JMIR Publications
The Leading eHealth Publisher

From Memory Loss to Dementia: Real-World Insights into Dementia Diagnosis Trajectory and Clinical Practice Patterns Unveiled by Natural Language Processing

Hunki Paek, Richard H. Fortinsky, Kyeryoung Lee, Liang-Chin Huang, Yazeed S. Maghaydah, George A. Kuchel, Xiaoyan Wang

Submitted to: JMIR Aging
on: August 08, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

From Memory Loss to Dementia: Real-World Insights into Dementia Diagnosis Trajectory and Clinical Practice Patterns Unveiled by Natural Language Processing

Hunki Paek[1*] PhD; Richard H. Fortinsky[2*] PhD; Kyeryoung Lee[1] PhD; Liang-Chin Huang[1] PhD; Yazeed S. Maghaydah[2] MD; George A. Kuchel[2] MD; Xiaoyan Wang[1, 3, 4] PhD

[1]IMO Health Rosemont US
[2]UConn Center on Aging, University of Connecticut School of Medicine Farmington US
[3]Center for Quantitative Medicine, University of Connecticut School of Medicine Farmington US
[4]Department of Health Policy and Management, Tulane University New Orleans US
[*]these authors contributed equally

Corresponding Author:
Kyeryoung Lee PhD
IMO Health
9600 West Bryn Mawr Avenue Suite 100
Rosemont
US

Abstract

Background: Understanding the dementia disease trajectory and clinical practice patterns in outpatient settings is vital for effective management. Knowledge about the path from initial memory loss complaints to dementia diagnosis remains limited

Objective: This study aims to 1) determine the time intervals between initial memory loss complaints and dementia diagnosis in outpatient care, 2) assess the proportion of patients receiving cognition-enhancing medication prior to dementia diagnosis, and 3) identify patient and provider characteristics that influence the time between memory complaints and diagnosis, and the prescription of cognition-enhancing medication.

Methods: This retrospective cohort study utilized a large outpatient EHR database from the University of Connecticut Health Center, covering 2010-2018, with a cohort of 581 outpatients. We employed a customized deep learning-based natural language processing (NLP) pipeline to extract clinical information from electronic health record (EHR) data, focusing on cognition-related symptoms, primary caregiver relation, and medication usage. We applied descriptive statistics, linear, and logistic regression for analysis.

Results: The NLP pipeline showed precision, recall, and F1 scores of 0.97, 0.93, and 0.95, respectively. The median time from the first memory loss complaint to dementia diagnosis was 342 days. Factors such as the location of initial complaints and diagnosis, and primary caregiver relationships significantly affected this interval. Around 25% of patients were prescribed cognition-enhancing medication before diagnosis, with the number of complaints influencing medication usage

Conclusions: Our NLP-guided analysis provided insights into the clinical pathways from memory complaints to dementia diagnosis and medication practices, which can enhance patient care and decision-making in outpatient settings.

(JMIR Preprints 08/08/2024:65221)
DOI: https://doi.org/10.2196/preprints.65221

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

From Memory Loss to Dementia: Real-World Insights into Dementia Diagnosis Trajectory and Clinical Practice Patterns Unveiled by Natural Language Processing

Hunki Paek, PhD[1*], Richard H. Fortinsky PhD[2*], Kyeryoung Lee, PhD[1], Liang-Chin Huang, PhD[1], Yazeed S. Maghaydah MD[2], George A. Kuchel MD[2*], Xiaoyan Wang PhD[1,3,4*]

[1]Intelligent Medical Objects, Rosemont, IL USA
[2]UConn Center on Aging, University of Connecticut School of Medicine, Farmington, CT USA
[3]Center for Quantitative Medicine, University of Connecticut School of Medicine, Farmington, CT USA
[4]Department of Health Policy and Management, Tulane University, New Orleans, LA USA

[*] Equal contributions

Word Counts: 3,484

Corresponding Author

Xiaoyan Wang, PhD
IMO health,
9600 West Bryn Mawr Avenue Suite 100,
Rosemont, IL 10018 USA
Tel: 847-613-6655

Abstract

Background: Understanding the dementia disease trajectory and clinical practice patterns in outpatient settings is vital for effective management. Knowledge about the path from initial memory loss complaints to dementia diagnosis remains limited.

Objective: This study aims to 1) determine the time intervals between initial memory loss complaints and dementia diagnosis in outpatient care, 2) assess the proportion of patients receiving cognition-enhancing medication prior to dementia diagnosis, and 3) identify patient and provider characteristics that influence the time between memory complaints and diagnosis, and the prescription of cognition-enhancing medication.

Methods: This retrospective cohort study utilized a large outpatient EHR database from the University of Connecticut Health Center, covering 2010-2018, with a cohort of 581 outpatients. We employed a customized deep learning-based natural language processing (NLP) pipeline to extract clinical information from electronic health record (EHR) data, focusing on cognition-related symptoms, primary caregiver relation, and medication usage. We applied descriptive statistics, linear, and logistic regression for analysis.

Results: The NLP pipeline showed precision, recall, and F1 scores of 0.97, 0.93, and 0.95, respectively. The median time from the first memory loss complaint to dementia diagnosis was 342 days. Factors such as the location of initial complaints and diagnosis, and primary caregiver relationships significantly affected this interval. Around 25% of patients were prescribed cognition-enhancing medication before diagnosis, with the number of complaints influencing medication usage.

Conclusion and Relevance: Our NLP-guided analysis provided insights into the clinical pathways from memory complaints to dementia diagnosis and medication practices, which can enhance patient care and decision-making in outpatient settings.

Introduction

The rising prevalence of dementia, driven by an aging population, presents a profound concern for society [1–4], placing substantial burdens on individuals but also imposing high financial costs [5–7]. Despite these challenges, no curative treatments are currently available [8,9], highlighting the critical need for early detection of prodromal symptoms of dementia, such as mild cognitive decline [10–13], and timely diagnosis. Early intervention can delay disease progression or alter the trajectory toward dementia [9,14,15]. However, dementia and its associated symptoms are frequently underreported and underdiagnosed in clinical practice [16,17]. As the condition progresses, patients commonly experience increased memory loss, deteriorating cognitive ability, heightened confusion, and changes in personality like agitation. The conversion from mild cognitive impairment to AD has been explored utilizing patient health data [18].

