Paper Review

Contents

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[1]; Vijay S Nori[1]; William H Crown[1]; Darshak M Sanghavi[1]; Paul Bleicher[1]

JMIR Medical Information (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)

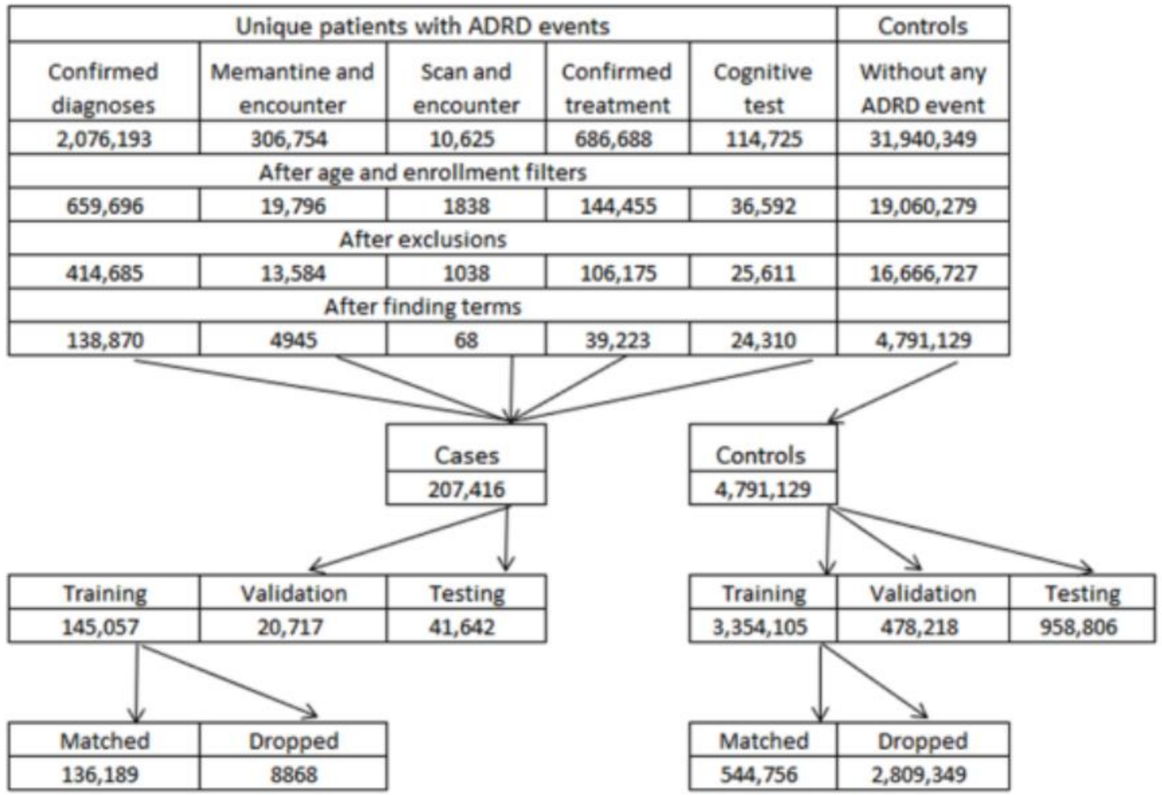
  - Claims Data (insurance, pharmacy, lab tests), EHR , De-identified patient records

- 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.



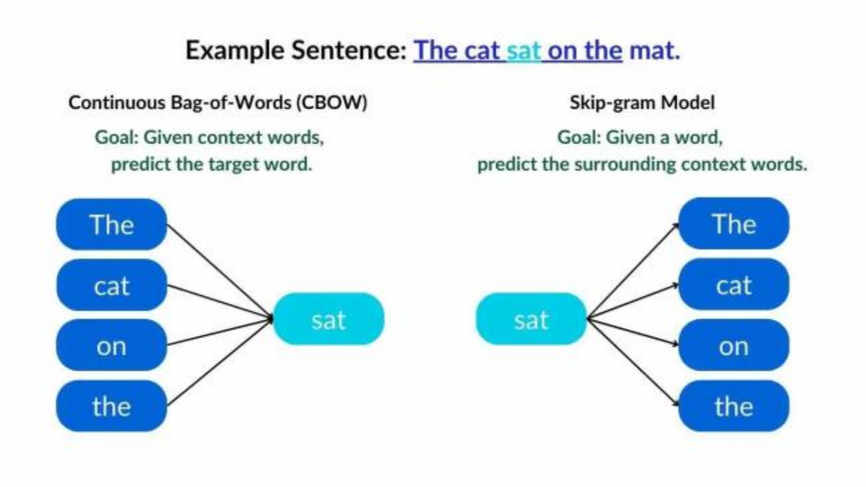
Why Did We Choose a 1:4 Matching Ratio?

