Disease Trajectories and End-of-Life Care for Dementias: Latent Topic Modeling and Trend Analysis Using Clinical Notes

Liqin Wang, PhD[1,2], Joshua Lakin, MD[1,2,3], Clay Riley, MS[1], Zfania Korach, MD[1,2], Laura N. Frain, MD[1,2], Li Zhou, MD, PhD[1,2]

[1]Brigham and Women's Hospital, Boston, MA, USA; [2]Harvard Medical School, Boston, MA, USA; [3]Dana-Farber Cancer Institute, Boston, MA, USA

Abstract

Despite the increasing prevalence, growing costs, and high mortality of dementia in older adults in the U.S., little is known about the course of these diseases and what care dementia patients receive in their final years of life. Using a large volume of clinical notes of dementia patients over the last two years of life, we conducted automatic topic modeling to capture the trends of various themes mentioned in care provider notes, including patients' physical function status, mental health, falls, nutrition and feeding, infections, hospital care, intensive care, end-of-life care, and family and social supports. Our research contributes to the adoption and evaluation of an unsupervised machine learning method using large amounts of retrospective free-text electronic health record data to discover and understand illness and health care trajectories.

1. Introduction and Background

Alzheimer's disease and related dementias (ADRD) affect more than 5.5 million Americans in 2017 and are the sixth leading cause of death in the US.[1] ADRD are prevalent in older adults.[2] Due to the loss of neurons and neuronal connections in the brain,[3] these patients often have a long-term and gradual decline in cognitive ability and memory problems. ADRD are among the costliest diseases: in the U.S. alone, the costs of caring for people over 65 with dementia were estimated to be as high as $259 billion in 2017.[1] These costs continue to rapidly increase as the population ages; the number of Americans aged 65 years and older is projected to more than double from 46 million in 2016 to 98 million by 2060.[1, 4]

End-of-life care for the dementia population presents unique challenges and burdens to their caregivers. As degenerative brain diseases, the symptoms of ADRD worsen over time. Memory progressively declines, and other cognitive functions like language and decision-making become more difficult and may drive distressing events such as delirium, agitation, and changes in personality and mood. Patients living with dementia can be completely reliant on others for assistance with even the most basic activities of daily living, such as eating and hygiene. One or many complications (such as pneumonia, dysphasia, and pain) often occur in dementia patients and may eventually lead to death.[5] Therefore, tracking and understanding the disease progression and complications of dementia near the end of life is important for providing the best care possible for patients and caregivers along the trajectory of the disease. In the past, people have studied the likelihood of developing mortality, morbidity, and various comorbidities among dementia patients;[6] however, there is a lack of an overall picture generated from large-scale clinical narrative data that describe disease trajectories, clinical complications, and treatments over patients' final years of life.

Importantly, people suffering from dementia often lose the ability to communicate their needs and make decisions, especially as the disease advances, complicating treatment and symptom management. Dementia patients may be more likely to receive intensive treatment because clinicians and caregivers are not aware of their preferences and wishes as they get sicker. However, substantial evidence shows that an intensive medical approach (e.g., systemic antibiotics) can be of limited efficacy in this patient population for particular illnesses, and may even cause numerous unintended consequences (e.g., renal failure).[7-9] Experts suggest that advance care planning (ACP), a process of discussing of patients' goals and values between patients and their clinicians in anticipation of possible future health deterioration, provides a foundation for making one's wishes, fears, and plans known.[10] However, given the cognitive decline associated with dementia, we currently do not have a clear path for effective ACP in this specific patient population.

With the national adoption of electronic health records (EHR), a significant amount of clinical data is recorded during a patient's lifetime. The data elements in patients' longitudinal EHR include administrative/claims data (e.g. demographics, encounters, and billing codes) and clinical information, which contains both structured data (e.g., problems, allergies, medications, vital signs, and laboratory data) and narrative documents (e.g., clinic visit notes, progress notes, and discharge summaries). In this study, we specifically focus on studying clinical narrative data, as this type of data conveys rich individualized patient history, assessments, and the provider's clinical decision-making process. We hypothesize that a statistical topic modeling approach using a longitudinal set of clinical notes will

discover critical themes and generate temporal patterns and trends, which in turn will provide important insights to study disease and care trajectories in the dementia population.

Topic modeling approaches apply statistical machine learning algorithms to discover abstract topics that occur in a collection of documents. Blei innovatively introduced probabilistic topic models based on latent Dirichlet allocation (LDA), a three-level hierarchical Bayesian model, to analyze the words of documents to discover the 'topics' that run through them.[11, 12] In the model, every document is modeled as a finite mixture over an underlying set of topic probabilities, while every topic is characterized by a unique Dirichlet distribution of the words that co-occur in them. Additionally, each document is treated as a 'bag of words' in which the linear order and the local structure of the contents of a document are ignored. With these assumptions in place, the corpus is sampled iteratively and its topic-word assignments are adjusted to better model their presumed latent distributions, causing groupings of words to emerge. Semantic content is frequently shared within these collections, allowing the resulting model to assign semantic domains to documents without supervision. Griffiths and Steyvers first applied topic models in scientific articles to discover temporal trends in topics.[13] Topic modeling is now a popular technique for deriving data semantics and has been successful in diverse fields,[14] such as bibliometrics,[13] social media,[15] and bioinformatics.[16] Like scientific articles, an individual clinical note can deal with various topics, such as patient medical history, physical exam, laboratory tests and results, diagnoses, treatments, medical reasoning, and care plans. Researchers have applied topic modeling algorithms to discover the themes embedded in clinical notes. Derived topics were used to summarize and organize documents or used for featurization and dimensionality reduction in late stages of a machine learning pipeline. For example, Zeng *et al.* applied topic modeling to the task of document retrieval.[17] Shao *et al.* used topic modeling for identifying indicators of frailty in clinical notes.[18] More recently, Jo *et al.* combined long short-term memory (LSTM) and a latent topic model for mortality prediction.[19]

In this study, we adopted and evaluated topic models for discovering important themes/topics mentioned in care provider notes about dementia patients. Further, we explored temporal patterns and trends of these topics over the final two years of life.