Electronic health records (EHRs), especially unstructured clinical notes, offer a valuable resource for enhancing the detection and management of disease by providing comprehensive data on patient health and history [19–23]. Natural language processing (NLP), a subfield of artificial intelligence that enables computers to understand, interpret, and generate human language, offers promise in extracting meaningful information from vast and complex free-text EHRs [24–27]. NLP has been instrumental in automatically extracting clinical information in various medical domains, including geriatric care [28–35]. For instance, Kharrazi et al. showcased higher rates of geriatric syndrome extraction from unstructured EHR using NLP compared to relying solely on claim data or structured EHR data [36]. Studies have successfully extracted cognitive status and measurement scores [37,38], as well as lifestyle exposures and discourse production for AD [39] from clinical documentation utilizing NLP. Additionally, multiple studies have applied NLP methods to extract neuropsychiatric symptoms, and cognitive or function impairment information [40–43]. While previous research has made significant strides in the earlier detection of cognitive decline utilizing EHR, most studies have focused on extracting symptoms or cognitive measurement scores rather than other clinical features that affect disease progression, such as the relationship of the primary caregivers with patients.

We aimed to investigate the time interval from initial memory loss complaints to dementia diagnosis and explore the association between various clinical features, including the family primary caregiver relationship using real-world outpatient clinical notes. Additionally, we aimed to analyze the pattern of cognition-enhancing medication prescriptions before diagnosis. To achieve this, we developed a customized NLP pipeline using deep-learning techniques, based on a prodromal dementia symptom ontology that we established.

Methods

Study Design

This retrospective study utilized data from the University of Connecticut (UConn) Health Center between 2010 and 2018 (Supplemental Figure 1A). The utilization of longitudinal EHR data allowed us to track all patients' clinical information including demographic characteristics, diagnoses, measurements, medications, and signs (Supplemental Figure 1B). The study cohort was defined as patients who met the following criteria: 1) received a dementia diagnosis, 2) had at least one outpatient visit per year 3) had at least one visit before the dementia diagnosis, and 4) had documented memory loss-related symptoms (e.g., memory loss, confusing, cognition impairment, trouble remembering, doesn't recall, forgets, blackout etc.) in the EHR. We analyzed demographic and clinical characteristics from structured EHR data, including insurance details, the initial location (medical unit) of memory loss complaints, and the location of the first dementia diagnosis. These locations encompass various settings within this healthcare system, such as geriatric medicine, internal medicine, and neurology outpatient clinics. NLP was utilized to extract symptoms and primary caregiver (family supporter) relationship information from clinical notes. Both diagnosis and cognition-improving medication information were extracted from both structured and unstructured data. Figure 1 provides an overview of the selection process of the study cohort and the information extraction process.

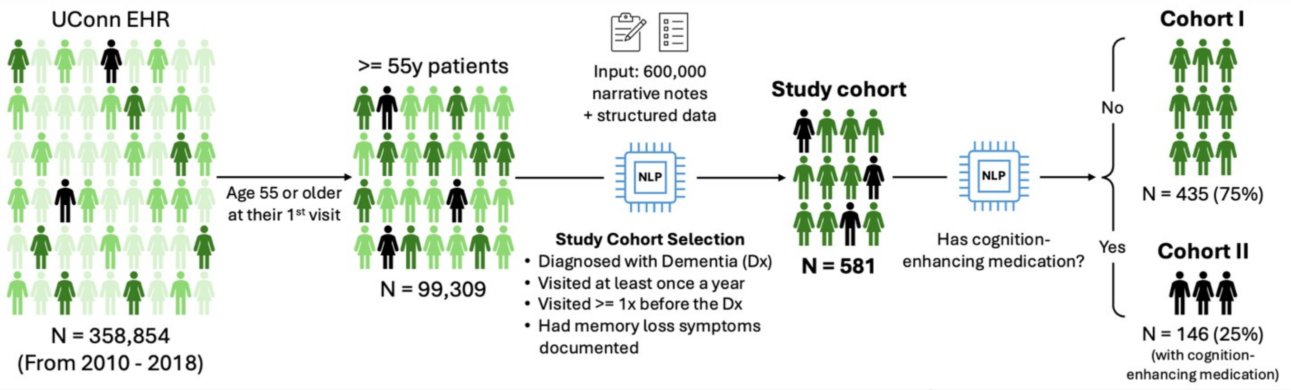

UConn EHR
N = 358,854
(From 2010 - 2018)
Age 55 or older at their 1st visit
>= 55y patients
N = 99,309
Input: 600,000 narrative notes + structured data
NLP
Study Cohort Selection
• Diagnosed with Dementia (Dx)
• Visited at least once a year
• Visited >= 1x before the Dx
• Had memory loss symptoms documented
Study cohort
N = 581
NLP
Has cognition-enhancing medication?
No
Yes
Cohort I
N = 435 (75%)
Cohort II
N = 146 (25%)
(with cognition-enhancing medication)

