PLOS ONE

RESEARCH ARTICLE

Identifying incident dementia by applying machine learning to a very large administrative claims dataset

Vijay S. Nori[1]*, Christopher A. Hane[1], David C. Martin[1], Alexander D. Kravetz[2], Darshak M. Sanghavi[1]

1 OptumLabs, Cambridge, MA, United States of America, **2** Devoted Health, Waltham, MA, United States of America

* vijay.nori@optum.com

OPEN ACCESS

Citation: Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM (2019) Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS ONE 14(7): e0203246. https://doi.org/10.1371/journal.pone.0203246

Editor: Kewei Chen, Banner Alzheimer's Institute, UNITED STATES

Received: August 12, 2018

Accepted: June 20, 2019

Published: July 5, 2019

Data Availability Statement: The data underlying the results of this study are third party data owned by OptumLabs and contain sensitive patient information; therefore the data is only available upon request. Interested researchers engaged in HIPAA compliant research may contact connected@optum.com for data access requests and to begin research collaborations with OptumLabs. These research collaborations require researchers to pay for rights to use and access the data. All interested researchers and partners of OptumLabs can access the data in the same

Abstract

Alzheimer's disease and related dementias (ADRD) are highly prevalent conditions, and prior efforts to develop predictive models have relied on demographic and clinical risk factors using traditional logistical regression methods. We hypothesized that machine-learning algorithms using administrative claims data may represent a novel approach to predicting ADRD. Using a national de-identified dataset of more than 125 million patients including over 10,000 clinical, pharmaceutical, and demographic variables, we developed a cohort to train a machine learning model to predict ADRD 4–5 years in advance. The Lasso algorithm selected a 50-variable model with an area under the curve (AUC) of 0.693. Top diagnosis codes in the model were memory loss (780.93), Parkinson's disease (332.0), mild cognitive impairment (331.83) and bipolar disorder (296.80), and top pharmacy codes were psychoactive drugs. Machine learning algorithms can rapidly develop predictive models for ADRD with massive datasets, without requiring hypothesis-driven feature engineering.

Introduction

As many as 35.6 million people lived with dementia worldwide in 2010 and those numbers are expected to double every 20 years to 115.4 million by 2050[1]. Within the United States, the annual number of incident cases is expected to more than double from 377,000 in 1995 to 959,000 yearly by 2050 [2] leading to a prevalence of 13.8 million [3]. The total health care and long-term care costs associated with dementia are expected to reach a historic high of over quarter of a trillion dollars in the US in 2018 [4]. Tools are needed to assist clinicians, public health workers, and epidemiologists in identifying individuals at risk for dementia and addressing this unfolding epidemic.

Although the upward trend in incidence of Alzheimer's disease and the ravaging effects it has on the subject and caregivers is well documented, there has been very little published work building predictive models which help with identifying people with high risk prior to the onset of the disease. Barnes, et al. [5] developed a late-life dementia risk index, using logistic regression model developed on a cohort of 3,375 people. The model had a cStatistic of 0.81 for

manner as the authors. Interested researchers can replicate the results of this study by following the protocol outlined in the Methods section.

Funding: Optum provided support in the form of salaries for authors [VSN, CAH, DCM, ADK, DMS], but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: The authors [VSN, CAH, DCM, ADK, DMS] are employees of Optum. However, this does not alter our adherence to PLOS ONE policies on sharing data and materials.

identifying people who were at risk of dementia in six years. The model used several important features such as demographics, cognitive scores, physical activity, results from magnetic resonance imaging (MRI) scans, social network behavior, etc. Exalto, et al. [6] developed a Cox proportional hazard model using 45 predictors with a cohort of 29,961 people with type 2 diabetes, using a combination of self-reported data, pharmacy fills for prescriptions, laboratory data, and hospitalization and outpatient records. The c-statistic for 10-year dementia risk on a validation cohort was reported as 0.74. While the results of both these studies are impressive, it is typically not possible to get such datasets including MRI scans, self-reported data, laboratory results, etc. on a large group of people. It is important to develop models using easily available data so that the model impact reaches a wider population. With rapid growth in the "baby boomer" population in the US and elsewhere requiring governments to spend billions of dollars on healthcare in an aging population, models need to be developed which can scale to large groups of people with available claims data, and can identify chronic conditions such as dementia before the onset of the disease. This early identification may assist in drug development and early treatment.

Machine learning algorithms have been used previously for developing predictive models on large administrative claims datasets using automatic feature selection. In a recent paper Rajkomar, et al. [7] developed a suite of models using machine learning algorithms on Electronic Health Records data for predicting tasks such as patient's final discharge diagnosis, 30-day unplanned re-admission, etc. with a cStatistic of 0.9 and 0.76, respectively. McCoy, et al. [8] developed a predictive model trained on medical and pharmacy claims for 473,049 people to identify those at risk for type 2 diabetes. The model with 48 variables had a cStatistic of 0.808.

Machine learning algorithms have the potential to use large datasets to rapidly develop predictive models without specific selection of predictor variables, allowing for automated selection of high value predictors from very large numbers of potential inputs. Such models can use hundreds or even thousands of input variables related to dementia and rapidly generate useful information for clinicians, patients, pharmaceutical companies, payers, and policy-makers. Such techniques can identify connections within large datasets and algorithms that are beyond the ken of even expert clinicians. Early applications have included developing predictive models of who might develop dementia based on risk scores incorporating clinical characteristics, lab tests, neuro-imaging, and neuropsychological testing. Less investigated has been the use of machine learning to identify incident cases of Alzheimer's disease and related dementias (ADRD) based on administrative claims data.

To that end, an effort was undertaken to develop and test a model which would predict incipient ADRD using machine learning and compare the performance of that model to previous models derived with traditional logistic regression techniques or based on individualized diagnostic testing.

Methods

Data set

This study utilizes data between 2001 and 2015 from the OptumLabs Data Warehouse (OLDW), [9] which includes a national de-identified dataset of more than 125 million privately insured individuals that is geographically and racially diverse, including individuals of all ages (including Medicare Advantage beneficiaries $\geq$65 years old) and from all 50 states, with greatest representation in the Midwest and South U.S. Census Regions [10]. OLDW provides full access to professional, facility, and outpatient prescription medication claims. Patient-identifying information was encrypted or removed from the study database prior to its

release to the study investigators, such that it is compliant with HIPAA and exempt from Institutional Review Board review.

Definition of outcome

The study outcome was a binary variable indicating a new diagnosis ADRD. The identification rules for diagnosed ADRD cases were developed based on past work [11], and extended in consultation with an Expert Advisory Panel consisting of clinicians and experts from academia. Individuals must have met at least one of the following criteria:

- a medical claim with ADRD diagnosis codes in any header position in an inpatient setting,
- a medical claim with an ADRD diagnosis code followed by another claim with an ADRD diagnosis code within 1 to 730 days; both claims can be in any setting, and the codes in any header position,
- a pharmacy claim for donepezil hydrochloride, galantamine hydrobromide, rivastigmine tartrate or tacrine hydrochloride, and
- a pharmacy claim for memantine hydrochloride along with a medical claim with an ADRD diagnosis code in any setting and any header position within 0 to 730 days. The confirmation with a diagnosis claim is required because memantine hydrochloride is also used as an augmentation therapy for anxiety disorders (OCD, ADHD, etc.) [12] as well as help slowing down the tolerance development to opioids [13].

