




# Paper Review





# Contents

- 01 Machine learning models to predict onset of dementia: A label learning approach
  - 02 Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study
- 

# Machine learning models to predict onset of dementia: A label learning approach



## Alzheimer's & Dementia: Translational Research & Clinical Interventions

Volume 5, 2019, Pages 918-925



"Alzheimers Dement (N Y) (IF: 6.34; Q1) . 2019 Dec 10:5:918-925. doi: 10.1016/j.trci.2019.10.006. eCollection 2019."

# Introduction

- ✓ Importance of early dementia prediction
- ✓ Early intervention, clinical trial recruitment
- ✓ Existing models rely on clinical data (AUC 0.60-0.78)
- ✓ Label inaccuracies (undercoding, miscoding)

# Objective

- ✓ Build a dementia risk prediction model using **large-scale claims + EHR data**
- ✓ Apply **Label Learning** to refine inaccurate case/control labels



Our hypothesis is that combining **large-scale complex data**, **sophisticated machine learning techniques**, and **label learning** will significantly enhance the performance of dementia prediction algorithms

# Data & Cohort

- ✓ Data Source: OptumLabs Data Warehouse (2007-2017)
  - Claims, EHR, demographic data
  - 45+ years old, 2-year clean period
- ✓ Cohorts: Claims-Only, SEHR, Open-World (OW-C, OW-E, OW-M)