Figure 1. The workflow for the study cohort selection and information extraction

Building a deep learning-based NLP pipeline

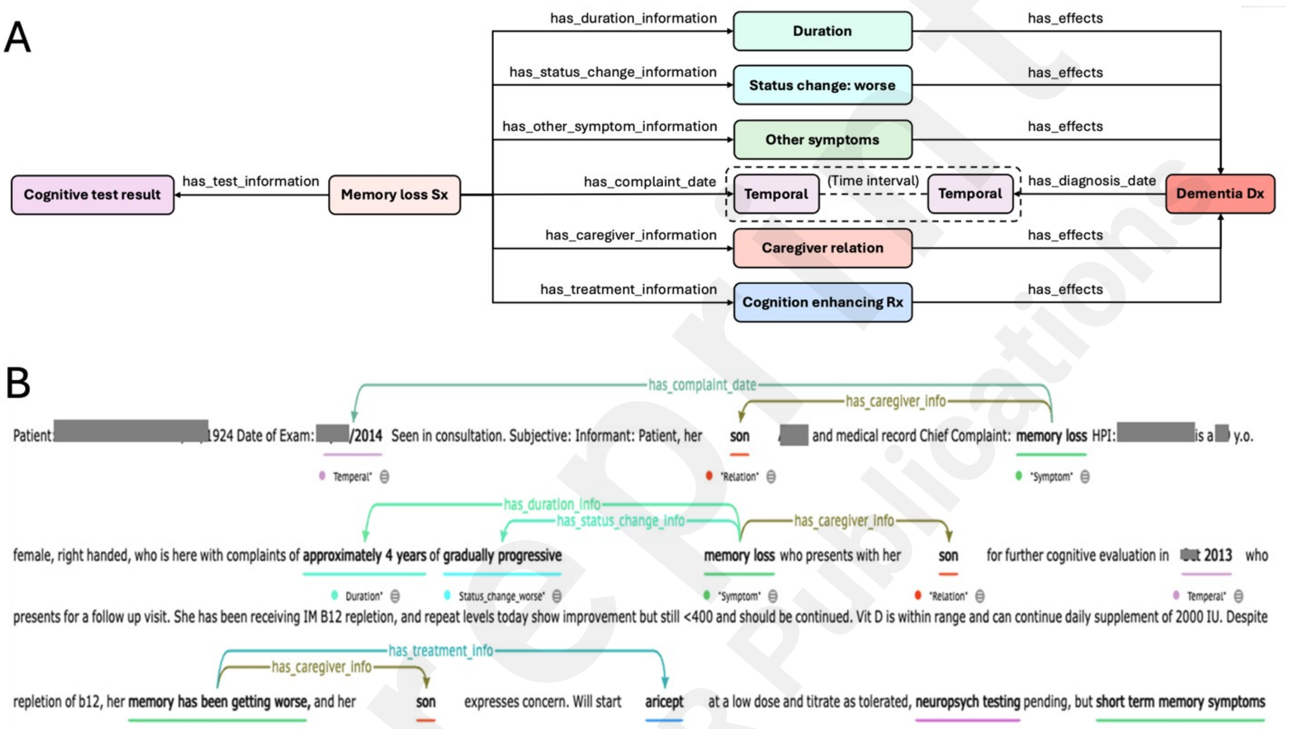
A framework was developed to curate the prodromal dementia symptoms comprising of four stages: 1). preprocessing and query expansion; 2) ontology construction and annotation; 3) NLP model development and 4) system evaluation.

Pre-processing and query expansion:

In this stage, a query was expanded to extract a comprehensive set of patient notes with documented memory loss symptoms. A list of seed terms was obtained through a manual survey of the literature and a clinical note review by a clinical researcher and domain expert. A bi-gram word2vec algorithm [44] was used to identify additional significant terms potentially related to memory loss symptoms to ensure the encapsulation of an expansive cohort. The expanded query terms were then applied to extract a comprehensive set of patient notes for NLP modeling.

Ontology construction and annotation:

This stage involves the simulation of an expert's knowledge and understanding of the free text. A prodromal dementia symptom ontology was built based on the physician's opinion, comprehensive literature review, and clinical note review. The ontology includes nine entities and nine relations. Nine entities are "Memory loss symptom (Sx)", "Dementia diagnosis (Dx)", "Temporal", "Duration", "Status change: worse", "Other symptoms", "Cognitive test result", "Caregiver relation", and "Cognition enhancing medication (Rx)". Nine relations are "has complaint date", "has diagnosis date", "has other symptom information", "has status change information", "has duration information", "has test information", "has caregiver information", "has treatment information", "has effects" as depicted in Figure 2A. Two independent annotators manually annotated notes using Clinical Language Annotation, Modeling, and Processing (CLAMP)[45], an NLP toolkit, guided by the constructed ontology. Figure 2B shows an example note with entities and relations annotated.



Patient:[redacted]1924 Date of Exam:[redacted]/2014 Seen in consultation. Subjective: Informant: Patient, her **son** [redacted] and medical record Chief Complaint: **memory loss** HPI:[redacted]is a[redacted] y.o. female, right handed, who is here with complaints of **approximately 4 years** of **gradually progressive** **memory loss** who presents with her **son** for further cognitive evaluation in [redacted]**t 2013** who presents for a follow up visit. She has been receiving IM B12 repletion, and repeat levels today show improvement but still <400 and should be continued. Vit D is within range and can continue daily supplement of 2000 IU. Despite repletion of b12, her **memory has been getting worse,** and her **son** expresses concern. Will start **aricept** at a low dose and titrate as tolerated, **neuropsych testing** pending, but **short term memory symptoms**

Figure 2. The ontology of memory loss and the annotated sample note. A) The ontology of prodromal dementia symptoms. Nine entities and nine semantic relations between entities were defined in the ontology. B) A sample note that has been de-identified and annotated with prodromal dementia symptoms. In this note, the following entities have been annotated and related to each other: Symptoms, Relation, Medication, Duration, Status change, Test, and Temporal.

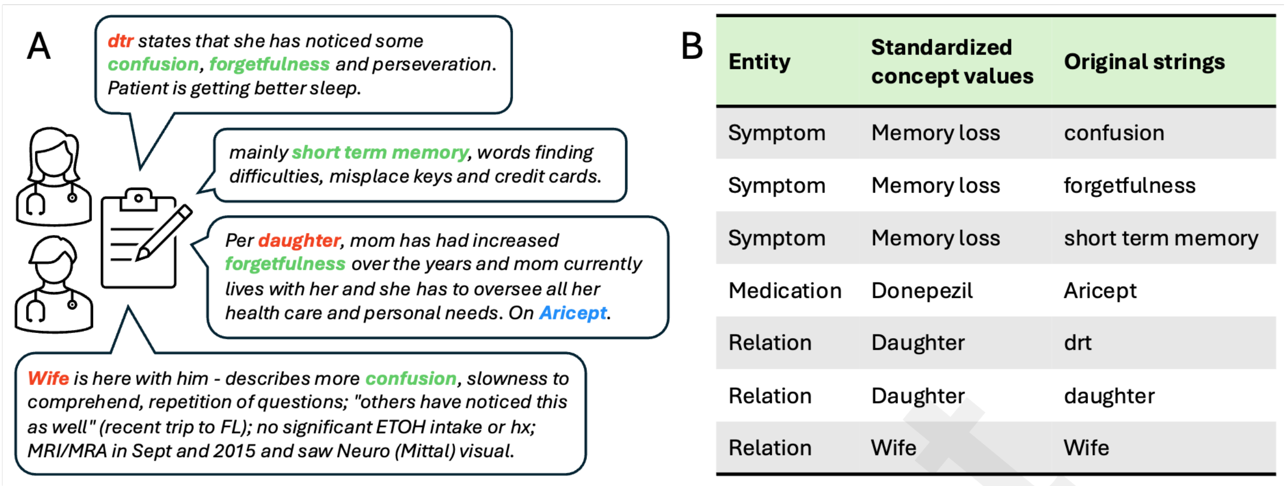
NLP model development:

The annotated notes obtained in the previous stage were utilized for NLP model training. A multi-layer deep learning architecture was adopted, which involved transforming the text into sequential vectors of characterization through the embedding step. The vectors were then fed into a Bidirectional Long-Short Term Memory (BiLSTM), a text classification architecture based on artificial neural networks, for pattern recognition in both forward and backward directions. Finally, the patterns were sent to the next layer of a Conditional Random Fields (CRF) model for prediction probability computation. 500 notes were annotated and 80% of annotated notes were used for the model training, while 20 % were reserved for independent model performance validation. The model was trained, calibrated, and tested for optimal performance.

NLP pipeline evaluation: The performance of the pipeline was evaluated in the validation set through Precision (positive predictive value (PPV)), Recall (sensitivity), and F1 score (a balanced score between false positives (FP) and false negatives (FN). Recall was calculated as the ratio of the number of entities that were identified by the pipeline over the total number of the corresponding entities in the manually annotated gold standard (i.e., true positive (TP)/ (TP + FN)). Precision was measured as the ratio of the number of distinct entities returned by the correct pipeline according to the gold standard divided by the total number of entities found by our pipeline (i.e., TP/ (TP + FP)). F1 score was calculated as the harmonic mean of PPV and sensitivity (i.e., 2 x PPV x sensitivity / (PPV + sensitivity). The manual annotation and training process was repeated with additional manually annotated notes until the model achieved the average F1 score >0.8.

Standardization of concept values leveraging NLP

Clinical notes contain abundant information but are often heterogeneous in form. To enable use case analysis, these heterogeneous entities needed to be standardized. Figure 3A illustrates the various forms in which cognition-related concepts like forgetfulness, memory loss, short-term memory, and confusion were documented in the clinical notes. Abbreviations (e.g., dtr for daughter) and mix-use of brand names and generic names of the same drugs (e.g., Aricept and Donepezil) were commonly found. Figure 3B provides examples of standardized concept values obtained through the NLP process. Cognition-enhancing medications described included donepezil, memantine, rivastigmine, and galantamine. For the analysis, the son, son-in-law, daughter-in-law, grandson, and granddaughter were classified as other adult children, while the nephew, niece, cousin, brother and sister were classified as other family support.



dtr states that she has noticed some confusion, forgetfulness and perseveration. Patient is getting better sleep.

mainly short term memory, words finding difficulties, misplace keys and credit cards.

Per daughter, mom has had increased forgetfulness over the years and mom currently lives with her and she has to oversee all her health care and personal needs. On Aricept.

Wife is here with him - describes more confusion, slowness to comprehend, repetition of questions; "others have noticed this as well" (recent trip to FL); no significant ETOH intake or hx; MRI/MRA in Sept and 2015 and saw Neuro (Mittal) visual.

Entity	Standardized concept values	Original strings
Symptom	Memory loss	confusion
Symptom	Memory loss	forgetfulness
Symptom	Memory loss	short term memory
Medication	Donepezil	Aricept
Relation	Daughter	drt
Relation	Daughter	daughter
Relation	Wife	Wife

Figure 3. Clinical note standardization via NLP. A) The diverse expressions used in clinical notes to depict concepts such as memory loss symptom, primary caregiver relationship, and cognition-enhancing medications. B) A sample of the standardized output generated by the NLP process, representing the standardized values of the various original expressions

Characterization of dementia cohorts and longitudinal analysis of dementia trajectory

The collection of clinical information of patients was accomplished through the extraction of both structured and unstructured EHR data. The statistical analysis was performed using R software (R core team). A logistic regression model was employed to assess the relationship between medication prescription history and other factors such as memory loss complaints and the primary caregiver relation information, by calculating the odd ratio (OR) and 95% confidence intervals (CI). 149 Patients who had memory loss complaints and diagnoses recorded on the same day were excluded from this study.

Results

Performance measures of memory loss NLP pipeline

The performance of our memory loss NLP pipeline was evaluated using precision, recall (sensitivity), and F1 score metrics, with detailed results presented in Table 1. Our system achieved high scores across all semantic types including "Memory Loss Symptom", "Dementia Diagnosis", "Duration", "Primary Caregiver", and "Status Change" (overall precision, recall (sensitivity), and F1 scores of 0.97, 0.93, and 0.95, respectively). For example, the precision for "memory loss symptoms" was 0.97, indicating that 97% of "memory loss symptoms" identified by our NLP system in patient's clinical notes, as verified against a manually curated gold standard. The recall (sensitivity) of 0.93 implies that our system correctly identified 93% of actual memory loss cases, with only 7% missed. Supplemental Table 1 provides comprehensive details including correct, predicted, and gold standard values for various semantic types, enabling a thorough understanding of the NLP pipeline's performance.

Clinical characterization of study cohorts

An average of 358,854 patients visited the UConn Health System for outpatient care between 2010 and 2018, with the number of visits increasing annually from 224,488 in 2010 to 1,024,349 in 2017. There were 8,686 patient visits in 2018 until the time point we collected data. Out of these patients, 99,039 patients were aged 55 or older at their first visit. From over 600,000 narrative records and coded data of those patients, 730 patients had documented memory loss symptoms in their clinical notes. Of these, 149 patients (20.4%) reported memory loss complaints and were diagnosed with dementia on the same day, while 581 patients (79.6%) had at least one-day gap between the complaint and diagnosis. For the following analysis, 581 patients with memory loss symptoms documented at different days were included in study cohorts (Figure 1). The demographic characteristics, primary insurance, and medication information of study cohorts are summarized in Table 1. Most cohort members were non-Hispanic White individuals (87.6%) and female (65.6%), with an age distribution of over 85 years (54.2%), 75-85 years (29.4%), 65-74(10.8%), and under 65 (5.5%). Out of 291 patients with primary caregiver relation information, adult children were the main caregivers (63.5%, with 46.3% being daughters and 17.2% being sons) followed by spouses (22.3%, with

13.4% being wives and 8.9% being husbands). Other family members (e.g., nephew) made up 14.1% of the cohort. Out of 581 patients, 146 patients (25%) had been prescribed cognition-improving medications prior to the dementia diagnosis.