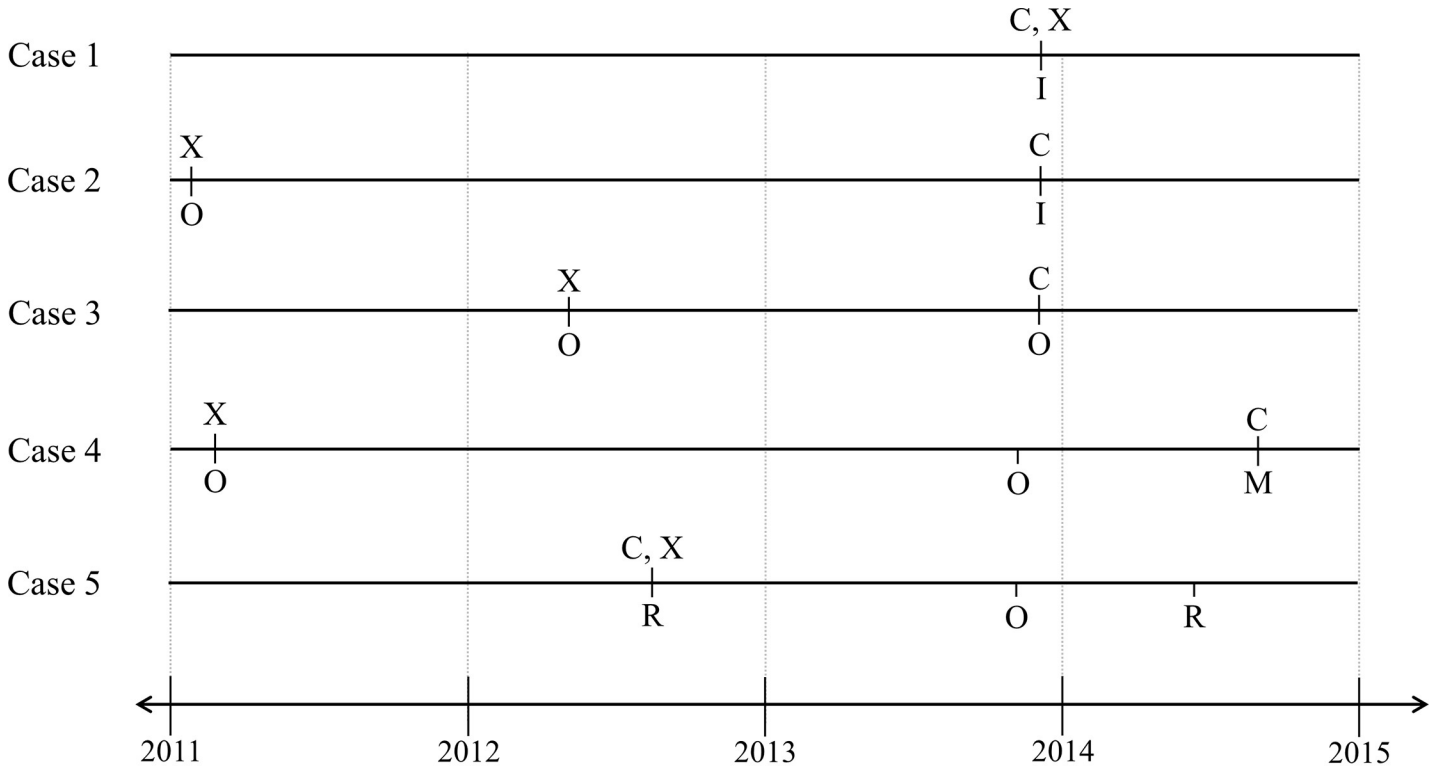
The index date for individuals (cases), whose diagnosis was confirmed using the above criteria, was set using the earliest occurrence of a medical claim with an ADRD diagnosis code in any setting or a pharmacy claim for any of the above drugs. To ensure that the identified individuals have incident rather than prevalent ADRD, we required a 60-month period of continuous enrollment without any of the above diagnosis or pharmacy claims before the index date. Fig 1 shows examples of how the index date and confirmation date are labeled in different situations with relevant inpatient, outpatient and pharmacy claims.

Training and testing

The ADRD predictive model was trained and tested using a nested case-control study design [8]. Step-wise demonstration of how the study population was assembled is depicted in Fig 2. OLDW contains 59,748,354 unique individuals with any period of medical and pharmacy coverage between 1/1/2007 and 6/30/2015. The end date was chosen so that the entire study could be done using medical claims before ICD-10 diagnosis codes were required. The start date was chosen to limit the population to a size amenable for processing on the available hardware. The final cohort size (over 200,000 training observations and nearly 600,000 in test) indicated that an earlier start date was not needed to have sufficient data.

We excluded 51,471,280 individuals who did not have at least one timespan with more than 5 years of continuous enrollment. Also, 740,363 individuals who are in long-term care facilities (LTC) were excluded, because LTC residence is strongly associated with pre-existing dementia [14] and also because dementia in LTC settings may be under-coded, leading the claims based identification algorithms to mark them incorrectly as cases or controls [15], [16]. This resulted in 8,257,557 eligible individuals of whom 238,336 individuals had a diagnosis of ADRD based on the outcome definitions described above.

In that ADRD cohort, we excluded 41,845 individuals who did not have a confirmatory diagnosis. We also excluded 150,127 individuals who had less than 5 years of continuous enrollment immediately prior to the index date within the current enrollment span or who



C	Diagnosis confirmation date	X	Index date
I	Claim with diagnosis from Table 1 in inpatient setting	M	Pharmacy claim for Memantine
O	Claim with diagnosis from Table 1 in outpatient setting	R	Pharmacy claim for Donepezil, Galantamine, Rivastigmine or Tacrine

Fig 1. Cases identified using different rules. Case 1 shows an individual who has an ADRD diagnosis in an inpatient setting and no previous relevant claims. So, the confirmation and index date are on the day of that claim. Case 2 has the same inpatient diagnosis claim and a previous claim in outpatient setting. So, the previous claim is used for as the index date, although it is over 730days prior to the claim in the inpatient setting. Case 3 has two claims in outpatient settings; the second claim is used as the confirmation and the first is used as the index date. Case 4 has a pharmacy claims for Memantine Hydrochloride and a diagnosis claim in an outpatient setting within 730 days. This case has a previous diagnosis claim in an outpatient setting which is used as the index date. Case 5 has multiple claims for Donepezil, Galantamin, Rivastigmin or Tacrin and the earliest of those is used for the confirmation and index dates.

https://doi.org/10.1371/journal.pone.0203246.g001

were less than 45 years old on the index date. To minimize bias from inclusion of individuals who had no encounters with the health care system, we excluded 1,419 people with no claims. These individuals may not be engaged in their health care management and would be poor participants in randomized controlled trials (RCT). Ultimately, there were 44,945 people in the incident ADRD cohort.