- Controlling for Confounding Variables
- Addressing Data Sparsity Issues
- Preventing Overfitting
- Potential Drawback:
  A high control-to-case ratio (like 4:1) can reduce model sensitivity, but we addressed this issue with additional methods (e.g., odds ratio filtering) to balance model performance effectively.

Clinical Notes & NLP Processing

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

FastText



- FastText, developed by Facebook AI Research.
- Key Process Steps

1. Transforming Clinical Terms into Sequences
2. Learning Semantic Representations
3. Model Configuration
4. Subword Information Disabled

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
  - Applied the same feature filtering method as proposed by Nori et al.
  - 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	__[a]	—	—	—
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	—	—	—	—

[a]

Evaluation Metric(AUC, Lift, PPV)

Table 3. Quality of model fit on the test data.

Year	Sensitivity		Specificity		Area under the curve		Lift	
	Baseline	Clinical notes	Baseline	Clinical notes	Baseline	Clinical notes	Baseline	Clinical notes
0	0.45	0.68	0.98	0.99	0.84	0.94	13.92	16.39
3	0.27	0.30	0.95	0.95	0.67	0.70	4.12	4.62
4	0.27	0.29	0.94	0.95	0.66	0.69	3.80	4.03
5	0.25	0.28	0.94	0.94	0.61	0.68	3.23	3.60
6	0.25	0.24	0.93	0.93	0.62	0.63	2.91	2.84
7	0.24	0.26	0.91	0.92	0.62	0.68	2.39	2.52
8	0.25	0.26	0.91	0.91	0.59	0.58	2.34	2.43

Evaluation Metric(AUC, Lift, PPV)