Table 1. Demographic characteristics, primary insurance, and medication information of study cohorts

Total	581(100%)
Age (years)	
Under 65	32 (5.5%)
65-74	63 (10.8%)
75-84	171 (29.4%)
85+	315 (54.2%)
Race	
White	509 (87.6%)
African American	33 (5.7%)
Other	39 (6.7%)
Sex	
Female	381 (65.6%)
Male	200 (34.4%)
Primary caregiver (family supporter) relation	
Husband	26 (8.9%)
Wife	39 (13.4%)
Daughter	135 (46.3%)
Son	50 (17.2%)
Other family members	41 (14.1%)
Missing	290
Prior medication before the first diagnosis of dementia	146 (25%)

Next, we investigated the outpatient care locations where the first memory loss complaints were reported and where dementia was diagnosed. Geriatrics is the most frequent location for both the first memory loss complaints made (53.0%) and the diagnosis of dementia (61.0%), followed by Primary care (31.8% and 27.4%, respectively) and Neurology (6.7% and 10.6%, respectively). The majority of the cohort was covered by Medicare (60.9%) or Medicaid (12.4%) as primary insurance, while 26.2% had commercial insurance. Only a small percentage of patients (0.3%) had no insurance coverage (Supplemental Table 2).

Distribution of time intervals between cognitive symptom complaints and dementia diagnosis and the number of complaints

The median time interval between the first memory loss complaints and dementia diagnosis was 342 days, ranging from a minimum of 1 day to a maximum of 1,458 days in our study cohort (n= 581) (Supplemental Figure 1A). Additionally, the number of complaints made before being diagnosed was analyzed, with a median of 3 complaints, ranging from a minimum of 1 complaint to a maximum of 18 complaints (Supplemental Figure 1B).

Association analysis for the earlier dementia diagnosis

We aimed to identify the clinical features that are associated with earlier diagnosis of dementia from the first memory loss complaints. Results indicated that the location of the first complaints made and the diagnosis, as well as the relation of the primary caregiver, were significantly associated with earlier diagnosis of dementia. Patients who made complaints in Geriatrics (- 141 days, p=<0.001) or Neurology (-158 days, p=0.016) were diagnosed with dementia earlier compared to those who made complaints in primary care. Furthermore, patients diagnosed with dementia in Geriatrics had a shorter interval of 152.9 days (p=<0.001) compared to those diagnosed in primary care. Additionally, having a wife or a daughter as a primary caregiver was associated with an earlier diagnosis of dementia, with a shorter interval of 249.6 days (p=0.010) and 176.8 days (p=0.04), respectively, compared to those who had a husband as a primary caregiver. However, factors such as age or insurance types were not found to have a significant impact on earlier diagnosis (Table 2).

Table 2. Statistical analysis of time intervals between the first complaints and dementia diagnosis.

Demographics	Estimate	P value
Age (years)		
Under 65	109.8	0.115
65-74	-61.6	0.2345
75-84	0.08	0.9982
85+	Ref	
Primary Insurance		

No insurance	36.8	0.2433
Medicaid	5.8	0.9653
Medicare	36.8	0.1387
Commercial	Ref	
The location of the first memory loss complaint		
Geriatrics	-141	<0.0001*
Neurology	-158	<0.0157*
Primary care	Ref	
other	81	0.1716
The location of the first diagnosis of dementia		
Geriatrics	-152.9	<0.001*
Neurology	-82.2	0.14
Primary care	Ref	
other	42	0.7582
Primary caregiver (family supporter) relation		
Wife	-249.6	0.0091*
Daughter	-176.8	0.0407*
other adult children	-127.4	0.0539
other family support	-257.5	0.1143
Husband	Ref	

*Statistically significant

Association analysis for the medication usage

Medication was prescribed in 25% of patients before dementia diagnosis. We next analyzed factors associated with the usage of cognition-enhancing medication before the diagnosis of dementia after the 1st complaints of memory loss. The only factor that was significantly associated with medication usage was the total number of memory loss complaints made; each additional memory complaint was associated with a 15% greater likelihood that cognition-enhancing medications were prescribed (OR=1.148; 95% CI:1.027-1.283) (Table 3).

Table 3. An analysis of the factors associated with the usage of medication before the diagnosis of dementia.

Demographics	**OR**	**95% Confidence Interval**	
Age (years)			
Under 65	3.827	0.403	23.32
65-74	1.251	0.507	3.727
75-84	1.127	0.615	2.445
85+	Ref		
The location of the first memory loss			

complaint			
Geriatrics	1.477	0.551	3.959
Neurology	2.124	0.449	10.05
Primary care	Ref		
other	0.331	0.103	1.058
The location of the first diagnosis			
Geriatrics	0.489	0.172	1.39
Neurology	0.65	0.182	2.319
Primary care	Ref		
Family Support			
Wife	4.367	0.850	22.447
Daughter	1.831	0.263	12.74
Other adult children	1.609	0.538	4.816
Other family support	1.033	0.276	3.871
Husband	Ref		
Total number of memory loss complaints before dx	1.148	1.027	1.283*

*Statistically significant

Discussion

We developed a high-performance deep learning-based NLP algorithm on an EHR dataset of dementia patients to delve into the real-world trajectory of dementia, starting from initial memory loss complaints to dementia diagnosis. Our investigation focused on the time interval from the first memory loss complaints to dementia diagnosis, the proportion of prescribed cognition-enhancing medication before diagnosis during this trajectory, and the clinical characteristics associated with these features.