2. Methods

2.1. Clinical Setting

This study is conducted at Partners HealthCare (PHS), an integrated system in the Greater Boston area in Massachusetts. The PHS care delivery network was founded by two large teaching hospitals affiliated with Harvard Medical School, Brigham and Women's Hospital (BWH) and Massachusetts General Hospital. PHS also includes multiple community hospitals, a physician network, and other health-related entities. In this study, we included patients with dementia from two PHS hospitals: BWH and Faulkner Hospital, a community teaching hospital.

2.2. Data Collection and Preparation

Patient data were collected from PHS' Research Patient Data Registry (RPDR) and Enterprise Data Warehouse (EDW). RPDR is a centralized clinical data registry, or data warehouse, designed for investigators with HIPAA-compliant access to data of 11 million patients, including demographics, clinical encounters, diagnoses, laboratories, medications, allergies, clinical notes, and many other data collected from Partners' EHR systems. EDW integrates data from multiple sources, including clinical and administrative claims data, and is made accessible for analytic work. We first identified a cohort of patients who meet all the following criteria: (1) any patient who was a patient of BWH or of Faulkner Hospital; (2) adults age 18 and older; (3) any patient who was diagnosed with ADRD based on the diagnosis codes listed in Table 1; (4) patients with a known date of death; and (5) patients with at least one note available since the diagnosis of ADRD. The data we accessed were documented between 01/01/2011 and 12/31/2017.

Table 1. ICD-9/10-CM codes used for identifying a cohort of Alzheimer's disease and related dementia

ICD-9-CM	290, 294.1, 294.2, 331.0, 331.1, 331.2, 331.82
ICD-10-CM	F00-F03, G30.0, G30.1, G30.8, G30.9, G31.0, G31.1, G31.83, G31.9

Patients' dates of death were obtained from two data sources: Partners' data repositories (i.e., EDW and RPDR) and Massachusetts Death Certificate files. The death information documented in Partners' data repositories contributed to the identification of a limited proportion of deceased patients. Massachusetts death data, which records deceased individuals in the state of Massachusetts, contributed to identifying more decedents. We requested the Massachusetts death data from the Department of Public Health for anyone who died between 2011 and 2017. To link the Massachusetts death records to Partners' patients, we used two sets of patient health identifiers with exact matching. The first set is a combination of social security number, gender and the date of birth. Spurious social security numbers

(e.g., 999-999-9999) were removed from both sources before linkage. The second set is a combination of first name, last name, middle names, date of birth, gender, and city of residence.

After we identified a list of ADRD patients with known dates of death, we retrieved their clinical notes documented during their last two years of life. We included all types of inpatient and ambulatory notes, such as office visit notes, progress notes, discharge summaries, emergency department notes, consultations, nutrition notes, social work notes, and phone calls. However, we excluded the notes generated from radiology, pulmonary, and cardiology exams. If one note had multiple versions, we only retrieved its latest version for analysis.

We divided patients' clinical notes into 24 subsets by month. Each note was labeled by the number of months until the date of death. For example, if a note was created during the last month of the patient's life, we labeled it as '01'. The date of a note used to calculate the time distance from death generally refers to the date of service. However, for inpatient encounters, discharge summaries could be signed off days or months later after the date of discharge, or the date could be missing or entered with a default date such as '1900-01-01'. For those exceptional cases, we imputed the date using either the encounter's start date or end date depending which one was closer to the originally recorded date of service. Also, for cases where the service, admission, and/or discharge dates were unavailable from structured tables, we extracted those dates from the clinical notes using regular expressions and rules.

Clinical notes intrinsically contain issues such as formatting, lexical variations, and misspellings. We cleaned the clinical notes with the following steps. First, we employed an existing natural language processing (NLP) tool, MTERMS,[20, 21] to prune the sections related to: medical records numbers, demographic information, dates (admission, birth date), names (e.g., attending physician, nurse), emergency contact information, hospital discharge instructions, immunizations, encounters, payors, and health insurance card numbers. Second, we used a natural language processing tool kit (NLTK)[22] to stem noun phrases and verbs. Third, we split the text into tokens and removed the following tokens: (1) non-alphabetic terms, (2) a subset of stop words from the Mallet package's stop words list, (3) words occurring in fewer than 25 documents (which are usually attributed to misspellings or noise), and (4) words occurring in more than half of the corpus documents (i.e., 'note', 'patient', 'pt').

2.3. Topic Modeling and Analysis

In this study, we used the LDA topic modeling implementation from the Mallet package.[23] We ran Mallet's topic modeling over the corpus three times using the same parameters listed as follows: number of topics = 250, alpha = 0.02, beta = 0.0025, and number of iteration of Gibbs sampling = 1000. From this step, it generated LDA models as well as three output files from each LDA model, including: (1) the 250 topics composed by a list of words with probabilistic distributions and the proportion of the topics over the entire corpus, (2) the predominant topic proportions within each note, and (3) the words and word counts in each topic. The topics generated from the LDA topic model vary each time when LDA is applied to a text corpus due to the separate randomized sampling processes for each. To align them, Shao *et al.* proposed a greedy algorithm to search over the space of triplets generated by combinations of topics from the three sets of topics, for which the max cosine distance of any two topics in the triplets should be smaller than an adjustable parameter.[18] We adopted this approach and set the parameter to 0.7, the same value used in Shao's work. This resulted in an alignment of the three models where spurious topics produced by random noise are discarded, leaving stable triplets identified by one topic from each of the component models. After the stabilization of topics, we calculated the counts of the words in each stable topic by taking the average counts of the words from the triplet of the original topics. In the same way, we also recalculated the proportion of the stable topics in each note by taking the averaged proportion from the three original topics.