In order to minimize the risk of confounding based on duration of enrollment, the index dates for the controls cohort were selected to match the distribution of the case's enrollment duration (Prince *et al*., 2013). Specifically, for each person in the case cohort, we computed the time from enrollment to the index date, and divided them in percentiles 0% (1,827 days), 1% (1,832 days), . . ., 99% (2,719 days), 100% (2,735 days). Each person in the eligible controls population was randomly assigned one of these enrollment durations. This enrollment duration was added to their span start date to compute the index date for the controls. Out of the 8,019,191 people, these index dates for 3,687,767 individuals were outside their coverage dates and hence they were dropped from the control cohort. Another 1,166,663 people were

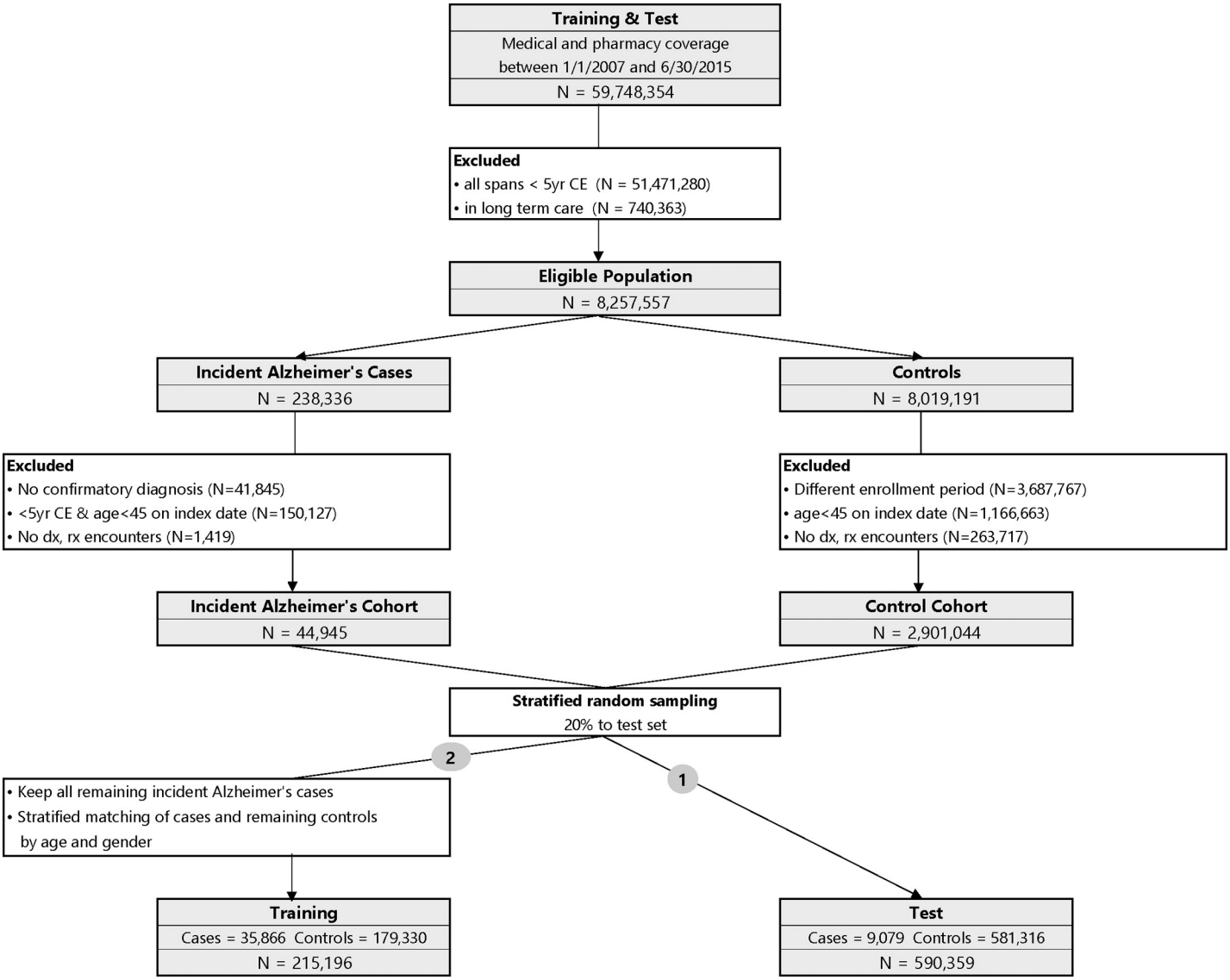


Fig 2. Study design with training and test cohorts.

https://doi.org/10.1371/journal.pone.0203246.g002

dropped because they were less than 45 years old on the index date. Additionally, 263,717 individuals were dropped from the controls since they had no claims. Finally, there were 2,901,044 people in the control cohort.

To assemble a test population, drawing from the incident ADRD and control cohorts, we randomly selected 30% of the available cases and enrollment-duration matched controls [17]; this yielded a population 9,079 cases and 581,316 controls (total 590,359). The above counts show that the original case/control cohort has about 1.5% cases. These type of two-class datasets (cases and controls) where one class constitutes a small minority of the entire data is called an imbalanced dataset [18]. Machine learning algorithms, like the one used in this work, do not accurately measure model performance when faced with unbalanced datasets [19]. They tend to predict the majority class data and the features in the minority class are treated as noise leading to mis-classification of the cases. The imbalanced dataset issue is handled by using a sample of the controls rather than all the controls. Such methods have been used previously in

other areas as reported by Brown & Mues [20]. King & Zeng [21] have presented a method for adjusting the weights in a logistic regression when one of the classes have been sampled.

The remaining 35,866 people within the ADRD incident cohort were assigned to the training population and were matched on age and gender along with enrollment duration to the 2,319,728 controls who are not in the test set so that each case could have up to 5 matched controls. The training dataset includes 215,196 individuals (35,866 cases and 179,330 controls). Thus the age/gender matching and selection of 5 matched controls implements an undersampling of the control class as the references cited above desire.

Independent variables

For training and test cohorts, we collected claims for individuals for the fourth and fifth year prior to the index date. While using just the fifth year would have resulted in a model which predicts four years prior to the onset of the disease, using just one year of claims data would not capture as much data on subjects with 13 months or more between encounters. Using claims in the fourth and fifth prior years leads to a dataset with more medical data per subject. Even with two years of data per subject in the model after feature selection, 90% of individuals had 4 or fewer unique medical codes. The claims included diagnoses (ICD-9 codes), Hierarchical Ingredient Code List (HICL) drug names, procedures (CPT codes for radiology [22]), and demographics (age and gender).

The diagnosis, drug and procedure codes were modeled using binary variables, with the value set to 1 if there was at least a single claim with a particular code. Age was modeled in ranges of 5 year increments from 40–44 to 85–89. Because of privacy concerns due to small numbers, ages greater than 89 years were mapped to 89 years. Using this process, the training model matrix had 10,363 clinical, pharmaceutical, and demographic variables (all binary) and 215,196 rows.

Analytic methods

We divided the analysis into two conceptual phases: a first phase that performs feature selection and a second phase that uses the best features to create a final model. In the first phase a Lasso logistic regression algorithm was run to identify the top 50 important predictors of the dependent variable. The Lasso algorithm outperforms other machine learning algorithms such as Random Forests and Regression Trees for variable selection [23]. Sensitivity tests were performed to show that using up to 500 variables did not demonstrably change the accuracy of the model.