We found that 20.4% of patients (n=149) had same-day documentation of memory loss complaints and dementia diagnosis. Among the remaining 79.6% of patients with at least a one-day gap between complaints and diagnosis, over half of the patients received a dementia diagnosis within a year of their initial memory loss complaints, with a median time of 342 days. The location of the first complaint and diagnosis and the relationship with the primary caregiver emerged as influential factors in achieving an earlier diagnosis. Notably, patients who initiated complaints or were diagnosed in Geriatrics or Neurology received earlier diagnoses compared to those in primary care. This underscores the important role of the initial complaint's location and the dementia diagnosis's setting in the early detection and management of dementia. Our findings align with previous research indicating missed and delayed diagnoses in primary cares [46]. Geriatricians and

Neurologists possess significantly more expertise and practical experience in diagnosing dementia and prescribing these meds than most primary care doctors. Additionally, understanding the factors that lead patients to receive care in a geriatric or neurological department rather than primary care would be an important question for further investigation. Similar to the previous study that identified dementia severity and marital status as independent predictors of receiving a clinical cognitive evaluation [47], other factors such as more complex medical needs (e.g., multiple chronic conditions and polypharmacy), severe function decline, and the primary caregiver's educational level or relationship with the patient could be associated with visits to geriatric or neurologic departments.

Remarkably, we found that patients with a wife or daughter as their primary caregiver were diagnosed earlier and more frequently used cognitive-improving medication before the dementia diagnosis. This emphasizes the vital role of primary caregivers in the diagnosis and treatment of dementia patients. Mahmoudi et al [35]. previously emphasized the importance of extracting caregiver information in dementia patient notes and developed the rule-based NLP algorithm to identify caregiver availability. In our work, we extended this by also extracting family-caregiver relationships with patients and analyzing their impact on the early diagnosis of dementia. Subsequent research should explore the underlying mechanisms and factors of caregivers in this context such as the association between the relationship of primary caregivers and visits to geriatric or neurologic departments. Contrary to expectations, our study revealed that age and insurance were not associated with earlier diagnoses. Surprisingly, the total number of memory loss complaints emerged as the sole factor significantly linked to medication usage, with other factors showing no significant association.

We demonstrated that extracting cognitive symptom-related terms from longitudinally documented patient notes before dementia diagnosis could be an alternative approach to analyzing documented cognitive measurement scores during patient visits, potentially aiding in identifying dementia patients. Previous studies have highlighted a significant lack of such documentation in clinical notes [37,41,48]. For instance, Harding et. al. found that cognitive measurement scores were rarely available in their cohorts when establishing the algorithm for identifying dementia patients in EHR [48]. Similarly, Maserejian et. al. demonstrated a low percentage of dementia (11%) or AD (24%) patients with cognitive measurement scores such as mini-mental state examination (MMSE), a recall test, a clock drawing, Montreal cognitive assessment (MoCA), Mini-Cog, or Saint Louis University Mental Status (SLUMS) documented and suggested prompts of cognitive measurement [37]. McCoy T.H. et al. attempted to extract cognitive symptom-related terms (e.g., impulsive, forgetful, cognitive, memory...) and converted them into scores, given the issue of reliability and scalability of the cognitive measurement test [41]. Consistently, the proportion of patient notes with cognitive test names, including MMSE, SLIMS, MOCA, Mini-cog, clock drawing, trail making, Boston naming test, and Wisconsin card sorting test, was very low (around 10%) in our study, so these were not used in further analysis.

Our NLP approach in automatically identifying cognitive symptom-related terms and primary caregivers, and systematically analyzing these factors along with other structured data, enhanced our understanding of dementia progression and management. Further exploration using this NLP method could significantly advance the field, providing deeper insights and more effective interventions for dementia care.

Strengths and Limitations

Our study has several strengths. Firstly, our work presents a novel approach to understanding clinical practices in dementia by examining the time interval from the symptom complaints to the diagnosis of dementia using real-world data. Secondly, our study highlights how the relationship between primary caregivers and patients, and the location (medical unit: geriatrics, neurology, and primary care) of complaints made may influence the time to diagnosis. Thirdly, our study developed and validated a customized NLP model to be used to predict an outcome in a clinical setting using EHRs.

Several limitations of this study should be considered. Firstly, the study relied on EHR data from a single healthcare system, which may limit the generalizability of the findings to other populations. Additionally, the patient population was predominantly white, female, and elderly patients over 85 years with Medicare or Medicaid insurance, which may limit the applicability of the findings to other demographic groups. Future studies should aim to include a more diverse patient population in terms of age, race, and insurance type. Furthermore, the study did not account for the potential impact of other medical conditions on the result.

Lastly, incorporating objective measures of cognitive decline and other co-occurring neuropsychiatric symptoms could enhance the assessment of dementia.

Conclusions

Our study highlights the importance of the location of initial memory loss complaints, the location of the dementia diagnosis, and the role of the primary caregiver in the early diagnosis and treatment of dementia patients. By analyzing complex clinical dementia care practice patterns within a real-world setting on a large scale using NLP, our exploratory analysis demonstrates the potential of advanced analytical techniques in achieving earlier and more accurate diagnoses of dementia.

References

1. Hebert, L. E., Weuve, J., Scherr, P. A. & Evans, D. A. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* **80**, 1778–1783 (2013).
2. Livingston, G. *et al.* Dementia prevention, intervention, and care. *Lancet Lond. Engl.* **390**, 2673–2734 (2017).
3. Matthews, K. A. *et al.* Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged ≥65 years. *Alzheimers Dement.* **15**, 17–24 (2019).
4. Plassman, B. L. *et al.* Prevalence of Dementia in the United States: The Aging, Demographics, and Memory Study. *Neuroepidemiology* **29**, 125–132 (2007).
5. Pedroza, P. *et al.* Global and regional spending on dementia care from 2000–2019 and expected future health spending scenarios from 2020–2050: An economic modelling exercise. *eClinicalMedicine* **45**, 101337 (2022).
6. GBD 2016 Dementia Collaborators. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 88–106 (2019).
7. Frankish, H. & Horton, R. Prevention and management of dementia: a priority for public health. *Lancet Lond. Engl.* **390**, 2614–2615 (2017).
8. Tisher, A. & Salardini, A. A Comprehensive Update on Treatment of Dementia. *Semin. Neurol.* **39**, 167–178 (2019).
9. Long, J. M. & Holtzman, D. M. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell* **179**, 312–339 (2019).
10. Gale, S. A., Acar, D. & Daffner, K. R. Dementia. *Am. J. Med.* **131**, 1161–1169 (2018).
11. 2021 Alzheimer's disease facts and figures. *Alzheimers Dement. J. Alzheimers Assoc.* **17**, 327–406 (2021).