# MACHINE LEARNING MODELS TO PREDICT ONSET OF DEMENTIA: A LABEL LEARNING APPROACH

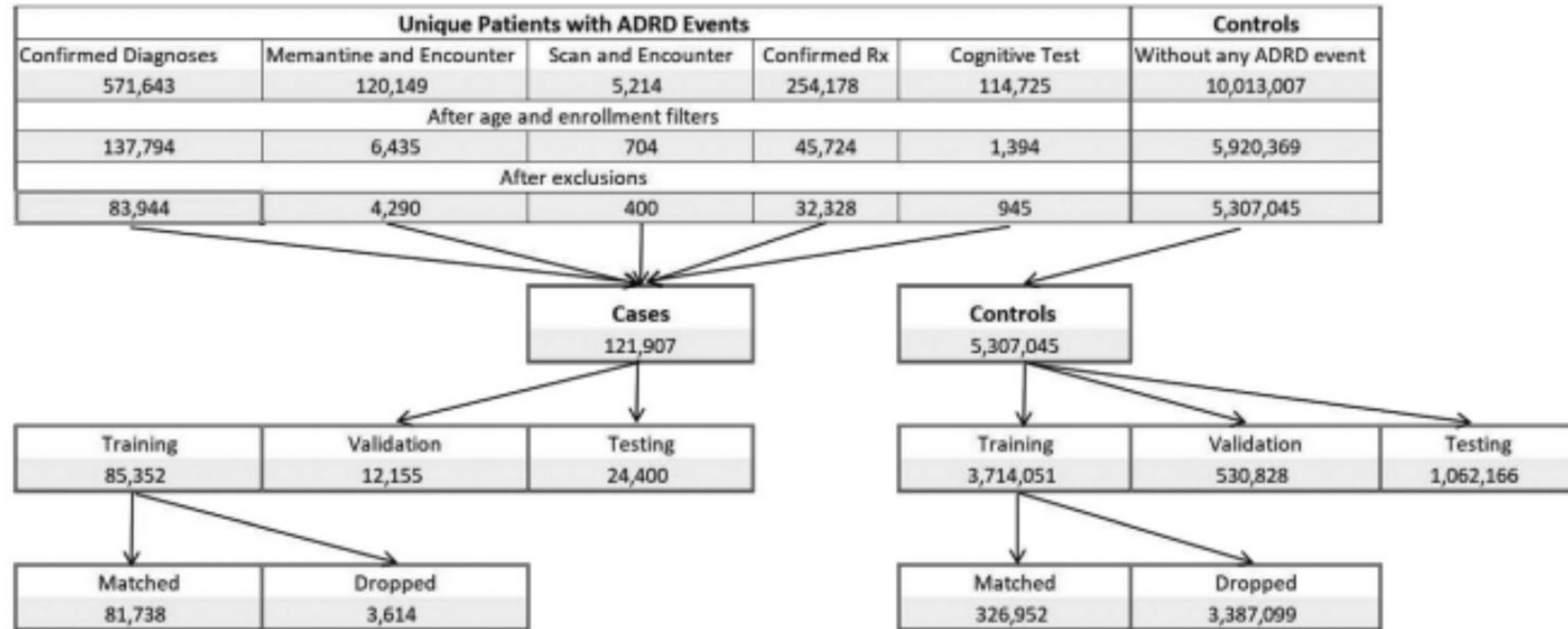


Fig. 1. Attrition of the two-year cohort into the training, validation and test data.

Table 1  
Data source sample sizes and summary statistics

Cohort	Subset	N	Age mean (SD)	Encounters mean (SD)	Case prevalence, %	Female, %	Cardiovascular disease prevalence, %	Mood disorder prevalence, %
Claims	ClaimsOnly	5,640,637	60.0 (10.7)	10.7 (21.6)	2.1	52.8	46.2	14.6
SEHR	ClaimsOnly	4,810,730	59.8 (10.6)	10.6 (21.0)	2.1	52.3	45.1	13.9
SEHR	Mixed	609,578	61.7 (11.3)	11.3 (29.4)	3.5	56.2	55.1	19.3
Open-World	ClaimsOnly	8,348,496	60.4 (10.7)	10.7 (24.1)	2.7	54.7	43.9	14.4
Open-World	EHROnly	7,276,426	62.6 (11.4)	11.4 (17.4)	3.7	59.0	34.2	14.1
Open-World	Mixed	1,602,898	60.6 (10.7)	10.7 (27.4)	4.1	57.6	47.7	19.1



# Method Overview

- ✓ Two-Stage Label Learning Approach
- ✓ Step 1: SEHR-based Label Learning Model

- Propensity score calculation
- Reclassification using IQR (25%-75%)

## Step 2: Prediction Model

- Applying learned labels to Claims-Only & Open-World cohorts
- ✓ Model algorithm : LightGBM (Gradient Boosting)

# Fitting Model

- ✓ All data was trained as a case/control classification model using LightGBM.
- ✓ Hyperparameter Tuning
  - Feature Fraction: 0.25, 0.2, 0.15
  - Learning Rate: 0.015, 0.01, 0.02
  - Minimum Data in Leaf: 1000, 800, 500