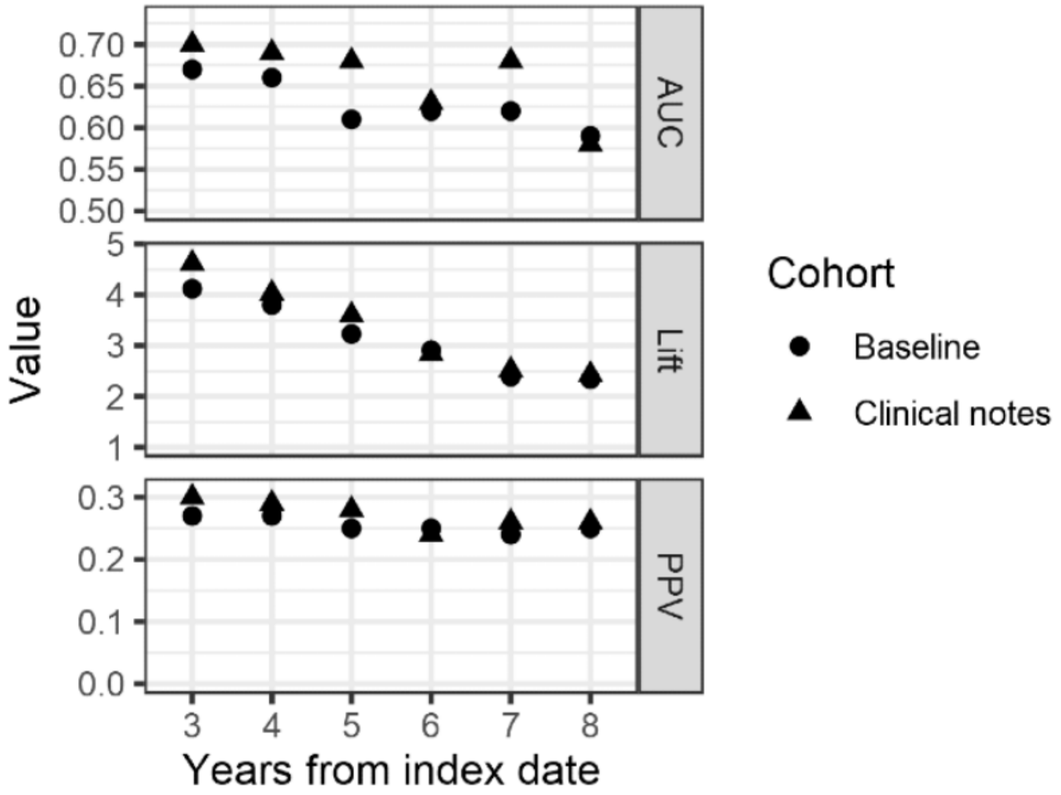
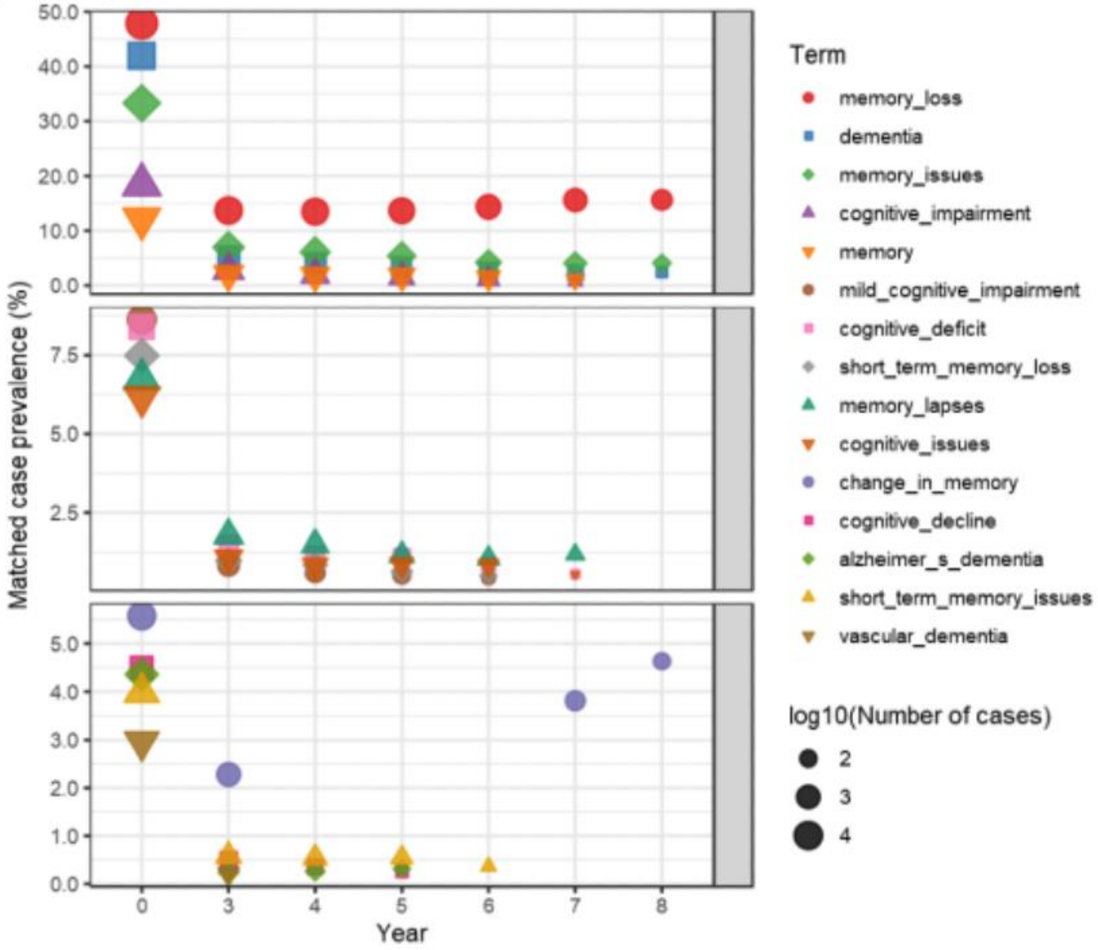


Table 4. Important variables at onset (year 0) Total Gain (N) is 22,040,569.