12. Ferencz, B. & Gerritsen, L. Genetics and underlying pathology of dementia. *Neuropsychol. Rev.* **25**, 113–124 (2015).
13. Flicker, C., Ferris, S. H. & Reisberg, B. Mild cognitive impairment in the elderly: predictors of dementia. *Neurology* **41**, 1006–1009 (1991).
14. Ossenkoppele, R. *et al.* Accuracy of Tau Positron Emission Tomography as a Prognostic Marker in Preclinical and Prodromal Alzheimer Disease: A Head-to-Head Comparison Against Amyloid Positron Emission Tomography and Magnetic Resonance Imaging. *JAMA Neurol.* **78**, 961–971 (2021).
15. Vermunt, L. *et al.* Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimers Dement. J. Alzheimers Assoc.* **15**, 888–898 (2019).
16. Amjad, H. *et al.* Underdiagnosis of Dementia: an Observational Study of Patterns in Diagnosis and Awareness in US Older Adults. *J. Gen. Intern. Med.* **33**, 1131–1138 (2018).
17. Maclagan, L. C. *et al.* Can Patients with Dementia Be Identified in Primary Care Electronic Medical Records Using Natural Language Processing? *J. Healthc. Inform. Res.* **7**, 42–58 (2023).
18. Cui, Y. *et al.* Identification of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Multivariate Predictors. *PLoS ONE* **6**, e21896 (2011).
19. Yao, L., Zhang, Y., Li, Y., Sanseau, P. & Agarwal, P. Electronic health records: Implications for drug discovery. *Drug Discov. Today* **16**, 594–599 (2011).
20. Ellsworth, M. A. *et al.* An appraisal of published usability evaluations of electronic health records via systematic review. *J. Am. Med. Inform. Assoc.* **24**, 218–226 (2017).
21. Wang, Y. *et al.* Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **77**, 34–49 (2018).
22. Bacigalupo, I. *et al.* Identification of dementia and MCI cases in health information systems: An Italian validation study. *Alzheimers Dement. N. Y. N* **8**, e12327 (2022).

23. Tjandra, D., Migrino, R. Q., Giordani, B. & Wiens, J. Cohort discovery and risk stratification for Alzheimer’s disease: an electronic health record-based approach. *Alzheimers Dement. Transl. Res. Clin. Interv.* **6**, (2020).

24. Fu, S. *et al.* Clinical concept extraction: A methodology review. *J. Biomed. Inform.* **109**, 103526 (2020).

25. Wang, X., Chused, A., Elhadad, N., Friedman, C. & Markatou, M. Automated knowledge acquisition from clinical narrative reports. *AMIA Annu. Symp. Proc. AMIA Symp.* 783–787 (2008).

26. Wang, X., Hripcsak, G., Markatou, M. & Friedman, C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *J. Am. Med. Inform. Assoc.* **16**, 328–337 (2009).

27. Koleck, T. A., Dreisbach, C., Bourne, P. E. & Bakken, S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J. Am. Med. Inform. Assoc. JAMIA* **26**, 364–379 (2019).

28. Dave, A. D., Ruano, G., Kost, J. & Wang, X. Automated Extraction of Pain Symptoms: A Natural Language Approach using Electronic Health Records. *Pain Physician* **25**, E245–E254 (2022).

29. Weegar, R. Applying natural language processing to electronic medical records for estimating healthy life expectancy. *Lancet Reg. Health - West. Pac.* **9**, 100132 (2021).

30. Maclagan, L. C. *et al.* Using natural language processing to identify signs and symptoms of dementia and cognitive impairment in primary care electronic medical records (EMR). *Alzheimers Dement.* **17**, (2021).

31. Al-Harrasi, A. M. *et al.* Motor signs in Alzheimer’s disease and vascular dementia: Detection through natural language processing, co-morbid features and relationship to adverse outcomes. *Exp. Gerontol.* **146**, 111223 (2021).

32. Byrd, R. J., Steinhubl, S. R., Sun, J., Ebadollahi, S. & Stewart, W. F. Automatic identification of

heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int. J. Med. Inf.* **83**, 983–992 (2014).

33. Divita, G. *et al.* General Symptom Extraction from VA Electronic Medical Notes. *Stud. Health Technol. Inform.* **245**, 356–360 (2017).
34. Fu, S. *et al.* Ascertainment of Delirium Status Using Natural Language Processing From Electronic Health Records. *J. Gerontol. A. Biol. Sci. Med. Sci.* **77**, 524–530 (2022).
35. Mahmoudi, E. *et al.* Identifying Caregiver Availability Using Medical Notes With Rule-Based Natural Language Processing: Retrospective Cohort Study. *JMIR Aging* **5**, e40241 (2022).
36. Kharrazi, H. *et al.* The Value of Unstructured Electronic Health Record Data in Geriatric Syndrome Case Identification. *J. Am. Geriatr. Soc.* **66**, 1499–1507 (2018).
37. Maserejian, N., Krzywy, H., Eaton, S. & Galvin, J. E. Cognitive measures lacking in EHR prior to dementia or Alzheimer's disease diagnosis. *Alzheimers Dement. J. Alzheimers Assoc.* **17**, 1231–1243 (2021).
38. Noori, A. *et al.* Development and Evaluation of a Natural Language Processing Annotation Tool to Facilitate Phenotyping of Cognitive Status in Electronic Health Records: Diagnostic Study. *J. Med. Internet Res.* **24**, e40384 (2022).
39. Zhou, X. *et al.* Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. *Int. J. Med. Inf.* **130**, 103943 (2019).
40. Topaz, M., Adams, V., Wilson, P., Woo, K. & Ryvicker, M. Free-Text Documentation of Dementia Symptoms in Home Healthcare: A Natural Language Processing Study. *Gerontol. Geriatr. Med.* **6**, 2333721420959861 (2020).
41. McCoy, T. H. *et al.* Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. *Alzheimers Dement. J. Alzheimers Assoc.* **16**, 531–540 (2020).
42. McCoy, T. H. *et al.* High Throughput Phenotyping for Dimensional Psychopathology in Electronic Health Records. *Biol. Psychiatry* **83**, 997–1004 (2018).

