13. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health*. 2023;2(1):e0000082.
14. Foraker RE, Yu SC, Gupta A, et al. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open*. 2020;3(4):557-566.
15. Logue MW, Miller MW, Sherva R, et al. Alzheimer's disease and related dementias among aging veterans: examining gene-by-environment interactions with post-traumatic stress disorder and traumatic brain injury. *Alzheimers Dement*. 2022.
16. Doraiswamy PM, Leon J, Cummings JL, Marin D, Neumann PJ. Prevalence and impact of medical comorbidity in Alzheimer's disease. *J Gerontol A Biol Sci Med Sci*. 2002;57(3):M173-177.
17. Duthie A, Chew D, Soiza RL. Non-psychiatric comorbidity associated with Alzheimer's disease. *QJM*. 2011;104(11):913-920.
18. Santiago JA, Potashkin JA. The impact of disease comorbidities in Alzheimer's disease. *Front Aging Neurosci*. 2021;13:631770.
19. Stampfer MJ. Cardiovascular disease and Alzheimer's disease: common links. *J Intern Med*. 2006;260(3):211-223.
20. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40(5):373-383.
21. Barnes DE, Kaup A, Kirby KA, Byers AL, Diaz-Arrastia R, Yaffe K. Traumatic brain injury and risk of dementia in older veterans. *Neurology*. 2014;83(4):312-319.
22. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care*. 1998;36(1):8-27.
23. Saunders GH, Dillard LK, Zobay O, Cannon JB, Naylor G. Electronic health records as a platform for audiological research: data validity, patient characteristics, and hearing-aid use persistence among 731,213 U.S. veterans. *Ear Hear*. 2021;42(4):927-940.
24. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom*. 1974;19(6):716-723.
25. Blanche P, Proust-Lima C, Loubere L, Berr C, Dartigues JF, Jacqmin-Gadda H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*. 2015;71(1):102-113.
26. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*. 2019;20(2):347-357.
27. Thomas JA, Foraker RE, Zamstein N, et al. Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *J Am Med Inform Assoc*. 2022;29(8):1350-1365.
28. Hung H, Chiang CT. Estimation methods for time-dependent AUC models with survival data. *Can J Stat*. 2010;38(1):8-26.
29. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatr Scand*. 2009;119(4):252-265.
30. Takeda M, Tanaka T, Okochi M, Kazui H. Non-pharmacological intervention for dementia patients. *Psychiatry Clin Neurosci*. 2012;66(1):1-7.
31. Aisen PS, Jimenez-Maggiora GA, Rafii MS, Walter S, Raman R. Early-stage Alzheimer disease: getting trial-ready. *Nat Rev Neurol*. 2022;18(7):389-399.
32. Guerreiro R, Bras J. The age factor in Alzheimer's disease. *Genome Med*. 2015;7:106.
33. Licher S, Leening MJG, Yilmaz P, et al. Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. *Am J Psychiatry*. 2019;176(7):543-551.
34. de la Torre JC. Alzheimer disease as a vascular disorder: nosological evidence. *Stroke*. 2002;33(4):1152-1162.
35. de la Torre JC. Is Alzheimer's disease a neurodegenerative or a vascular disorder? Data, dogma, and dialectics. *Lancet Neurol*. 2004;3(3):184-190.
36. Iturria-Medina Y, Sotero RC, Toussaint PJ, Mateos-Perez JM, Evans AC. Alzheimer's disease neuroimaging I. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. *Nat Commun*. 2016;7:11934.
37. Park KW, Yoon HJ, Kang DY, Kim BC, Kim S, Kim JW. Regional cerebral blood flow differences in patients with mild cognitive impairment between those who did and did not develop Alzheimer's disease. *Psychiatry Res*. 2012;203(2-3):201-206.
38. Hirao K, Ohnishi T, Hirata Y, et al. The prediction of rapid conversion to Alzheimer's disease in mild cognitive impairment using regional cerebral blood flow SPECT. *Neuroimage*. 2005;28(4):1014-1021.
39. Migrino RQ, Karamanova N, Truran S, et al. Cerebrovascular medin is associated with Alzheimer's disease and vascular dementia. *Alzheimers Dement (Amst)*. 2020;12(1):e12072.
40. Emmerzaal TL, Kiliaan AJ, Gustafson DR. 2003-2013: a decade of body mass index, Alzheimer's disease, and dementia. *J Alzheimers Dis*. 2015;43(3):739-755.
41. Besser LM, Gill DP, Monsell SE, et al. Body mass index, weight change, and clinical progression in mild cognitive impairment and Alzheimer disease. *Alzheimer Dis Assoc Disord*. 2014;28(1):36-43.
42. Tyas SL. Alcohol use and the risk of developing Alzheimer's disease. *Alcohol Res Health*. 2001;25(4):299-306.
43. Luchsinger JA, Tang MX, Siddiqui M, Shea S, Mayeux R. Alcohol intake and risk of dementia. *J Am Geriatr Soc*. 2004;52(4):540-546.
44. Leng Y, McEvoy CT, Allen IE, Yaffe K. Association of sleep-disordered breathing with cognitive function and risk of cognitive impairment: a systematic review and meta-analysis. *JAMA Neurol*. 2017;74(10):1237-1245.
45. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2020;396(10248):413-446.
46. Vickers AJ. Making raw data more widely available. *BMJ*. 2011;342:d2323.
47. Boyd AD, Hosner C, Hunscher DA, Athey BD, Clauw DJ, Green LA. An 'Honest Broker' mechanism to maintain privacy for patient care and academic medical research. *Int J Med Inform*. 2007;76(5-6):407-411.
48. Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak*. 2019;19(1):44.
49. Aguilar BJ, Miller D, Jasuja G, et al. Rule-based identification of individuals with mild cognitive impairment or Alzheimer's disease using clinical notes from the United States veterans affairs healthcare system. *Neurol Ther*. 2023;12(6):2067-2078.

50. Beach TG, Adler CH, Sue LI, et al. Arizona study of aging and neurodegenerative disorders and brain and body donation program. *Neuropathology*. 2015;35(4):354-389.
51. Karantzoulis S, Galvin JE. Distinguishing Alzheimer's disease from other major forms of dementia. *Expert Rev Neurother*. 2011;11(11):1579-1591.

SUPPORTING INFORMATION

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

How to cite this article: Irwin C, Tjandra D, Hu C, et al. Predicting 5-year dementia conversion in veterans with mild cognitive impairment. *Alzheimer's Dement*. 2024;16:e12572. https://doi.org/10.1002/dad2.12572