The Lasso algorithm works by simultaneously maximizing the likelihood function (fitting the data well) and minimizing the sum of the absolute value of the coefficients (choosing a small set of features). The tradeoff between fitting the data and the number of features is managed with a regularization parameter, lambda, weighting the sum of the coefficient's absolute values. Setting lambda very high results in no feature selection–predicting only using the prevalence of the outcome–and gradually smaller lambda values allow additional features into the model. To evaluate which value of lambda works best, 4-way cross-validation was enabled so that each fit with n variables was evaluated against 3 overlapping data sets. A full regularization path is computed starting with highest values of lambda down to lowest values of lambda on a log scale [24]. The search for optimal lambda is an efficient approach to handle wide datasets with several features since it helps filter the noise and retain features with high predictive power. Since objective function used by the Lasso algorithm is a regularized version of Logistic Regression, the final list of features selected by Lasso will be a good set of predictors for the Logistic Regression algorithm. A final fit using logistic regression was used to fit all the training

data (removing the cross-validation) and remove the influence of lambda on the 50 final coefficients. The Lasso algorithm was run by applying the glm function [25] in the h2o package (R version 3.5.0) which was called with parameters "nfolds = 4, lambda_search = TRUE and max_active_predictors = 50." Logistic Regression was trained using the lrm function (a logistic regression algorithm) from the rms package [26]. The mathematical formulae and expressions used in these packages to obtain the optimal lambda are presented in [24], [27] and [28].

The 50-variable ADRD model was used to compute scores for each individual in the training and test datasets. Scores need to be converted into a threshold above which an action will be taken with the individual. A common method to choose the threshold is to set it so the fraction of scores above the threshold is equal to the case prevalence in the population. Using this method, the amount of outreach effort that would be expended toward intervening on a potentially "at risk" group is proportional to the prevalence. Scores which were at or above the threshold were classified as at-risk for ADRD, while scores which were below the threshold were classified as not at-risk. The sensitivity, specificity and lift of the model at this threshold for training and test populations were also computed. Lift [29] is a measure of model effectiveness at a threshold and is defined as the positive predictive value of this model divided by the case prevalence. For example, if the prevalence of a disease in a population is 1% then a *random* outreach effort would need to target 1000 people to get to 10 cases. However, if predictions from a model with a lift of 5 were used, one would need to target just 200 people to get to the 10 cases.

Because the prevalence and biological causes of dementia vary based on age, the classification into predicted outcome was repeated after stratifying the individual scores using three age ranges *viz*., 40–64, 65–74 and 75–89, and computing thresholds for each range based on the prevalence of cases in that range. The sensitivity, specificity and lift for each of the age ranges, as well as for an entire cohort using three thresholds were computed.

Results

Cohort characteristics

Baseline demographic and clinical characteristics of the training and test study population for cases and controls are summarized in Table 1. The age and gender-matched training population was on average 77.24 years old (SD 6.95) for cases and controls. The cases for the unmatched test set had a similar average age of 77.19 (SD 6.99), the unmatched controls were younger with an average age of 58.71 (SD 11.35).

Fig 3 shows the number of all codes in the model matrix versus the cumulative percentage of observations separately for cases and controls. This figure helps to illustrate the very sparse nature of the dataset because over 55% of the cases and 80% of the controls have fewer than 3 codes (if a case or control has 3 codes, then there will be 3 ones in the row for that observation and rest zeros). This shows that three years prior to the index data, there are only a few codes in claims for majority of the members in the cohort. The data sparsity makes this a very hard binary classification problem as evidenced further in the results below.

Model characteristics

The 50-variable ADRD model had a mean AUC (area under the receiver operating characteristic curve) of 64.26% within [63.88%, 64.58%] for 4-fold cross-validation, and 64.25% when refit to the entire training population. The AUC for the test data, which was matched only on enrollment duration and not on age and gender, was 69.3%. (The improved AUC on the test data is expected because the test data set is not age matched, making the test data set easier to predict after training on age-matched data.)

Table 1. Demographic and clinical profiles of cohorts.

	Training		Validation	
	Cases (N = 35,866)	Controls (N = 179,330)	Cases (N = 9,079)	Controls (N = 581,316)
Age, years, mean (SD)	77.24 (6.95)	77.24 (6.95)	77.19 (6.99)	58.71 (11.35)
Gender, male, N (%)	13,794 (38.46)	68,970 (38.46)	3,432 (37.80)	267,801 (46.07)
Diagnosis Codes, N(%)				
HYPERTENSION (401.9)	21,580 (60.17)	98,672 (55.02)	5,419 (59.69)	187,144 (32.19)
HYPERLIPIDEMIA (272.4)	17,229 (48.04)	82,314 (45.90)	4,357 (47.99)	191,833 (33.00)
HYPERCHOLESTEROLEMIA (272.0)	10,768 (30.02)	58,244 (32.48)	2,814 (30.99)	116,730 (20.08)
PAIN IN SOFT TISSUES OF LIMB (729.5)	10,112 (28.19)	42,702 (23.81)	2,566 (28.26)	91,039 (15.66)
OTHER MALAISE AND FATIGUE (780.79)	9,540 (26.60)	34,465 (19.22)	2,403 (26.47)	87,362 (15.03)
Procedure Codes N (%)				
RADIOLOGIC EXAMINATION, CHEST (71020)	13,915 (38.80)	61,303 (34.18)	3,518 (38.75)	14,3310 (24.65)
DIAGNOSTIC RADIOGRAPHIC PROC (76499)	12,759 (35.57)	64,037 (35.71)	3,235 (35.63)	205,253 (35.31)
SCREENING MAMMOGRAPHY (77052)	8,884 (24.77)	51,472 (28.70)	2,287 (25.19)	171,086 (29.43)
BONE DENSITY STUDY (77080)	6,815 (19.00)	38,011 (21.20)	1,746 (19.23)	85,077 (14.64)
CT, HEAD OR BRAIN (70450)	6,266 (17.47)	18,047 (10.06)	1,609 (17.72)	28,423 (4.89)
Drug Codes N (%)				
HYDROCODONE BIT/ACETAMINOPHEN	4,198 (11.70)	12,686 (7.07)	1,046 (11.52)	72,548 (12.48)
SIMVASTATIN	3,548 (9.89)	12,584 (7.02)	968 (10.66)	37,149 (6.39)
AZITHROMYCIN	3,221 (8.98)	12,561 (7.00)	823 (9.06)	69,553 (11.96)
LISINOPRIL	3,211 (8.95)	10,926 (6.09)	824 (9.08)	31,572 (5.43)
LEVOTHYROXINE SODIUM	2,663 (7.42)	9,623 (5.37)	682 (7.51)	29,602 (5.09)

https://doi.org/10.1371/journal.pone.0203246.t001

Table 2 shows the sensitivity and specificity of the model for when a single threshold was used for the entire training or test cohort. The threshold for the training set is 0.20 and is computed using the 16.67% prevalence of cases due to the 1:5 matching. The sensitivity, specificity and lift are 31.9%, 86.4% and 1.9, respectively. Using the lower prevalence of 1.54% in the unmatched test cohort, a new threshold of 0.37 is computed and the same metrics are now computed as 9.9%, 98.6% and 6.4, respectively. In this cohort, the lift indicates the model reaches 6.4 times more true cases than outreach to a similarly sized random group.

To enhance the model predictions, we chose thresholds that varied based on age-based prevalence. The performance can be computed as shown in Table 2. The table shows that the sensitivity increases with age for the test cohort. The sensitivity (specificity) for the test cohort increases from 9.9% (98.6%) to 16.4% (98.7%) by using the age-based thresholds. Thus the age specific thresholds increase the sensitivity 64% without a demonstrable change in the specificity.

Fig 4 shows a density plot of the distribution of scores for the cases and controls in the training and test cohorts, for the different groups of ages and for all ages. The dashed line indicates the threshold; all scores greater than or equal (lesser than) that line would be classified as cases (controls). These plots depict separation of the cases and controls at higher score range, with a large overlap in scores at the lower range.