To draw the trends of those stable topics over the last two years of life, we calculated the proportion of the topics in the corpus by time slice with the following formula:

$$p(t_{D_i}) = \frac{\sum_{d \in D_i} p(t_d) * s_d}{\sum_{d \in D_i} s_d}$$

where t denotes a topic, i is time slice $i \in \{1, 2, ..., 24\}$, D_i denotes a corpus of time slice i, $p(t_{D_i})$ denotes the proportion of topic t within corpus D_i, and $p(t_d)$ denotes the proportion of topic t within a document d, and s_d denotes the size (word count) of a document d.

2.4. Evaluation

We hypothesize that the trends generated based on LDA-style topic modeling using the clinical text are correlated with the trends derived from well-structuralized data, when data are available in both sources. We used the Pearson's r correlation to statistically test the hypothesis, with an addition of graphically drawing the trendlines for visual comparison. For the evaluation, we selected two LDA-generated topics with one related to a specific diagnosis and

another one related to a medication. We extracted the diagnosis and medication from the structured fields in the EHR and generated the trends for those two selected topics both from clinical notes and structured data based on the same population over the same time span.

2.5. Descriptive Topic Analysis

As a descriptive study, we analyzed the topics with the following steps. First, we provided an overall summary of the topics that had been covered in the clinical notes for the dementia population. We invited two clinical experts, a geriatrician (LF) and a palliative care physician (JL), to classify the topics into different categories as they perceived to be clinically relevant (e.g., physical functional status, end-of-life care). Any disagreements between the two experts were resolved by consensus. Second, we invited the clinicians to identify the topics most relevant to the disease course and patient care near the end of patients' lives and conducted a trend analysis over patients' last two years of life. Third, we conducted a trend analysis across categories to discover potential correlation between topics.

3. Results

3.1. Summary of Dementia Cohort and Clinical Notes

We identified 47,462 patients who were assigned one or more of the ICD9/10 codes listed in Table 1 between 2011 and 2017. Of these patients, 19,845 were deceased (based on the available records) and date of death was available for 19,167 patients. Because we focused on patients' last two years of life, we found that 7,875 (41.1% of 19,167) decedents had clinical notes generated during that period that were available in our data sources. Figure 1 shows the number of eligible patients in each time window, counting back monthly from the time of death, along with the number of patients who had a new diagnosis of dementia recorded in our system, as well as the number of inpatients and outpatients. As shown in Figure 1, the volume of patients who accessed our healthcare services significantly increases in their final stage of life. For example, during the last month of life, 3,841 patients visited our hospitals and clinics, and 36.8% (n=1,414) of them were newly diagnosed with dementia. Table 2 shows the characteristics of the dementia cohort when comparing them to non-dementia decedents in our intuitions in terms of age, gender, race and ethnicity, education attainment, and marital status. The marital status of our studied population had higher proportion in the 'single' and 'widowed' categories comparing to the non-dementia population. As to education level, our studied dementia population had a lower proportion in 'college and above' and a higher proportion in 'high school and equivalent' than the non-dementia group.
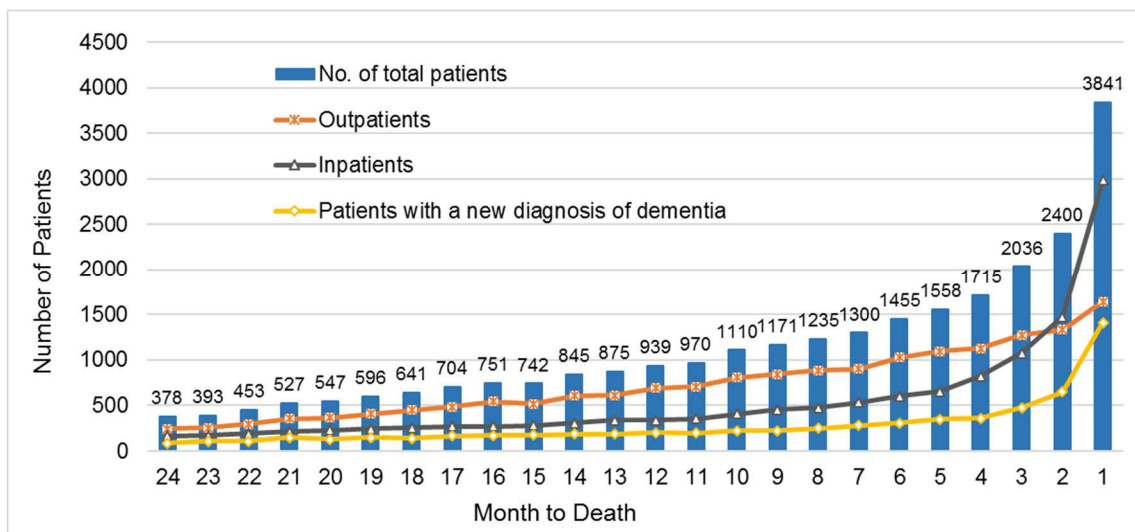

Figure 1. Among the total of 7,875 patients with dementia who had at least one clinical note generated over the last two years of their life, this figure demonstrates the number of patients per month, along with the number of patients with a new diagnosis of dementia recorded in our system, the number of inpatients, and the number of outpatients. This patient cohort was identified from the BWH and the Faulkner Hospital between 2011-2017.