  - Number of Trees: 300
  - Tree Size: 127, 63
- ✓ Optimization Goals
  - Maximize Sensitivity while preventing Overfitting
  - Ensure Generalizability → Performance difference between training and validation data < 1%
  - Prevent excessive tuning that may degrade real-world performance

# Label Learning

- ✓ Why is Label Learning necessary?
  - Diagnostic errors (undercoding, miscoding)
  - Case/control misclassification
- ✓ Process:
  - SEHR data → Similarity score calculation → Reclassification
- ✓ Model Implementation
  - Expand Feature Space → Short-term (60 days) + Long-term (730 days) data
  - Train model only with SEHR data (Excluding Claims & Open-World data)

# Label Learning

Table 2  
Label Learning Model results by age group

Age group	Sensitivity	AUC	Lift	True positives	False positives	True negatives	False negatives	Case, %	Case count	Total count
45,55	0.29	0.89	94.0	403	977	444,939	977	0.31	1380	447,296
55,60	0.34	0.90	63.8	325	622	175,999	622	0.53	947	177,568
60,64	0.39	0.90	48.8	374	583	117,945	583	0.80	957	119,485
64,70	0.38	0.88	24.0	844	1396	139,148	1396	1.57	2240	142,784
70,75	0.43	0.85	10.7	1499	1998	81,507	1998	4.02	3497	87,002
75,80	0.49	0.83	5.2	2722	2818	49,961	2818	9.50	5540	58,319
80,99	0.53	0.81	2.9	5205	4634	39,639	4634	18.18	9839	54,112
Summary	0.47	0.87	20.9	11,372	13,028	1,049,138	13,028	2.25	24,400	1,086,566

Abbreviation: AUC, area-under-the-curve.