Table 3 shows the coefficients, significance and variance inflation factor (VIF) for each variable in the model. The intercept term is the model constant representing the default risk for a male patient aged 65–69 with no other features. The maximum VIF for any variable is 1.46 indicating that the model has minimal collinearity. The top four diagnosis codes with a

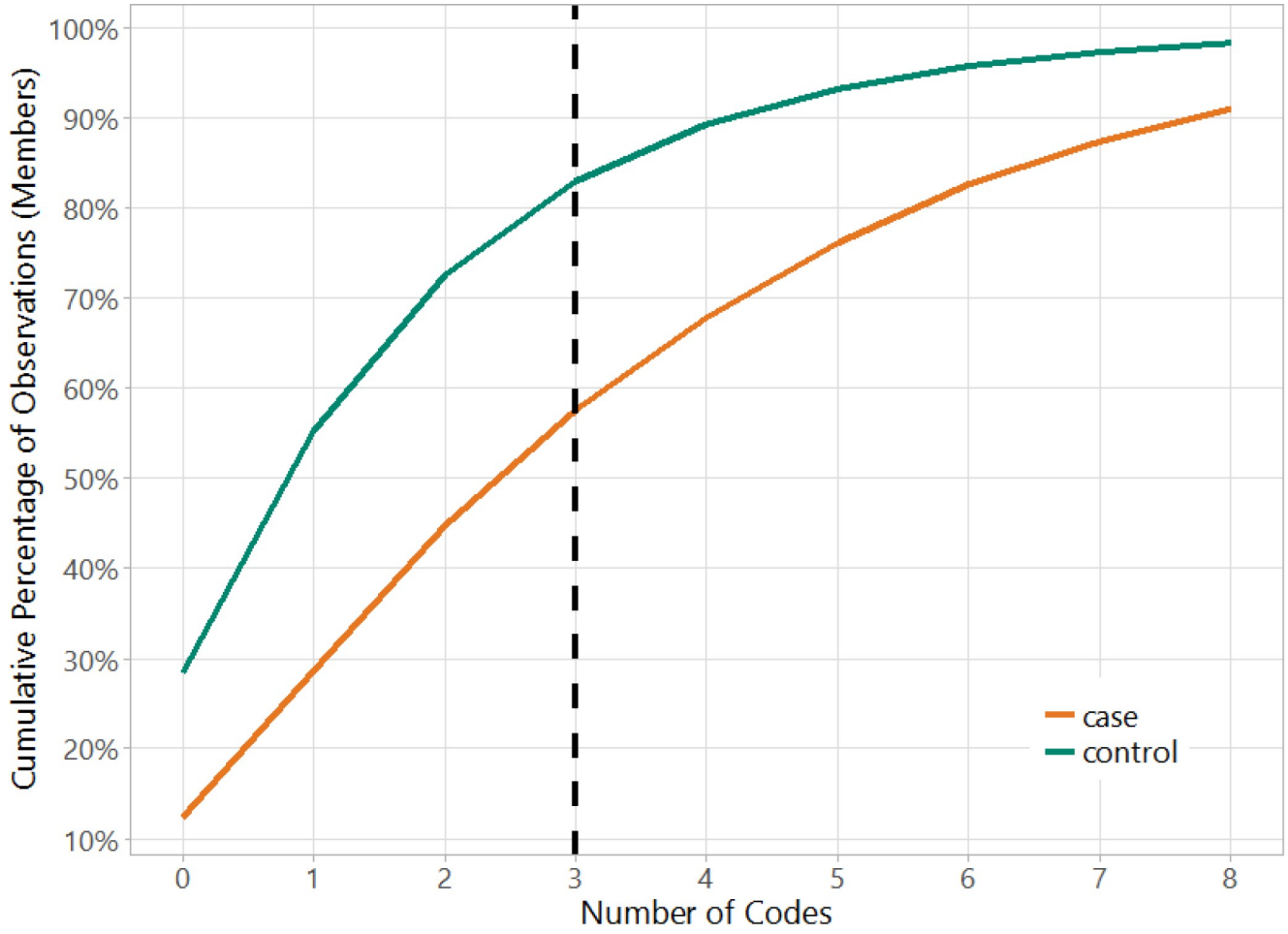


Fig 3. Data sparsity in the cohort. Over 55% of the cases and 80% of the controls have fewer than 3 codes.

https://doi.org/10.1371/journal.pone.0203246.g003

coefficient greater than 1 are Memory loss (780.93), Parkinson 's disease (332.0), Mild Cognitive Impairment (331.83) and Bipolar Disorder (296.80). The top 5 pharmacy codes are psychoactive drugs.

Table 2. Model sensitivity and specificity computed using a single threshold for the entire cohort based on the prevalence of the cases and for each age range based on the prevalence of the cases in those age ranges. Age and gender matching in training yields different prevalence and measures.

Cohort	Age	Prevalence	Threshold	Sensitivity	Specificity	Lift
Training	all	16.67%	0.20	31.9%	86.4%	1.9
	15–64	16.67%	0.18	41.8%	88.4%	2.5
	65–74	16.67%	0.19	39.1%	87.8%	2.3
	75–99	16.67%	0.21	29.6%	85.9%	1.8
	computed	16.67%		32.1%	86.4%	1.9
Test	all	1.54%	0.37	9.9%	98.6%	6.4
	15–64	0.14%	0.61	3.3%	99.9%	23.3
	65–74	1.77%	0.40	9.8%	98.4%	5.6
	75–99	8.46%	0.26	19.2%	92.5%	2.3
	computed	1.54%		16.4%	98.7%	10.7