These patients had a total of 432,007 notes. Figure 2.A shows the distribution of the total number of inpatient and outpatient notes by month to death; and Figure 2.B is a box-and-whisker plot that shows the distribution of the number of notes per patient in each time window. Dementia patients had a significantly larger proportion of inpatient notes than outpatient notes in their final stage of life. The number of notes per patient was also increasing while closer to death. During the last month of life, a dementia patient had, on average, 100 notes.

Table 2. Characteristics of decedents with and without dementia

Variables	Dementia	Non-Dementia	Variables	Dementia	Non-Dementia
Overall no.	7,875	133,394	**Marital status** - no. (%)		
Age of death - year	84.3 ± 9.5	71.9 ± 16.5	Married	3,224 (40.9)	63,641 (47.7)
Female gender - no. (%)	4,290 (54.5)	67,844 (50.9)	Single	1,204 (37.4)	26,277 (19.7)
Race or ethnic group - no. (%)			Divorced	536 (6.8)	10,931 (8.2)
White	6,620 (84.1)	101,988 (76.4)	Widowed	2,386 (30.3)	19,907 (14.9)
Black	517 (6.6)	6,419 (4.8)	Others/unknown	525 (6.7)	12,639 (9.5)
Asian	148 (1.9)	1,695 (1.3)	**Educational attainment**- no. (%)		
Other/unknown race	590 (7.5)	23,157 (17.4)	College and above	1,891 (11.3)	28,825 (21.6)
Non-Hispanic*	7,242 (92.0)	110,406 (82.8)	High school or equivalent	2,942 (37.4)	33,050 (24.9)
			Did not complete high school	793 (10.1)	6,694 (5.0)
			Unknown	2,249 (28.5)	64,826 (48.6)

* Ethnicity and race were reported separately

A. Total number of notes by month to death

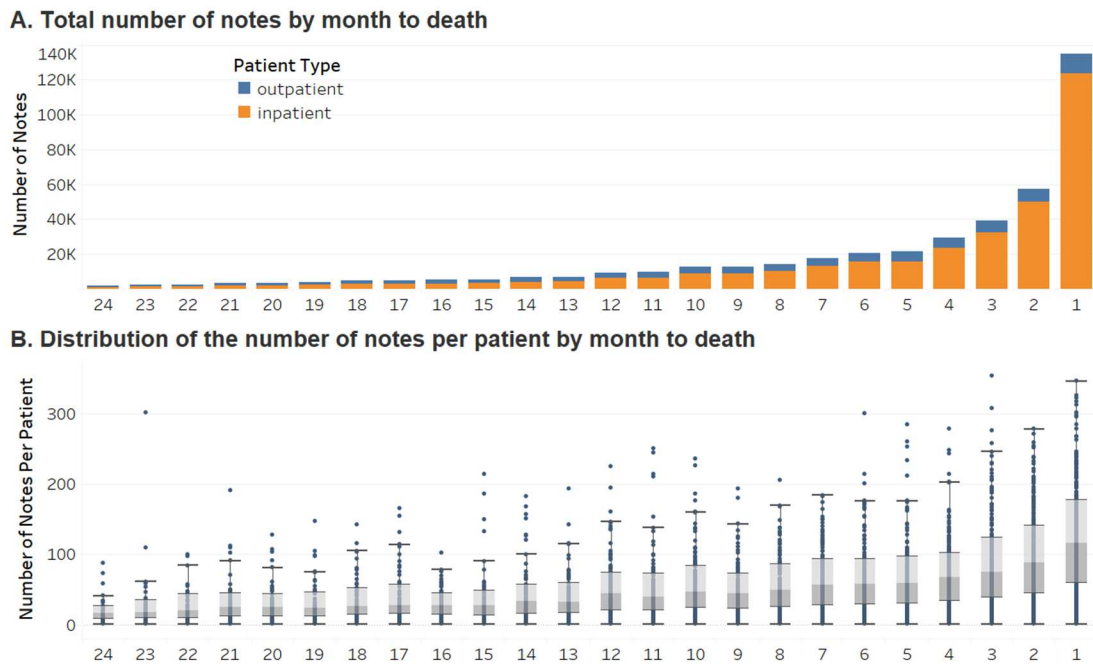**B. Distribution of the number of notes per patient by month to death**

Figure 2. A. Total number of outpatient and inpatient notes by month over the last two years of patients' life; B. A box-and-whisker plot to show the distribution of the number of notes per patient by month.

3.2. Topic Analysis

Among the three sets of 250 topics generated by LDA models, we identified 224 stable topics. Some topics convey similar themes, and the domain experts analyzed all stable topics and classified them into 72 unique categories. For example, 26 topics were categorized as *medication delivery*; 19 topics were categorized as *hospital care*; and other categories composed of multiple topics include: *functional status* (11 topics), *discharge* (11), *nursing care* (10), *laboratory* (10), *intensive care* (8), *physical exam* (7), and *general care* (7). Furthermore, the 72 unique categories can be summarized into following groups: (1) categories related to the disease course such as chronic or other diseases (e.g., *heart failure*, *diabetes*, *cancer*) or acute morbidity (e.g., *aspiration pneumonia, gastrointestinal infection*); (2) categories related to medical care, treatment, and procedures (e.g., *intravenous line*, *skin care*, *biliary care*, *pain management, transplant, physical therapy*); (3) patient assessment (e.g., *imaging*, *cognitive assessment, social history*); (4) categories related to significant events (e.g., *falls*, *fracture*, *stroke, seizure*); (5) categories related to family patient care and support (e.g., *family support*, *end of life care*, *decision making*); (6) categories related to patients' physical functional status (e.g., *mobility*, *swallowing function*), and cognitive function status (e.g., *altered mental status*, *psychiatric disorder*); and (7) other categories related to documentation, insurance, caregiver contact, and so on. Table 3 lists a collection of sample clinician assigned categories along with the top 15 most probable words and the trend analysis groups that they were aggregated into.