# Label Learning

Table 3  
Comparison of onset model quality for original versus learned labels

Original labels					Learned labels				
Prediction threshold	Sensitivity of ADRD	Specificity of ADRD	Positive predictive value of ADRD	Proportion of cohort over threshold	Prediction threshold	Sensitivity of ADRD	Specificity of ADRD	Positive predictive value of ADRD	Proportion of cohort over threshold
Choosing by threshold greater than									
0.75	0.060	1.000	0.857	0.002	0.75	0.075	1.000	1.000	0.002
0.50	0.180	0.998	0.681	0.006	0.50	0.283	1.000	1.000	0.006
0.20	0.388	0.987	0.405	0.021	0.20	0.619	0.991	0.604	0.021
Choosing by sensitivity									
0.102	0.50	0.971	0.282	0.040	0.328	0.50	0.999	0.926	0.011
0.007	0.90	0.545	0.043	0.465	0.040	0.90	0.921	0.196	0.096
0.004	0.95	0.325	0.031	0.681	0.031	0.95	0.893	0.160	0.124
Choosing by specificity									
0.064	0.572	0.95	0.209	0.061	0.061	0.830	0.95	0.262	0.066
0.223	0.353	0.99	0.448	0.018	0.189	0.634	0.99	0.576	0.023
0.610	0.128	0.999	0.747	0.004	0.326	0.503	0.999	0.916	0.012

Abbreviation: ADRD, Alzheimer's disease or related dementias.

# Cohorts & Time Windows

Table 4  
Sensitivity (area-under-the-curve) scores over different time windows

Time window	Outcome label	SEHR	OW-C	OW-E	OW-M	Claims
Label Learning	Original	0.47 (0.87)	0.49 (0.87)	0.41 (0.83)	0.50 (0.86)	0.46 (0.87)
3 year	Original	0.26 (0.70)	0.29 (0.70)	0.26 (0.67)	0.29 (0.68)	0.23 (0.69)
3 year	Learned	0.24 (0.71)	0.28 (0.73)	0.27 (0.72)	0.30 (0.72)	0.24 (0.71)
4 year	Original	0.27 (0.67)	0.29 (0.68)	0.26 (0.66)	0.29 (0.66)	0.25 (0.69)
4 year	Learned	0.21 (0.68)	0.27 (0.72)	0.26 (0.72)	0.29 (0.71)	0.20 (0.71)
5 year	Original	0.25 (0.64)	0.27 (0.63)	0.24 (0.61)	0.26 (0.62)	0.25 (0.67)
5 year	Learned	0.22 (0.66)	0.24 (0.71)	0.21 (0.71)	0.25 (0.70)	0.23 (0.68)
6 year	Original	0.26 (0.68)	0.27 (0.65)	0.23 (0.64)	0.27 (0.64)	0.25 (0.69)
6 year	Learned	0.22 (0.67)	0.24 (0.69)	0.23 (0.69)	0.25 (0.69)	0.23 (0.69)
7 year	Original	0.25 (0.65)	0.25 (0.67)	0.21 (0.64)	0.26 (0.66)	0.26 (0.68)
7 year	Learned	0.22 (0.67)	0.21 (0.69)	0.20 (0.67)	0.22 (0.68)	0.18 (0.68)
8 year	Original	0.25 (0.63)	0.25 (0.63)	0.22 (0.60)	0.28 (0.62)	0.24 (0.59)
8 year	Learned	0.15 (0.72)	0.21 (0.65)	0.21 (0.63)	0.25 (0.66)	0.18 (0.70)

Abbreviations: SEHR, structured electronic health record data; OW-C, Open World claims only data; OW-E, Open World EHR data; OW-M Open World mixed data.

# Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study



## Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study

Christopher A Hane<sup>1</sup> ; Vijay S Nori<sup>1</sup> ; William H Crown<sup>1</sup> ; Darshak M Sanghavi<sup>1</sup> ; Paul Bleicher<sup>1</sup>



JMIR Med Inform (IF: 2.58; Q4), 2020 Jun 3;8(6):e17819. doi: 10.2196/17819.

# Introduction

- ✔ Undiagnosed dementia cases lead to delayed treatment & management
- ✔ Difficulty in recruiting participants for clinical trials
- ✔ Previous studies relied on small-scale clinical data or claims data alone
- ✔ NLP-based EHR processing has the potential to improve predictive accuracy



# Objective

- ✔ Combine claims data and EHR (clinical notes) to improve prediction accuracy
- ✔ Apply NLP techniques to extract meaningful terms from clinical notes



To enhance model accuracy by incorporating **clinical notes data** and analyze the frequency of cognitive concerns appearing in patients' clinical records up to 10 years before ADRD onset.

And The processing of the clinical notes in this study favors **automation, not clinical insight and expertise.**

# Data & Cohort



Data Source: OptumLabs Data Warehouse (2007-2017)

- Claims Data (insurance, pharmacy, lab tests), EHR , De-identified patient records

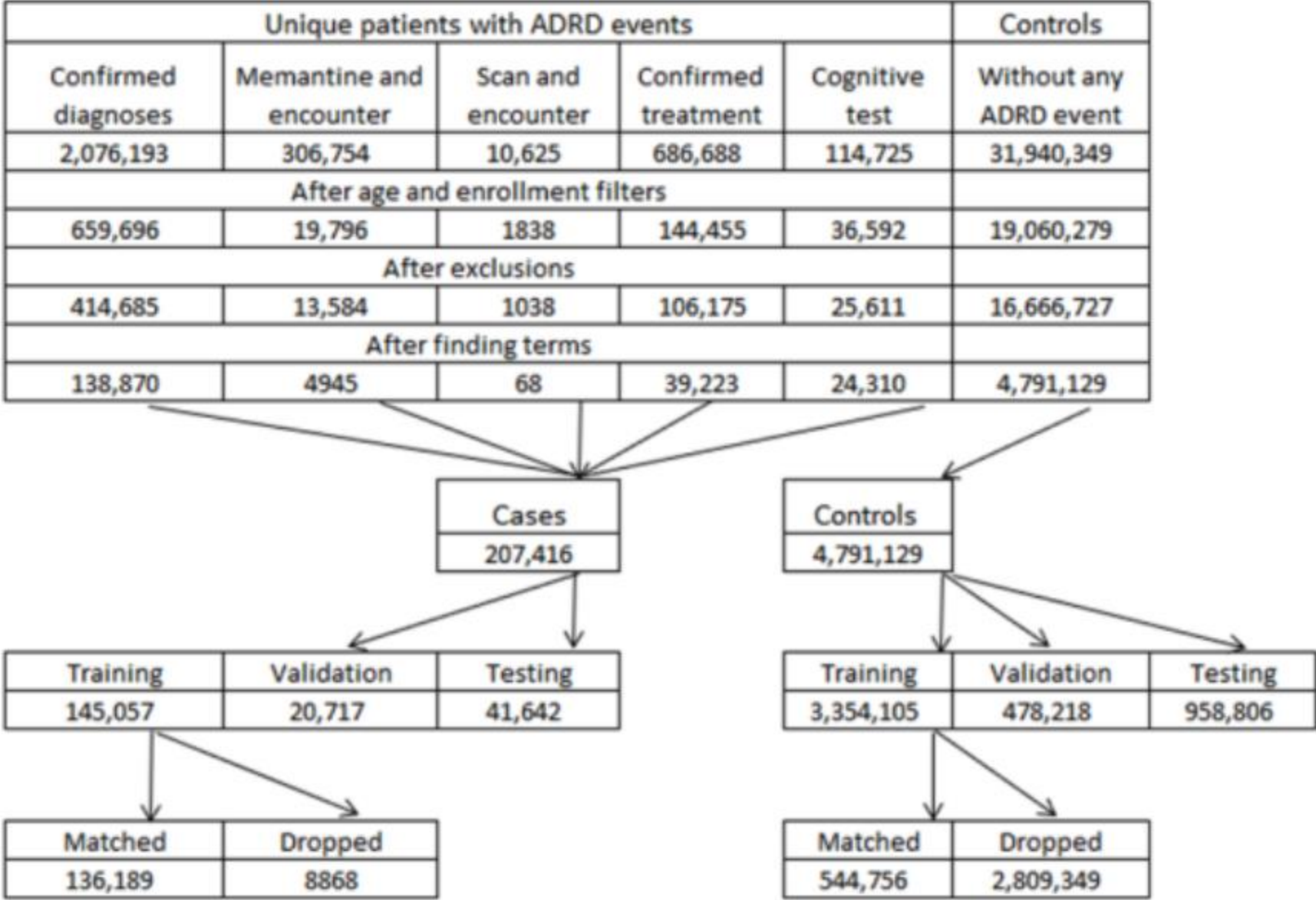


Cohort Selection

- ADRD Diagnosed Patients (Cases) vs. Dementia-Free Controls
- Utilized 2 years of data
- Applied 1:4 Matching Criteria(Age, Gender, Visit Frequency, Index Year)

# DataSet

Figure 1. Attrition table. ADRD: Alzheimer disease and related dementias.



# Clinical Notes & NLP Processing

- ✓ Proprietary **NLP from Optum** extracts medical concepts
  - Includes medications, symptoms, family history
- ✓ Clustering similar terms using FastText + hclust

# Clustering

- ✓ Clinical records contain various expressions with the same meaning (abbreviations, spelling variations)  
Examples:(MI, AMI, Acute Myocardial Infarction), (HTN, Hypertension), (Alzheimer Disease vs. Alzheimer Dementia)
- ✓ This study uses **FastText NLP** to automatically cluster similar terms
- ✓ Fully automated clustering without clinical expert intervention(hclust)