https://doi.org/10.1371/journal.pone.0203246.t002

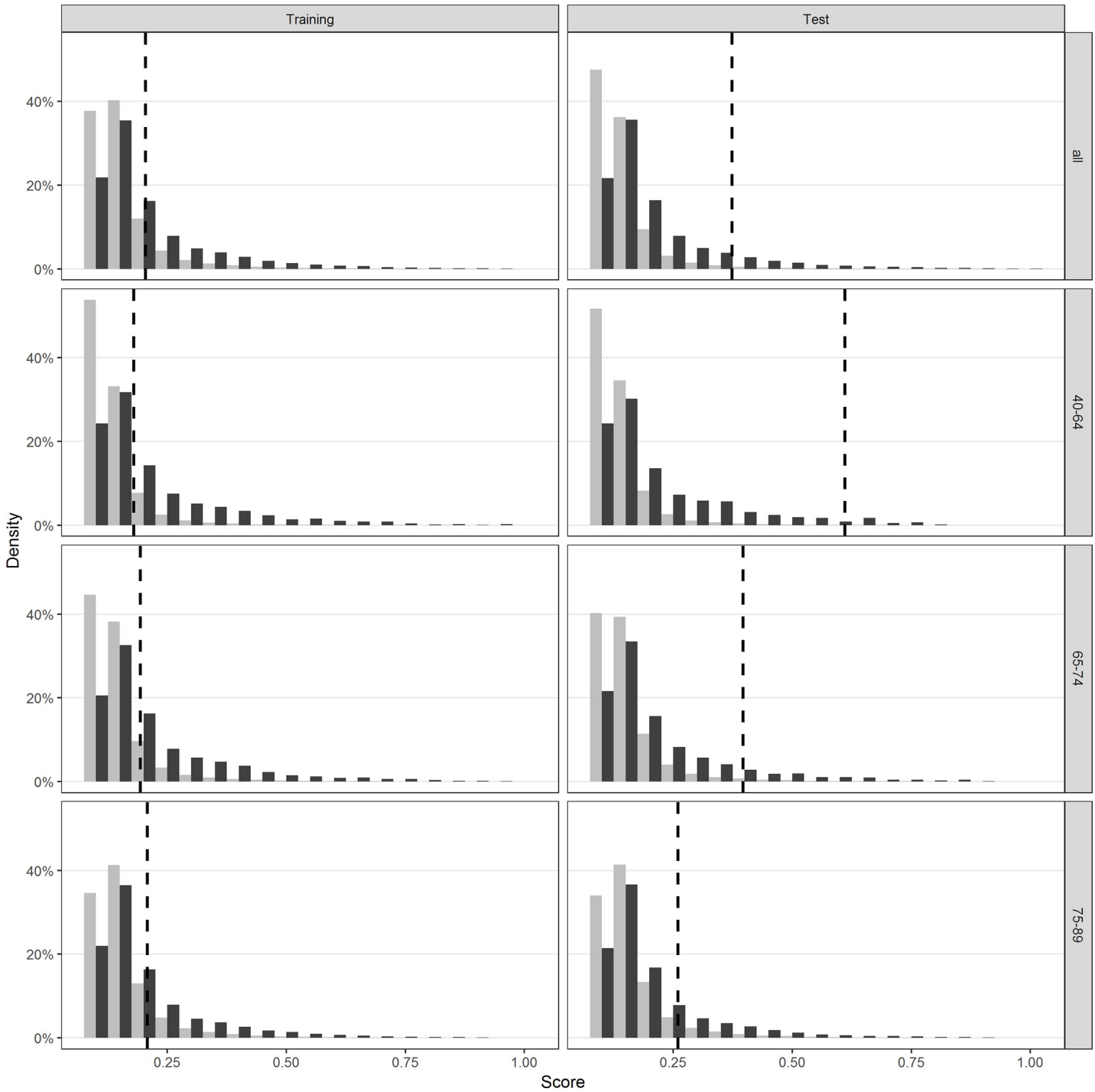


Fig 4. Distribution of scores for cases and controls for different age ranges. Figure shows substantial overlap in scores between cases and controls. The vertical lines show a proposed cut point for classification.

https://doi.org/10.1371/journal.pone.0203246.g004

Discussion

A recent systematic review of dementia risk prediction models [30] found models that could be grouped into five categories: (1) demographic factors only; (2) cognitive based (cognitive

Table 3. Clinical diagnosis, procedure and pharmacy variables included in the model, coefficients and variance inflation factors (VIF). The intercept has no VIF because it is a constant and does not vary across the observations due to age/gender matching. The specific ICD-9 codes for diagnosis and CPT-4 codes for procedures used to identify these variables are shown in parenthesis.

Type	Variable Description	Coefficient	Pr(>|Z|)	VIF
	INTERCEPT (MALE, 65–69)	-1.96	<0.0001	
ICD-9-CM	MEMORY LOSS (780.93)	1.33	<0.0001	1.02
	PARALYSIS AGITANS (332.0)	1.17	<0.0001	1.01
	MILD COGNITIVE IMPAIRMENT, SO STATED (331.83)	1.14	<0.0001	1.01
	BIPOLAR DISORDER, UNSPECIFIED (296.80)	1.00	<0.0001	1.01
	UNSPECIFIED PSYCHOSIS (298.9)	0.44	<0.0001	1.06
	LOSS OF WEIGHT (783.21)	0.42	<0.0001	1.02
	DEPRESSIVE DISORDER, NOT ELSEWHERE CLASSIFIED (311)	0.37	<0.0001	1.10
	ALTERED MENTAL STATUS (780.97)	0.33	<0.0001	1.16
	PERSONAL HISTORY OF FALL (V15.88)	0.32	<0.0001	1.05
	OTHER CONVULSIONS (780.39)	0.32	<0.0001	1.04
	UNSPECIFIED FALL (E888.9)	0.27	<0.0001	1.06
	OTHER CHRONIC PAIN (338.29)	0.28	<0.0001	1.03
	ACUTE, BUT ILL-DEFINED, CEREBROVASCULAR DISEASE (436)	0.24	<0.0001	1.33
	URGE INCONTINENCE (788.31)	0.24	<0.0001	1.07
	OTHER ALTERATION OF CONSCIOUSNESS (780.09)	0.23	<0.0001	1.09
	UNSPECIFIED CONSTIPATION (564.00)	0.21	<0.0001	1.05
	UNSPECIFIED URINARY INCONTINENCE (788.30)	0.21	<0.0001	1.08
	ENCOUNTER FOR LONG-TERM (CURRENT) USE OF OTHER MEDICATIONS (V58.69)	0.17	<0.0001	1.03
	LACK OF COORDINATION (781.3)	0.15	<0.0001	1.09
	OTHER MALAISE AND FATIGUE (780.79)	0.11	<0.0001	1.15
	DIABETES MELLITUS (250.02)	0.12	<0.0001	1.30
	ABNORMALITY OF GAIT (781.2)	0.10	<0.0001	1.20
	DIZZINESS AND GIDDINESS (780.4)	0.11	<0.0001	1.13
	UNSPECIFIED CEREBRAL ARTERY OCCLUSION WITH CEREBRAL INFARCTION (434.91)	0.11	0.0083	1.35
	DIABETES MELLITUS (250.00)	0.07	<0.0001	1.42
	EDEMA (782.3)	0.07	<0.0001	1.10
	MUSCLE WEAKNESS (GENERALIZED) (728.87)	0.06	0.0277	1.14
	URINARY TRACT INFECTION, SITE NOT SPECIFIED (599.0)	0.03	0.0526	1.12
CPT	SCREENING MAMMOGRAPHY; COMPUTER-AIDED DETECTION (77052)	-0.25	<0.0001	1.03
	COMPUTED TOMOGRAPHY, HEAD OR BRAIN (70450)	0.16	<0.0001	1.42
	RADIOLOGIC EXAMINATION, CHEST; SINGLE VIEW, FRONTAL (71010)	0.05	0.0011	1.28
Medication (HICL Description)	VENLAFAXINE HCL	0.52	<0.0001	1.02
	DULOXETINE HCL	0.45	<0.0001	1.05
	TOLTERODINE TARTRATE	0.32	<0.0001	1.07
	SERTRALINE HCL	0.29	<0.0001	1.05
	CITALOPRAM HYDROBROMIDE	0.25	<0.0001	1.06
	POTASSIUM CHLORIDE	0.24	<0.0001	1.35
	OXYBUTYNIN CHLORIDE	0.22	0.0001	1.08
	HYDROCODONE BIT/ACETAMINOPHEN	0.19	<0.0001	1.27
	PROPOXYPHENE/ACETAMINOPHEN	0.15	<0.0001	1.09
	SULFAMETHOXAZOLE/TRIMETHOPRIM	0.15	<0.0001	1.12
	METFORMIN HCL	0.13	0.0001	1.40
	BLOOD SUGAR DIAGNOSTIC	0.12	0.0011	1.38
	LISINOPRIL	0.11	<0.0001	1.17
	CEPHALEXIN MONOHYDRATE	0.10	0.0003	1.12
	SIMVASTATIN	0.10	<0.0001	1.17
	CLOPIDOGREL BISULFATE	0.09	0.0054	1.11
	TRAMADOL HCL	0.09	0.0097	1.13
	GABAPENTIN	0.07	0.0914	1.12
	FUROSEMIDE	0.00	0.9396	1.46

https://doi.org/10.1371/journal.pone.0203246.t003

test scores); (3) health variables and risk factors; (4) genetic risk scores; and (5) multi-variable models which combined demographic with health and lifestyle factors. Previously, machine

learning techniques have been primarily being applied to clinical data to successfully identify early cases of ADRD [31], to cluster patients into fast versus slow progression sub-types [32], to distinguish mild cognitive impairment or normal aging from early dementia [31], [33], and to assist in the interpretation and clinical significance of findings from neuro-imaging studies [34], [35], [36], [37], [38] [39]. Performing such work using administrative claims data may offer a larger pool for analyses and identification, since such claims are more widely available for large populations.