Table 3. Examples of stable topics (with top 15 most probable words), categories, and groups

Group	Category	Top 15 Most Probable Words of a Topic
End of life care	Family/hospice care	care family hospice home dnr dementia palliative dni discussion goal intake daughter failure admission thrive
	Comfort care	comfort prn care cmo morphine family hospice measure transition pain comfortable palliative dni dilaudid dnr
	Palliative care	care palliative pain prn continue family delirium time comfort review symptom well management agitation follow
Mental health	Psychiatric disorder*	dementia disorder agitation psych day depakote bipolar home behavioral Seroquel alzheimer sleep husband behavior medication
	Cognitive assessment	cognitive memory year difficulty word test recall dementia history impairment disease attention mild speech function
	Altered mental status*	agitation delirium Haldol seroquel continue prn sleep qhs agitate dementia trazodone medication dose olanzapine psych
	Altered mental status*	status mental alter baseline change acute dementia improve delirium due question command unable recent head
	Altered mental status*	refuse nurse agitate time bed attempt staff continue state call unable haldol progress agitation sleep
Support	Family support	support family provide care social work team meet health contact illness report plan time adjustment
	Spiritual support	spiritual support consult care spirituality outcome chaplain family assessment emotional faith prayer source religion religious
	Caregiver contact	contact phone emergency relation state home unite america mobile information primary address extend prophylaxis access
	Social history	son year report wife home life live state history daughter ago deny alcohol care past
Physical functional status	Mobility*	mobility mouth care medication sit functional bed assist stand transfer assistance balance phone day status
	Sit stand assistance*	sit stand assist assistance level bed ambulation mobility supine cue transfer roll device walker chair
	Gait*	gait normal tremor hallucination day increase decline movement mild memory history wife hand time dementia
Falls, pressure ulcer, and skin care	Fall*	fracture fall rib spine leave injury hematoma head collar trauma chest acute displace compression surgery
	Pressure ulcer	wind ulcer sacral dress decubitus heel stage vac debridement care osteomyelitis foot surgery pressure change
	Skin care*	lesion skin apply rash application well back dress scalp area day dermatology face change top
Nutrition and feeding	Swallow function*	swallow liquid slp aspiration oral solid speech thin dysphagia diet puree consistency thick language nectar
	Artificial feeding	tube gastrostomy feed peg fee jejunostomy hour gtube trach site bid change water goal route
	Nutritional status*	nutrition intake supplement day protein weight diet goal food status meal time malnutrition date continue
Infection	Aspiration pneumonia	pneumonia pna aspiration cxr fever influenza cough vancomycin cefepime admission culture negative day course chest
	Gastro-Intestinal (GI) infection	diarrhea admission diff daily stool hold colitis infection pain abdominal increase day negative secondary home
	Urinary infection	uti urinary infection urine tract day culture status ceftriaxone mental treat antibiotic recurrent daughter acute
	Bacteremia	susceptible culture vancomycin bacteremia resistant blood infection antibiotic mic continue negative ampicillin picc line staph
	Insulin*	insulin unit give diabetes lantus scale subcutaneously slide glucose blood nph meal continue sugar dose

* indicates that these categories are composed of multiple topics that were not exclusively listed here due to the space limitation.

3.3. Evaluation

We compared the trends of the topics generated from the clinical notes to the trends generated from the structured data. Two categories of topics generated from the clinical notes were chosen, and we identified equivalent fields from the structured fields. The topic of aspiration pneumonia was mapped to two ICD-9/10-CM codes: 507.0 and J69.0. We queried the diagnosis tables to identify those patients who were assigned any of those two diagnosis codes among our study cohort. We calculated the proportion of patients who had aspiration pneumonia by time slice. The second set of topics we chose was related to insulin (Table 3). We queried the medication order tables to calculate the proportion of patients in our study cohort that had been prescribed medications under the pharmaceutical class 'insulins'. The trends of aspiration pneumonia and the trends of insulin were compared between two sources of data. The Pearson's r correlation was 0.69 for aspiration pneumonia, indicating a moderate positive correlation, while for insulin it was -0.37, indicating weak negative correlation. As shown in Figure 3.A, the trend based on the topic modeling was similar to the trend based on the ICD codes, although the rate of increase/decrease is slightly different. From Figure 3.B, the patterns of the trends for insulin were similar except in the last 7 months of patients' life.