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

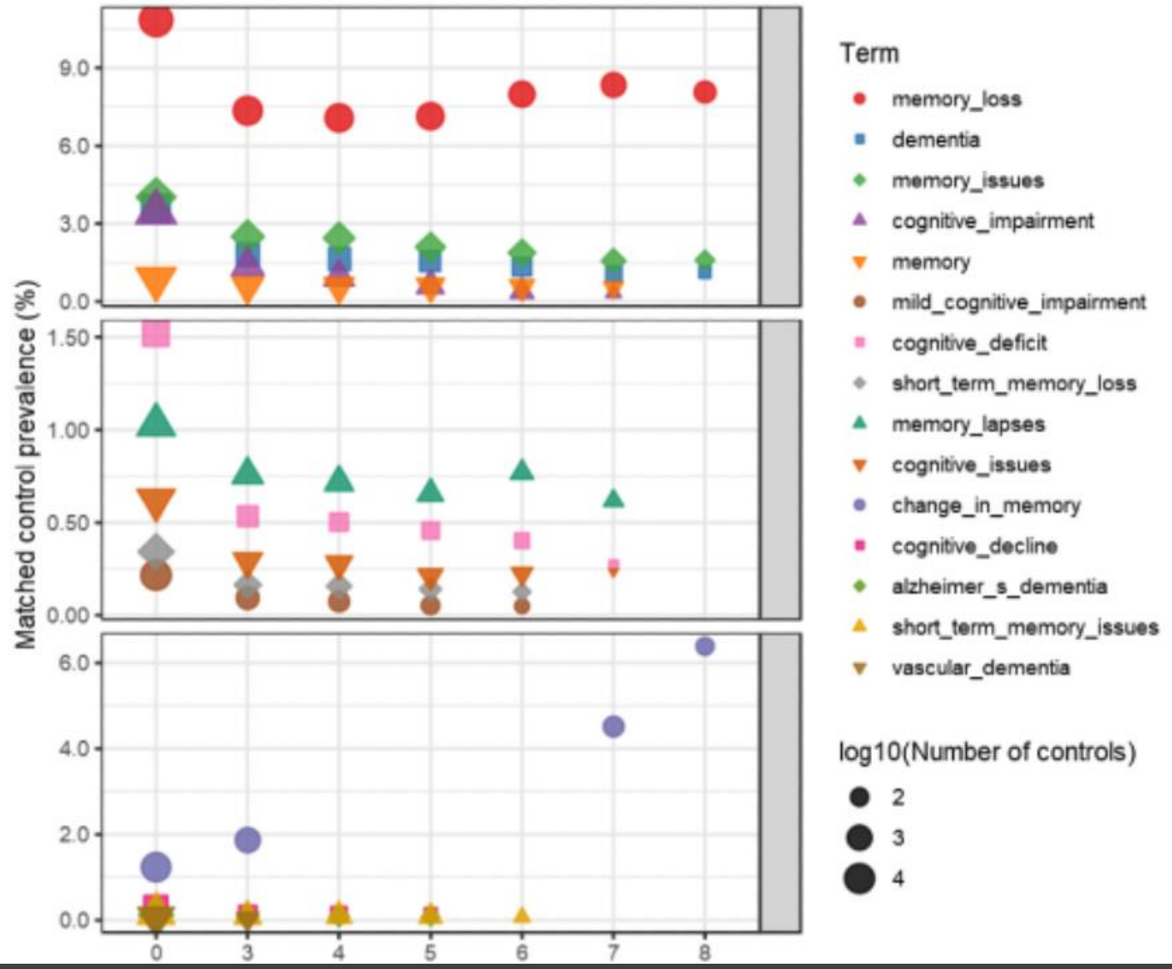
Frequency of congnitive terms in cases

Figure 3. Frequency of cognitive terms in cases.



Frequency of congnitive terms in controls

Figure 4. Frequency of cognitive terms in controls.



Conclusion

- Clinical Notes Play a Crucial Role in ADRD Prediction
- Early Symptoms Appear in Clinical Notes Up to 6 Years Before Diagnosis
- Under-Coding of ADRD in Medical Records
- Family History Had Minimal Impact on Model Performance

Limitations

- Diagnosis Coding Issues
- Generalizability of Findings
- Enhancing Long-Term Prediction Performance