As noted earlier, a national de-identified dataset was used to train and test the models. While the census region-level information for members in the cohort is available, it was not used to control for potential regional differences in the model for the following reasons. First is that while it may help with adjusting for access to care, it will not help with addressing differences in coding behavior between providers, difference in insurance plan types, etc. Second, we wanted the model to be usable in a variety of production settings with claims coming in from different sources without the need for that additional information. This dataset has a combination of so many benefit plans, regions, etc. potential bias in any one of them would not effect the overall model.

The performance of this claims-based model with an AUC of 64% compares favorably with other models based more directly on clinical data collected about risk factors and health variables. The inclusion of a validation component in this study represents a refinement not often found in previous studies. The aforementioned meta-analysis of 21 papers on dementia risk prediction found only 4 with validation components [30]. AUC values in these papers ranged from 49% to 78%. These models used various non-claims based variables including genotype. These extra variables come at high cost and based on those published AUC values, do not always lead to better models.

A review of the medications and diagnosis codes that remained in the final model (Table 3) suggests that helpful inferences can be drawn from machine-learning derived predictors. For example, six variables (12% of the model) related to falls, dizziness, gait disorders, or weakness, suggesting a subacute phase of progression. Additionally, diagnoses and medications related to vascular disease and to diabetes mellitus were represented, which supports clinical data suggesting overlap of risk factors for cardiovascular disease with vascular dementia [40], [41] and an association of dementia risk with diabetes mellitus [42], [43], [44]. Psychiatric symptoms and psychoactive medications were also unsurprisingly present. Screening mammography proved to be the only code with a negative coefficient indicative of a protective effect. Because the application of screening procedures can be tempered by clinical judgment which incorporates such factors as life expectancy and quality of life, the presence of this screening test might be a marker for healthier and less impaired individuals who would therefore be at lower risk for developing dementia.

This model becomes increasingly useful as potential disease-modifying treatments for dementia are developed to a stage for clinical testing. Thus, the ability to achieve a lift of 6.4 means that a patient identified by the model will be 6.4 times more likely to be diagnosed in the near-term with dementia. An identified cohort with such enhanced prior probability could be much more cost-effectively screened for clinical research than an unselected population.

A limitation of our study is that dementia may be undercoded, presenting a challenge for training models; in one study, Alzheimer's disease and related dementias was recorded as a diagnosis for less than 25% of patients with moderate to severe cognitive impairment [45]; and in another, physicians were unaware of cognitive impairment in more than 40% of their cognitively impaired patients [46]. Among participants in a Medicare Alzheimer's Disease Demonstration, less than 20% of participants were classified with dementia of the Alzheimer type based on a year's worth of claims data, although 68% carried that diagnosis upon referral [47].

A review of seven studies examining the extent to which dementia is omitted as a cause of death, found that the reporting on death certificates ranged from a 7.2% to 41.8% [48].

Author Contributions

Conceptualization: David C. Martin, Alexander D. Kravetz, Darshak M. Sanghavi.

Data curation: Vijay S. Nori.

Formal analysis: Vijay S. Nori, Christopher A. Hane.

Investigation: Vijay S. Nori, Christopher A. Hane, Alexander D. Kravetz.

Methodology: Christopher A. Hane.

Supervision: Christopher A. Hane, Darshak M. Sanghavi.

Validation: Christopher A. Hane.

Visualization: Vijay S. Nori.

Writing – original draft: Vijay S. Nori, Christopher A. Hane, David C. Martin, Darshak M. Sanghavi.

Writing – review & editing: Vijay S. Nori, Christopher A. Hane, David C. Martin, Darshak M. Sanghavi.

References

1. Prince MJ, Bryce RM, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: A systematic review and metaanalysis. Alzheimer's & Dementia. 2013; 9(1):63–75. https://doi.org/10.1016/j.jalz.2012.11.007 PMID: 23305823
2. Hebert LE, Beckett LA, Scherr PA, Evans DA. Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050. Alzheimer Disease and Associated Disorders. 2001; 15(4):169–73. https://doi.org/10.1097/00002093-200110000-00002 PMID: 11723367
3. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology. 2013; 80(19):1778–83. https://doi.org/10.1212/WNL.0b013e31828726f5 PMID: 23390181
4. Alzheimer's Association. 2018 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2018; 14(3):367–429. https://doi.org/10.1016/j.jalz.2018.02.001
5. Barnes DE, Covinsky KE, Whitmer RA, Kuller LH, Lopez OL, Yaffe K. Predicting risk of dementia in older adults: The late-life dementia risk index. Neurology. 2009; 73(3):173–9. https://doi.org/10.1212/WNL.0b013e3181a81636 PMID: 19439724
6. Exalto LG, Biessels GJ, Karter AJ, Huang ES, Katon WJ, Minkoff JR, et al. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: A cohort study. The Lancet Diabetes & Endocrinology. 2013; 1(3):183–90. https://doi.org/10.1016/S2213-8587(13)70048-2 PMID: 24622366
7. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018; 1(1). https://doi.org/10.1038/s41746-018-0029-1
8. McCoy RG, Nori VS, Smith SA, Hane CA. Development and Validation of HealthImpact: An Incident Diabetes Prediction Model Based on Administrative Data. Health Services Research. 2016; 51(5):1896–918. https://doi.org/10.1111/1475-6773.12461 PMID: 26898782
9. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: Building A Novel Node In The Learning Health Care System. Health Affairs. 2014; 33(7):1187–94. https://doi.org/10.1377/hlthaff.2014.0038 PMID: 25006145
10. OptumLabs, OptumLabs Data Warehouse Technical Specifications. 2015. url:https://www.optum.com/content/dam/optum/resources/productSheets/5302_Data_Assets_Chart_Sheet_ISPOR.pdf
11. Taylor DH Jr, Østbye T, Langa KM, Weir D, Plassman BL. The Accuracy of Medicare Claims as an Epidemiological Tool: The Case of Dementia Revisited. Journal of Alzheimer's Disease. 2009; 17(4):807–15. https://doi.org/10.3233/JAD-2009-1099 PMID: 19542620