3.4. Trend Analysis

As shown in Figure 4.A, documentation of altered mental status increased towards the end of life, while the topic trends for cognitive assessment and psychiatric disorder decreased. Figure 4.B shows the trends for physical functional status, including the themes of mobility, sit-standing assistance, and gait. Functional status dropped as patients got closer the death as indicated by decreasing trends of mobility and increasing trend of sit-standing assistance. Figure 4.C shows the trends of patient falls and fracture, skin care, and pressure ulcers. The trend for falls and fracture, while
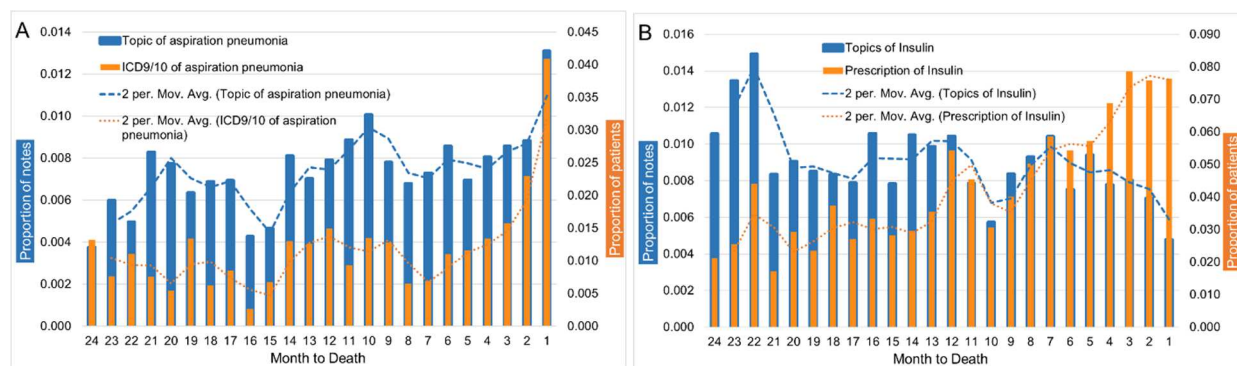
Figure 3. A. Comparison of the trends of *aspiration pneumonia* captured by topic modeling from clinical notes and by ICD9/10 from diagnosis tables; B. Comparisons of the trends of *Insulin* captured by topic modeling from clinical notes and by medication orders.

volatile, was slightly decreasing over time. We saw increasing trends of pressure ulcer while the trend for skin care was stable over the last two years of life. Figure 4.D shows the trends of swallow function, nutritional status, and artificial feeding, all of which were increasing over time. Figure 4.E shows the trends of acute infection in dementia patients during their end of life. Aspiration pneumonia, urinary infection, and bacteremia showed increasing trends while GI infection decreased. In addition, aspiration pneumonia had a higher proportion of notes than other infections. From Figure 4.F, in terms of hospital care, we found that the proportion of notes discussing themes with the topic of discharge and general care decreased as the patient neared death, while topics related to intensive care were increasingly documented in the clinical notes, especially during the last two months of life. Notes contained relatively stable documentation of topics of nursing care and hospital care over time. Figure 4.G shows that all categories related to the end of life care were trending up in the last months of life. The topics about family/hospice care began to increase around a year before death, but documentation of palliative care and comfort care topics rose in the last 2 months of life. From Figure 4.H, the trends for support are generally flat throughout the last 2 years of life. Notably, spiritual support trended up in the last few months of life, especially the last month. The trend for social history was decreasing.

4. Discussion

In this present study, we applied a topic modeling approach to study the clinical notes of ADRD patients at our institution to analyze the disease course during the last two years of life. We generated 224 stable topics covering a wide variety of subjects, including mental status, functional status, pains, chronic diseases (e.g., heart failure, diabetes), acute diseases, falls, physical exam, laboratory, treatments, general care, hospital care, nursing care, intensive care, end-of-life care, and spiritual and family support.

By grouping multiple categories of topics together for analysis, we identified trends among certain topics over the last two years of life. The direction of the trend differed between the various categories: as the patients approached death, an increasing trend was observed for end-of-life care (i.e., hospice/family, palliative care, comfort care), altered mental status, swallow function, artificial feeding, nutritional status, aspiration pneumonia, sit-stand assistance, and pressure ulcers. A decreasing trend was observed for discharge, psychiatric disorder, cognitive assessment, social history, mobility, falls and fracture, and GI infection. An interpretation of these trends follows.