43. Oh, I. Y. *et al.* Extraction of clinical phenotypes for Alzheimer's disease dementia from clinical notes using natural language processing. *JAMIA Open* **6**, ooad014 (2023).
44. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. (2013) doi:10.48550/ARXIV.1301.3781.
45. Soysal, E. *et al.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* **25**, 331–336 (2018).
46. Bradford, A., Kunik, M. E., Schulz, P., Williams, S. P. & Singh, H. Missed and Delayed Diagnosis of Dementia in Primary Care: Prevalence and Contributing Factors. *Alzheimer Dis. Assoc. Disord.* **23**, 306–314 (2009).
47. Kotagal, V. *et al.* Factors associated with cognitive evaluations in the United States. *Neurology* **84**, 64–71 (2015).
48. Harding, B. N. *et al.* Methods to identify dementia in the electronic health record: Comparing cognitive test scores with dementia algorithms. *Healthc. Amst. Neth.* **8**, 100430 (2020).

Authors' Contributions

H. Paek, G.A. Kuchel, K. Lee, R.H. Fortinsky, and X. Wang designed the study and wrote the manuscript. H. Paek, K. Lee and X. Wang reviewed the literature, and patient notes and constructed the dementia ontology. H. Paek, K. Lee, L-C Huang, and X. Wang were involved in the model training, the post-processing, and the data analysis. Y.S. Maghaydah, G.A. Kuchel, R.H. Fortinsky, and X. Wang discussed the project and reviewed the manuscript.

Financial Support: (grant information):

No financial support to declare.

Disclosures: (add grant or other information):

HP, KL, LHC, and XW are current employees of IMO Health. All authors declare no other competing financial or non-financial interests.

Reproducible Research Statement:

The data used in this study is not open access due to patient privacy, security and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requirement.

Corresponding Author:

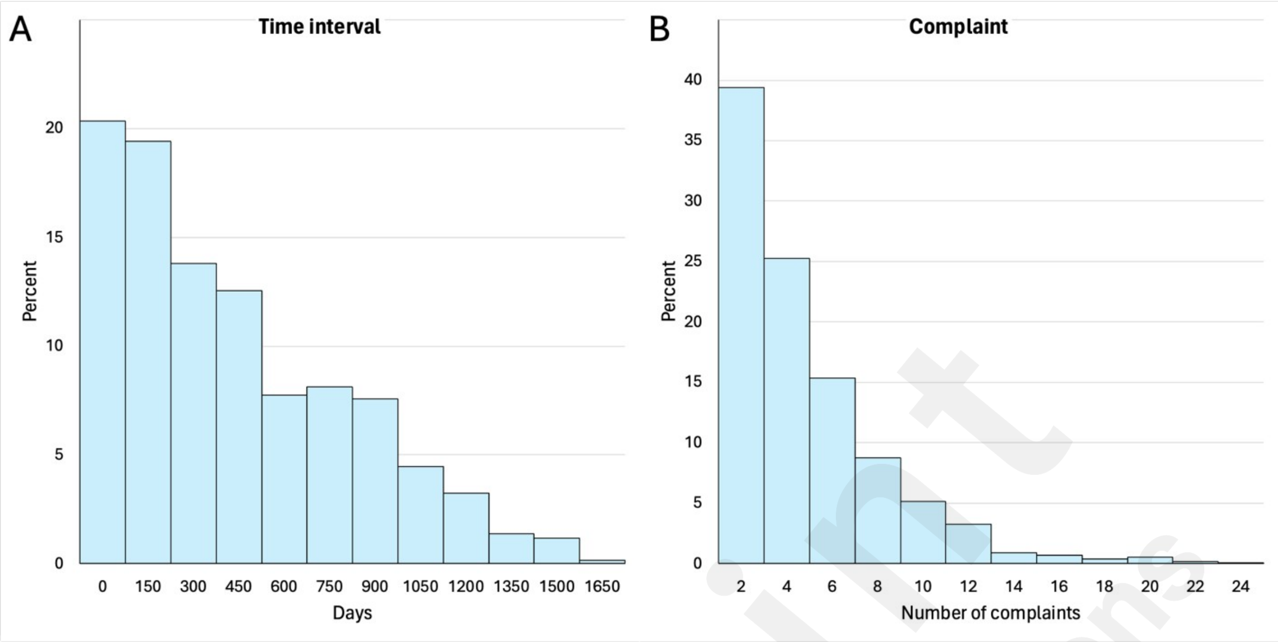
Xiaoyan Wang, PhD
9600 West Bryn Mawr Avenue Rosemont, IL 60018
Phone: 2012828098
Email: xw108@caa.columbia.edu

Supplemental Table 1. Evaluation of memory loss NLP pipeline.

Semantic	Correct	Predict	Gold	Precision	Recall	F-Score
Memory Loss Symptom	**4178**	**4249**	**4327**	**0.98**	**0.97**	**0.97**
Dementia Diagnosis	**873**	**881**	**908**	**0.99**	**0.96**	**0.98**
Duration	**275**	**300**	**350**	**0.92**	**0.79**	**0.85**
Primary Caregiver (family supporter) Relation	**1460**	**1503**	**1510**	**0.97**	**0.97**	**0.97**
Status Change	**264**	**266**	**274**	**0.99**	**0.96**	**0.98**

Supplemental Table 2. Descriptive statistics of providers and insurance information.

Total	**581 (100%)**
The location of the first memory loss complaint	
Geriatrics	308 (53%)
Neurology	39 (6.7%)
Primary care	185 (31.8%)
other	49 (8.4%)
The location of the first diagnosis of dementia	
Geriatrics	350 (61.0%)
Neurology	61 (10.6%)
Primary care	163 (27.4%)
other	7 (1%)
Primary Insurance	
Medicaid	72 (12.4%)
Medicare	354 (60.9%)
Commercial	152 (26.2%)
No insurance	2 (0.3%)



Supplemental Figure 1. The distribution time intervals and complaints. A) Distribution of the time intervals between the first memory loss complaints and the diagnosis of dementia B) Distribution of the number of complaints made before the diagnosis of dementia

Supplementary Files

Multimedia Appendixes

Supplemental Figure 1. The distribution time intervals and complaints. A) Distribution of the time intervals between the first memory loss complaints and the diagnosis of dementia B) Distribution of the number of complaints made before the diagnosis of dementia.
URL: http://asset.jmir.pub/assets/c05381e983f07006e68805d72e1de674.docx