12. Zdanys K, Tampi RR. A systematic review of off-label uses of memantine for psychiatric disorders. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2008; 32(6):1362–74. https://doi.org/10.1016/j.pnpbp.2008.01.008 PMID: 18262702
13. Grande L, O'Donnell B, Fitzgibbon D, Terman G. Ultra-Low Dose Ketamine and Memantine Treatment for Pain in an Opioid-Tolerant Oncology Patient. Anesthesia & Analgesia. 2008; 107(4):1380–3. https://doi.org/10.1213/ane.0b013e3181733ddd PMID: 18806055
14. Harris-Kojetin L, Sengupta M, Park-Lee E, Valverde R, Caffrey C, Rome V, et al. Long-Term Care Providers and Services Users in the United States: Data From the National Study of Long-Term Care Providers, 2013–2014. Vital & Health Statistics. 2016; 3(38):1–105.
15. Boustani M, Zimmerman S, Williams CS, Gruber-Baldini AL, Watson L, Reed PS, et al. Characteristics Associated With Behavioral Symptoms Related to Dementia in Long-Term Care Residents. The Gerontologist. 2005; 45(suppl_1):56–61. https://doi.org/10.1093/geront/45.suppl_1.56 PMID: 16230750
16. Smith M, Buckwalter KC, Kang H, Ellingrod V, Schultz SK. Dementia Care in Assisted Living: Needs and Challenges. Issues in Mental Health Nursing. 2008; 29(8):817–38. https://doi.org/10.1080/01612840802182839 PMID: 18649209
17. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer. 2009.
18. Drummond C, Holte RC. C4.5, ClassImbalance, and CostSensitivity: Why Under-Sampling beats Over-Sampling. In Proceedings of the ICML'03 Workshop on Learning from Imbalanced Data Sets. 2003.
19. Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning. 1997.
20. Brown I, Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. 2012; 39(3):3446–53. https://doi.org/10.1016/j.eswa.2011.09.033
21. King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001; 9(2):137–63. https://doi.org/10.1093/oxfordjournals.pan.a004868
22. CPT 2014, Current Procedural Terminology, Professional Edition. American Medical Association. 2013.
23. Lu F, Petkova E. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. Statistics In Medicine. 2014; 33(3):401–21. https://doi.org/10.1002/sim.5937 PMID: 23934941
24. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33(1):1–22. PMID: 20808728
25. Nykodym T, Kraljevic T, Wang A, Wong W. Generalized Linear Modeling with H2O. 2018. url:https://CRAN.R-project.org/package=h2o.
26. Harrell FE Jr. Package 'rms'. 2018. url:https://CRAN.R-project.org/package=rms.
27. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of The Royal Statistical Society. 2005; 67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x
28. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. Journal of The Royal Statistical Society. 2012; 74(2):245–66. https://doi.org/10.1111/j.1467-9868.2011.01004.x PMID: 25506256
29. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier. 2011.
30. Tang EYH, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. PLOS One. 2015; 10(9). https://doi.org/10.1371/journal.pone.0136181 PMID: 26334524
31. So A, Hooshyar D, Park KW, Lim HS. Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques. Applied Sciences. 2017; 7(7):1–17. https://doi.org/10.3390/app7070651
32. Gamberger D, Lavrač N, Srivatsa S, Tanzi RE, Doraiswamy PM. Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. Scientific Reports. 2017; 7(1):1–12. https://doi.org/10.1038/s41598-016-0028-x
33. Shankle WR, Mani S, Pazzani MJ, Smyth P. Detecting very early stages of dementia from normal aging with Machine Learning methods. In Proceedings of the 6th Conference on Artificial Intelligence in Medicine Europe. 1997.
34. Adaszewski S, Dukart J, Kherif F, Frackowiak R, Draganski B. How early can we predict Alzheimer's disease using computational anatomy?. Neurobiology of Aging. 2013; 34(12):2815–26. https://doi.org/10.1016/j.neurobiolaging.2013.06.015 PMID: 23890839
35. Dyrba M, Barkhof F, Fellgiebel A, Filippi M, Hausner L, Hauenstein K, et al. Predicting Prodromal Alzheimer's Disease in Subjects with Mild Cognitive Impairment Using Machine Learning Classification of

Multimodal Multicenter Diffusion Tensor and Magnetic Resonance Imaging Data. Journal of NeuroImaging. 2015; 25(5):738–47. https://doi.org/10.1111/jon.12214 PMID: 25644739

36. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. Brain. 2008; 131(3):681–9. https://doi.org/10.1093/brain/awm319 PMID: 18202106

37. Lao Z, Shen D, Xue Z, Karacali B, Resnick S, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. NeuroImage. 2004; 21(1):46–57. PMID: 14741641

38. Li S, Shi F, Pu F, Li X, Jiang T, Xie S, et al. Hippocampal Shape Analysis of Alzheimer Disease Based on Machine Learning Methods. American Journal of Neuroradiology. 2007; 28(7):1339–45. https://doi.org/10.3174/ajnr.A0620 PMID: 17698538

39. Zhang Y, Dong Z, Phillips P, Wang S, Ji G, Yang J, et al. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Frontiers In Computational Neurosciene. 2015; 9(66):1–15. https://doi.org/10.3389/fncom.2015.00066 PMID: 26082713

40. Adelborg K, Horváth-Puhó E, Ording A, Pedersen L, Sørensen HT, Henderson VW. Heart failure and risk of dementia: a Danish nationwide population-based cohort study. European Journal of Heart Failure. 2017; 19(2):253–60. https://doi.org/10.1002/ejhf.631 PMID: 27612177

41. Rusanen M, Kivipelto M, Levälahti E, Laatikainen T, Tuomilehto J, Soininen H, et al. Heart Diseases and Long-Term Risk of Dementia and Alzheimer's Disease: A Population-Based CAIDE Study. Journal of Alzheimer's Disease. 2014; 42(1):183–91. https://doi.org/10.3233/JAD-132363 PMID: 24825565

42. Abner EL, Nelson PT, Kryscio RJ, Schmitt FA, Fardo DW, Woltjer R, et al. Diabetes is associated with cerebrovascular but not Alzheimer's disease neuropathology. Alzheimer's and Dementia. 2016; 12(8):882–9. https://doi.org/10.1016/j.jalz.2015.12.006 PMID: 26812281

43. Haan MN. Therapy Insight: type 2 diabetes mellitus and the risk of late-onset Alzheimer's disease. Nature Clinical Practice Neurology. 2006; 2(3):159–66. https://doi.org/10.1038/ncpneuro0124 PMID: 16932542

44. Pugazhenthi S, Qin L, Reddy PH. Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease. Biochimica et Biophysica Acta (BBA)—Molecular Basis of Disease. 2017; 1863(5):1037–45. https://doi.org/10.1016/j.bbadis.2016.04.017 PMID: 27156888

45. Callahan CM, Hendrie HC, Tierney WM. Documentation and Evaluation of Cognitive Impairment in Elderly Primary Care Patients. Annals of Internal Medicine. 1995; 122(5):422–9. https://doi.org/10.7326/0003-4819-122-6-199503150-00004 PMID: 7856990

46. Chodosh J, Petitti DB, Elliott M, Hays RD, Crooks VC, Reuben DB, et al. Physician Recognition of Cognitive Impairment: Evaluating the Need for Improvement. Journal of The American Geriatrics Society. 2004; 52(7):1051–9. https://doi.org/10.1111/j.1532-5415.2004.52301.x PMID: 15209641

47. Newcomer R, Clay T, Luxenberg JS, Miller R. Misclassification and Selection Bias When Identifying Alzheimer's Disease Solely from Medicare Claims Records. Journal of the American Geriatrics Society. 1999; 47(2):215–9. https://doi.org/10.1111/j.1532-5415.1999.tb04580.x PMID: 9988293

48. Romero JP, Benito-León J, Mitchell AJ, Trincado R, Bermejo-Pareja F. Under Reporting of Dementia Deaths on Death Certificates using Data from A Population-Based Study (NEDICES). Journal of Alzheimer's Disease. 2014; 39(4):741–8. https://doi.org/10.3233/JAD-131622 PMID: 24254704