First, the gradual deterioration of dementia patients' mental problems towards death is indicated by the increasing trend of altered mental status (manifested by words such as 'agitation', 'delirium', 'Haldol', 'seroquel' and 'alter') and decreasing trend of cognitive assessment ('cognitive', 'memory', 'year', 'difficulty', 'word' and 'test') and psychiatric issues ('deny', 'depression', 'history', 'mood' and 'psychiatric'). Second, as patients get closer to death, their decline in physical function is indicated by decreasing trends of mobility and an increasing trend of sit-standing assistance. As patients are less mobile, we found a decrease in the documentation of falls and increasing documentation of pressure ulcer likely due to prolonged sitting or decubitus. Third, the increasing trend of swallow function and nutritional status topics findings is also in line with advanced dementia patients' worsening ability to feed themselves and to cope with eating difficulties and weight loss found in previous clinical studies.[3] While trend lines of swallow function and nutritional status are almost parallel (Figure 4.D), the magnitude of increasing notes for swallow function is larger compared to artificial feeding. This discrepancy may reflect the scientific evidence about the lack of benefit of artificial feeding in this population,[18, 19] and the resulting decline in the use of feeding tubes in US nursing homes in the past two decades,[20] demonstrating the coherence between clinical practice and the topics learned from the text.
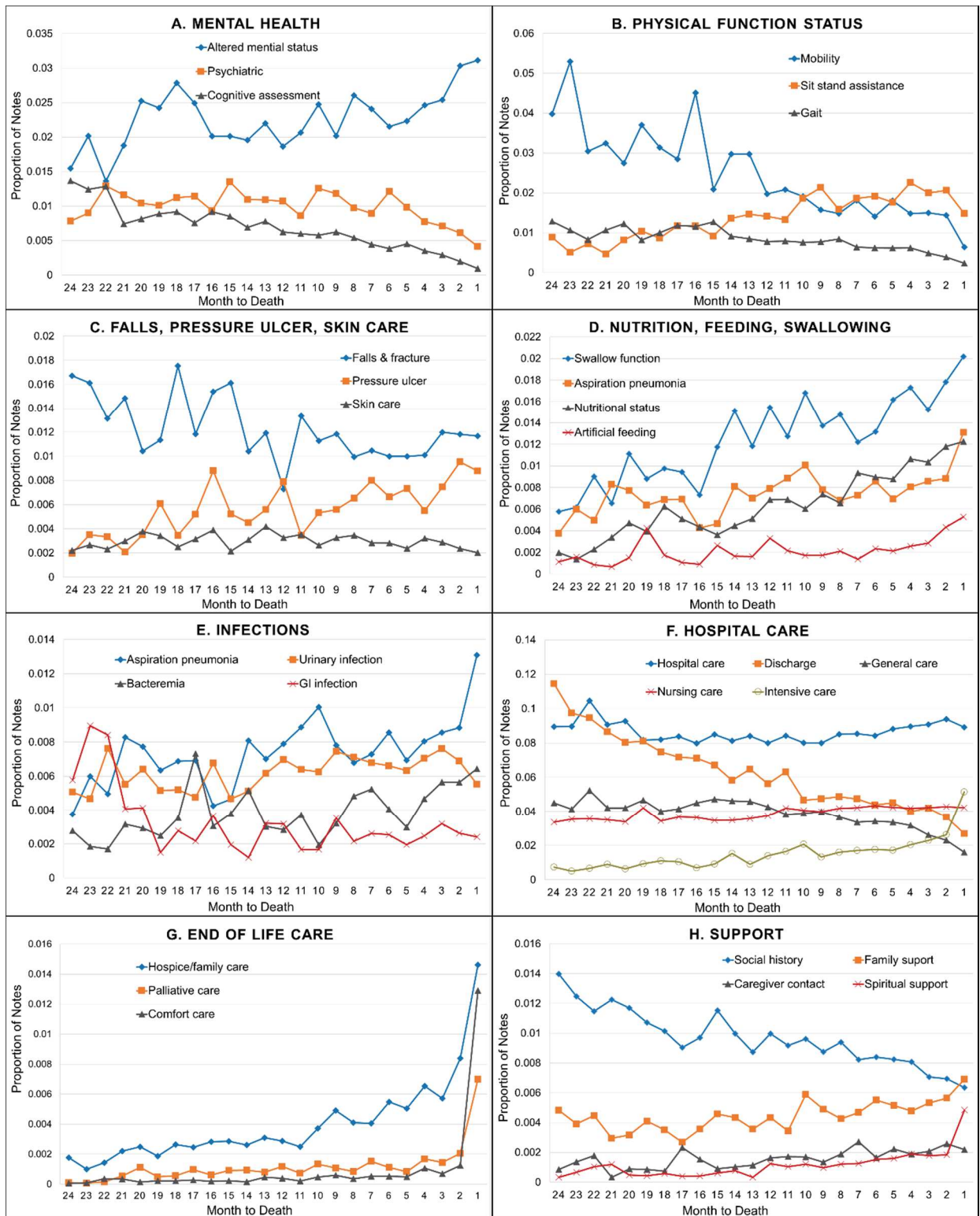
Figure 4. Trend analysis of different groups of topics, including: A. mental health, B. physical function status, C. falls, pressure ulcers, and skin care, D. nutrition, feeding and swallowing, E. infections, F. hospital care, G. end of life care, and H. support.

The concomitant increasing trend of aspiration pneumonia provides an additional possible example for this coherence, reflecting the causal relationship between impaired swallow function and aspiration pneumonia resulting from

inhalation of food or saliva. Fourth, prior literature suggests that dementia patients experience an increasing rate of acute infections in last year of life.[8] In the present study, we found that the major infection documented in the notes is aspiration pneumonia, which is followed by urinary infections, bacteremia, and GI infections. Specifically, the trends for aspiration pneumonia and bacteremia increase during the last five months, while those of GI and urinary infections somewhat surprisingly decrease in the last three months of life. Lastly, trends for intensity of care and end-of-life care topics match what has been demonstrated for other diseases. We found an increasing trend of topics revolving around what may be considered burdensome treatments, such as intensive care and artificial feeding, during dementia patients' last stages of life. Additionally, we also found that documentation of palliative care and comfort care appeared late, during the last two months of life, while documentation of hospice started earlier but rose sharply in the last months of life. This pattern is in line with previous research showing that among dementia patients, conversations about advance care planning typically happen late in the disease course or do not happen at all.[24] Thus, our data suggest an opportunity in our institution for conducting such conversations with dementia patients and their families earlier in the course of their illness.

The trends of the topics generated by our method correlated with those of their respective structured data elements, including diagnoses (aspiration pneumonia) and medications (insulin), demonstrating the ability of this method to accurately capture the patient's clinical trend beyond the content of the structured elements. Mismatch between the trends of the generated topics and the structured element can be attributed to various causes. First, certain structured information elements, such as medications, can be more uniformly mentioned in specific types of notes, such as discharge summaries compared to other types of notes. Thus, the trend of such elements can reflect note type distribution rather than the entities' importance. Second, structured data elements like diagnoses suffer from a well-recognized problem of high false positive rate, which may cause their trend to deviate from that of the narrative documentation. Finally, the trend of topics in clinical notes should be complementary to other analyses used together to capture the disease population more comprehensively.

The current study presents several limitations. While a minimal experiment was conducted to compare the trends discovered from the clinical notes using topic modeling to the trends generated from structured data, this is mainly a descriptive study. Future work may consider a comparison to a control group to distinguish whether the trends identified from this study are unique to dementia or not. The retrospective cohort we studied came from a single healthcare organization (Partners HealthCare), and thus the topics we generated might be organization-specific; however, our results corroborate research done in this population in other settings. Among all patients in our network with a diagnosis of dementia in their last two years of life, only 41% had clinical notes. The absence of this large subpopulation of dementia patients that did not receive subsequent substantial care in our network may simply reflect a transfer to another network,[25] but could also reflect the existence of a subset of dementia patients with a different pattern of healthcare utilization, affecting the generalizability of our findings. Finally, the trending analysis assumes that a mention of a concept reflects its existence or occurrence in the patient. However, linguistic phenomena such as negation, hedging and subject variation can undermine this assumption. In future work, these issues can be mitigated using the appropriate NLP techniques.