# Machine Learning

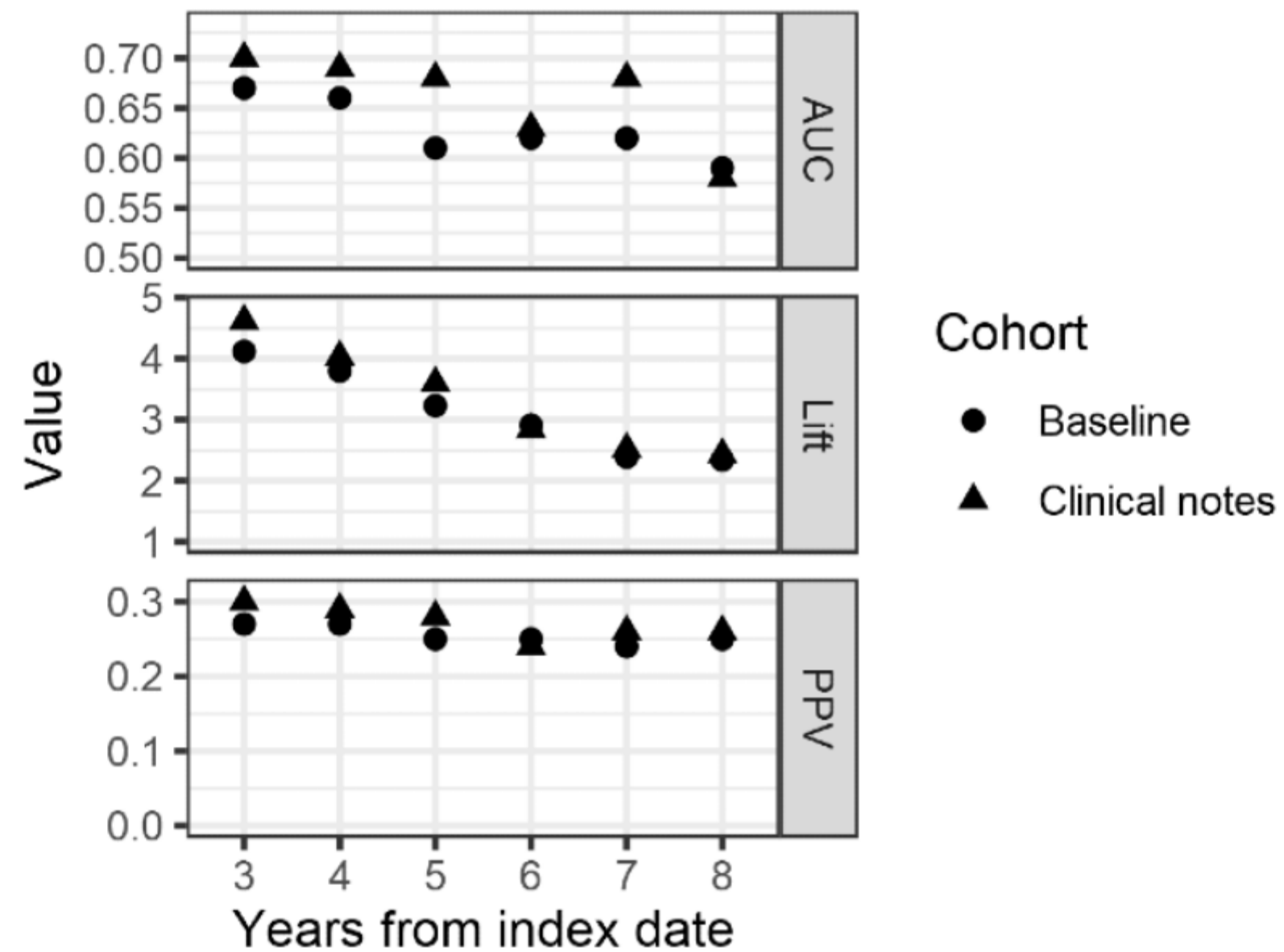
- ✔ Feature Filtering
  - Removed extreme features based on Matched to Unmatched Odds Ratio
  - Eliminated features without sufficient support in the data
  - Excluded terms highly correlated with age to prevent bias in the model
- ✔ Machine Learning Model with LightGBM
- ✔ Model Optimization in LightGBM

# Top 20 Relative Risk Diagnoses

Table 2. Top 20 relative risks of diagnosis.

Diagnosis	International Classification of Diseases, Ninth Revision code	Relative risk at years to index date				
		0	3	4	5	6
Wandering in diseases classified elsewhere	V403.1	21.57	— <sup>a</sup>	—	—	—
Unspecified senile psychotic condition	290.9	19.26	—	—	—	—
Unspecified persistent mental disorders due to conditions classified elsewhere	294.9	17.38	6.78	5.89	6.20	5.07
Senility without mention of psychosis	797.	16.48	—	—	—	—
Other general symptoms	780.9	16.27	—	—	—	—
Unspecified nonpsychotic mental disorder following organic brain damage	310.9	16.25	5.92	—	—	—
Other specified nonpsychotic mental disorders following organic brain damage	310.89	15.99	—	—	—	—
Other specified nonpsychotic mental disorder following organic brain damage	310.8	15.78	—	—	—	—
Other signs and symptoms involving cognition	799.59	15.06	4.45	—	—	—
Frontal lobe executive functional deficit	799.55	15.01	—	—	—	—
Dissociative amnesia	300.12	13.52	—	—	—	—
Personality change due to conditions classified elsewhere	310.1	13.52	4.42	—	—	—
Factitious disorder with predominantly psychological signs and symptoms	300.16	13.39	—	—	—	—
Psychotic disorder with delusions in conditions classified elsewhere	293.81	12.86	—	—	—	—
Confusional arousals	327.41	12.48	—	—	—	—
Visuospatial deficit	799.53	12.42	—	—	—	—
Reactive confusion	298.2	12.21	4.54	—	—	—
Subacute delirium	293.1	12.15	—	—	—	—
Alcohol-induced persisting amnestic disorder	291.1	12.07	—	—	—	—
Frontal lobe syndrome	310.0	12.07	—	—	—	—

# Evaluation Metric(AUC, Lift, PPV)



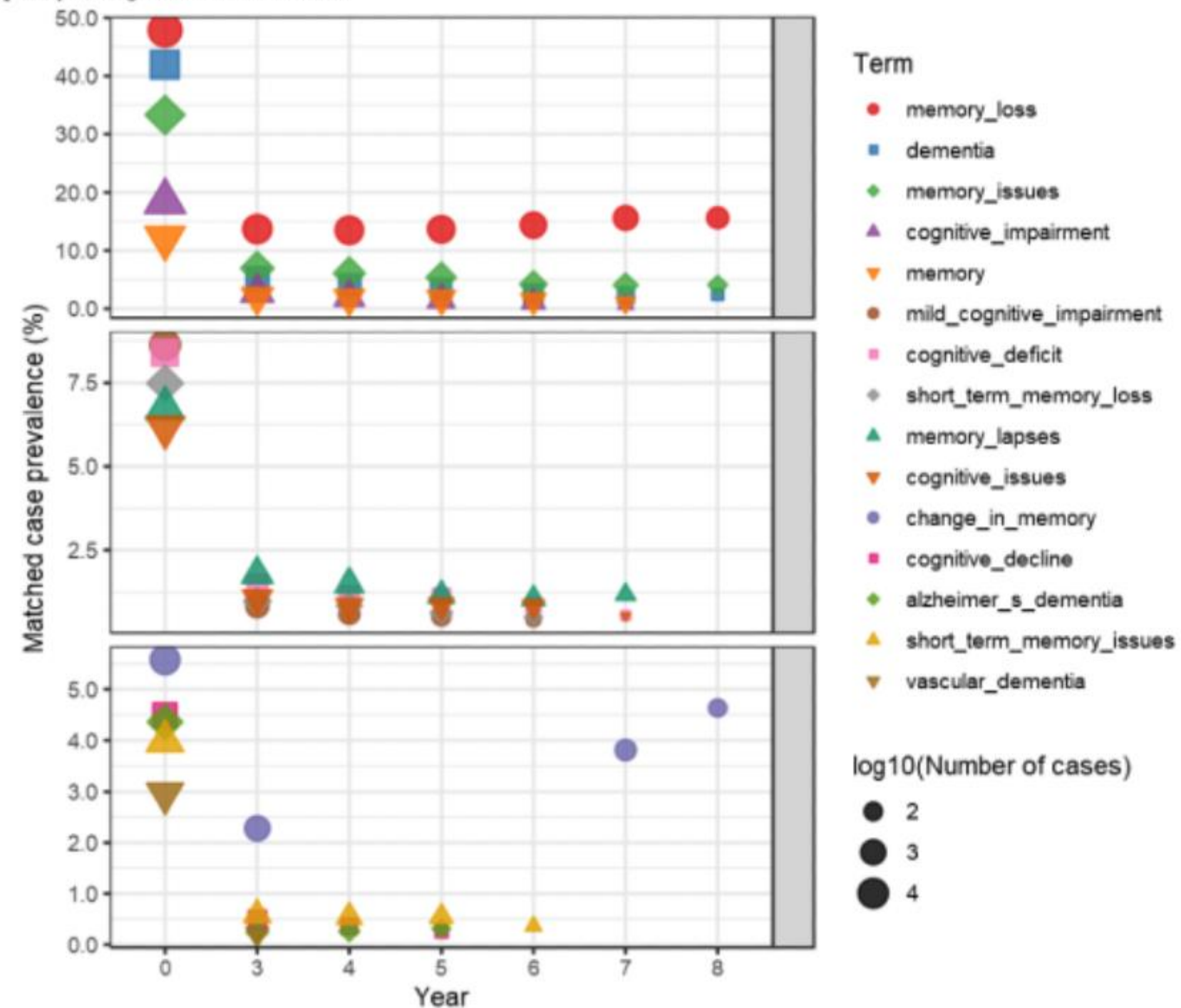


**Table 4.** Important variables at onset (year 0) Total Gain (N) is 22,040,569.

Variable type	Variable name	Gain, n	Percent gain	Cumulative percent gain
cls	Dementia and Alzheimer dementia	3,298,549	15.0	15.0
cls	Memory loss and memory issues	2,833,536	12.9	27.8
idv	Dementia	2,162,843	9.8	37.6
idv	memory_issues	1,525,697	6.9	44.6
idv	memory_loss	1,498,113	6.8	51.4
idv	mild_cognitive_impairment	459,131	2.1	53.4
idv	Forgetful	419,780	1.9	55.3
cls	Alzheimer disease and other family memory issues	382,811	1.7	57.1
ETG	Neurological diseases signs and symptoms	378,955	1.7	58.8
idv	cognitive_impairment	346,991	1.6	60.4
idv	Memory	337,701	1.5	61.9
ICD	Altered mental status	275,533	1.3	63.2
idv	memory_lapses	256,076	1.2	64.3
idv	short_term_memory_loss	252,683	1.1	65.5
CPT	Neuropsychological testing (eg, Halstead-Reitan neuropsychological battery, Wechsler memory scales, and Wisconsin card sorting test), per hour of the psychologist's or physician's time, both face-to-face time administering tests to the patient and time interpreting these test results and preparing the report	245,279	1.1	66.6
cls	Cognitive impairment and hearing impairment	232,700	1.1	67.6
idv	Alzheimers_disease	221,324	1.0	68.6
cls	Cognitive issues and cognitive disorder	214,553	1.0	69.6
ETG	Mood disorder, depressed	214,171	1.0	70.6