5. Conclusion

LDA-based topic modeling is a feasible approach to discover the illness trajectories and end-of-life care for dementia patients using a large corpus of patients' longitudinal clinical notes. The patterns and trends of the generated topics provide unique findings and insights that are often not documented in the structured data fields in the EHR, such as functional status, mental status, and palliative care. In the next step, we plan to study those topics in prediction algorithms to identify patients in need of earlier palliative care interventions, something that appeared to happen late in our study population. Such interventions, when provided to patients earlier in the disease course, are shown to improve patient care and family bereavement outcomes and result in more appropriate use of healthcare resources[26, 27]. The work may provide important insights and guidance for timely and effective palliative care interventions for patients with dementia.

References

1. Association As. 2017 Alzheimer's disease facts and figures. Alzheimer's & Dementia. 2017;13(4):325-73.
2. Aronson MK, Ooi WL, Geva DL, Masur D, Blau A, Frishman W. Dementia: age-dependent incidence, prevalence, and mortality in the old old. Archives of Internal Medicine. 1991;151(5):989-92.
3. Whitehouse PJ, Price DL, Clark AW, Coyle JT, DeLong MR. Alzheimer disease: evidence for selective loss of cholinergic neurons in the nucleus basalis. Annals of neurology. 1981;10(2):122-6.
4. Alzheimer's A. 2016 Alzheimer's disease facts and figures. Alzheimers Dement. 2016;12(4):459-509.
5. Mitchell SL, Teno JM, Kiely DK, Shaffer ML, Jones RN, Prigerson HG, et al. The clinical course of advanced dementia. New England Journal of Medicine. 2009;361(16):1529-38.
6. Tolea MI, Morris JC, Galvin JE. Trajectory of Mobility Decline by Type of Dementia. Alzheimer Dis Assoc Disord. 2016;30(1):60-6. PMCID: PMC4592781
7. Evers MM, Purohit D, Perl D, Khan K, Marin DB. Palliative and aggressive end-of-life care for patients with dementia. Psychiatric Services. 2002;53(5):609-13.
8. Morrison RS, Siu AL. Survival in end-stage dementia following acute illness. JAMA. 2000;284(1):47-52.
9. Hurley AC, Volicer BJ, Volicer L. Effect of fever-management strategy on the progression of dementia of the Alzheimer type. Alzheimer Dis Assoc Disord. 1996;10(1):5-10.
10. Jethwa KD, Onalaja O. Advance care planning and palliative medicine in advanced dementia: a literature review. BJPsych Bull. 2015;39(2):74-8. PMCID: PMC4478901
11. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.
12. Blei DM. Probabilistic topic models. Communications of the ACM. 2012;55(4):77-84.
13. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National academy of Sciences. 2004;101(suppl 1):5228-35.
14. Hofmann T. Probabilistic latent semantic analysis. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence; Stockholm, Sweden. 2073829: Morgan Kaufmann Publishers Inc.; 1999. p. 289-96.
15. Stieglitz S, Dang-Xuan L. Social media and political communication: a social media analytics framework. Social Network Analysis and Mining. 2013;3(4):1277-91.
16. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus. 2016;5(1):1608. PMCID: PMC5028368
17. Zeng QT, Redd D, Rindflesch T, Nebeker J, editors. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annual Symposium Proceedings; 2012.
18. Shao Y, Mohanty AF, Ahmed A, Weir CR, Bray BE, Shah RU, et al. Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure. AMIA Annu Symp Proc. 2016;2016:1110-8. PMCID: PMC5333331
19. Jo Y, Lee L, Palaskar S. Combining LSTM and Latent Topic Modeling for Mortality Prediction. arXiv preprint arXiv:170902842. 2017.
20. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc. 2011;2011:1639-48. PMCID: PMC3243163
21. Zhou L, Baughman AW, Lei VJ, Lai KH, Navathe AS, Chang F, et al. Identifying Patients with Depression Using Free-text Clinical Documents. Stud Health Technol Inform. 2015;216:629-33.
22. Bird S, Loper E, editors. NLTK: the natural language toolkit. Proceedings of the ACL 2004 on Interactive poster and demonstration sessions; 2004: Association for Computational Linguistics.
23. McCallum AK. MALLET: A Machine Learning for Language Toolkit; 2002 [cited Feburary 5, 2018]. Available from: http://mallet.cs.umass.edu/.
24. Detering KM, Hancock AD, Reade MC, Silvester W. The impact of advance care planning on end of life care in elderly patients: randomised controlled trial. Bmj. 2010;340:c1345.
25. Gozalo P, Teno JM, Mitchell SL, Skinner J, Bynum J, Tyler D, et al. End-of-life transitions among nursing home residents with cognitive issues. N Engl J Med. 2011;365(13):1212-21. PMCID: PMC3236369
26. Zhang B, Wright AA, Huskamp HA, Nilsson ME, Maciejewski ML, Earle CC, et al. Health care costs in the last week of life: associations with end-of-life conversations. Arch Intern Med. 2009;169(5):480-8. PMCID: PMC2862687
27. Mack JW, Weeks JC, Wright AA, Block SD, Prigerson HG. End-of-life discussions, goal attainment, and distress at the end of life: predictors and outcomes of receipt of care consistent with preferences. J Clin Oncol. 2010;28(7):1203-8. PMCID: PMC2834470