CPT	Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material	213,180	1.0	71.5
ICD	Unspecified persistent mental disorders due to conditions classified elsewhere	174,643	0.8	72.3
cls	Memory lapses and concentrating	163,176	0.7	73.1
idv	getting_lost	159,014	0.7	73.8
CPT	Computed tomography, head or brain; without contrast material	150,658	0.7	74.5
cls	Family dementia and memory disturbance	125,350	0.6	75.1
ETG	Psychotic and schizophrenic disorders	121,595	0.6	75.6
dem	Age	118,598	0.5	76.1
RXG	Atypical antipsychotics	115,677	0.5	76.7
ETG	Mental disorders, organic and drug-induced	114,621	0.5	77.2
cls	Pain and tenderness	106,109	0.5	77.7
CPT	Neuropsychological testing (eg, Halstead-Reitan neuropsychological battery, Wechsler memory scales, and Wisconsin card sorting test), with qualified health care professional interpretation and report, administered by technician, per hour of technician time, face-to-face	100,425	0.5	78.1
dem	Number of encounters	94,671	0.4	78.6
RXG	Selective serotonin reuptake inhibitors	93,374	0.4	79.0
OFam	informant	92,567	0.4	79.4
idv	relaxing_issues	84,741	0.4	79.8
ICD	Depressive disorder, not elsewhere classified	77,831	0.4	80.1

# Frequency of cognitive terms in cases

Figure 3. Frequency of cognitive terms in cases.



# Frequency of cognitive terms in controls

Figure 4. Frequency of cognitive terms in controls.

