Original Research Article

Using personal writings to detect dementia: A text mining approach

Beni Asllani[ID] and Deborah M Mullen[ID]

Department of Management, The University of Tennessee at Chattanooga, Gary W. Rollins College of Business, Chattanooga, TN, USA

Abstract

A novel text mining pilot for dementia detection using Linguistic Inquiry and Word Count (LIWC) was tested on public figures' writings looking at word choice and affect compared to those with and without dementia. The differences found in this analysis mirror the expected patterns where writings of people with dementia reflect significantly more analytical thinking words, but significantly less authentic and emotional tone. In addition, the analysis found that people with dementia use significantly less functional words, such as grammar, and affections (happiness, sadness, anger, sadness), but tend to use significantly more pronouns in their writings. Written samples of those with dementia also use significantly less time-oriented words that indicate past, present, or future. The analysis of free form text suggests a potential avenue for detecting early changes that correlate with dementia, allowing for early preventative treatment before noticeable cognitive impairment occurs.

Keywords

dementia, linguistic inquiry and word count, text-mining

Introduction

Globally dementia is one of the leading causes of disability in older people.[1] Dementia, characterized by a progressive decline of cognitive functions in thinking and memory, leads to behavioral changes and the inability to engage in daily activities.[1] The most common type of dementia, Alzheimer's disease, affects 60%–80% of all patients.[2] Risk factors for dementia include family

Corresponding author:
Deborah M Mullen, Department of Management, The University of Tennessee at Chattanooga, Gary W. Rollins College of Business, 615 McCallie Ave, Chattanooga, TN 37403, USA.
Email: deborah-mullen@utc.edu

history, age, brain trauma, genetic mutations, cardiovascular disease, chronic use of medications that mimic dementia (anticholinergic agents), and lower educational attainment.[3] Dementia diagnoses require a thorough review of personal health history, physical exam, lab. values, brain scans, and functional movement assessments, as well as cognitive testing such as a verbal fluency test, the "Mini-Cog," or other instruments.[2,3]

Clinical use of the "Mini-Cog" mini or other verbal diagnostic tests asks patients to recall words, items, and concepts from short or long-term memory as indicators of cognitive impairment. Patients with dementia often struggle with object names, active verb usage and are prone to repetitive words and linguistically useless phrases such as "um".[4,5] The verbal fluency test asks patients to list as many animals as they can in 60 s with <=15 animals listed signaling potential dementia (<=12 if only 1–7 years of education and <=9 with no education).[3] The "Mini-Cog" test asks patients to remember three unrelated words (i.e., river, nation, finger), draw the time of day on a clock face, and then recall the three words. A Mini-Cog score of 0–2 correlates highly with a diagnosis of dementia (each item recalled and an accurate clock is each worth one point apiece).[3] Verbal cognitive testing, widely used in healthcare, is unreliable in cases of minimal (onset) impairment and is prone to average age-related cognitive decline.[2,3] Commonly patients and family members do not mention cognitive impairment to clinicians until the middle or late stages of dementia, where cognition and memory issues have already compromised their health and well-being.[6] While there are no cures for dementia, early detection is critical for delaying the progression through active health management, including promising new drugs that may postpone impartment.

Early detection allows for more intensive interventions before disability. Researchers have been searching for methods to detect minimal early cognitive effects as an early warning sign.[7,8] Efforts have focused on three specific approaches: medical records data mining,[9,10] general internet text data mining,[11] and population-wide preventative individual screening.[12] Researchers have examined clinician visit notes within electronic medical records with limited success for reported observations of dementia-like pre-curser patient behaviors combined with sequential brain images.[9,10,13,14] Regardless of the model or analysis technique (machine learning, predictive models, structured, and unstructured analysis), these approaches utilize limited medical record data. In most cases, once dementia is suspected, data collection occurs without a comparative baseline measurement. In clinical practices, there is little time given to collection of baseline data on cognition or functional baseline measurement for two reasons: time constraints and a lack of reimbursement for collecting baseline data that may or may not be needed later, thus the absence of a comparison data set.

Bull et al. (2016) propose including data collected from individuals outside of medical care such as email, documents, and diaries as a method to diagnose dementia.[11] Bull et al.[11] suggest that these electronic writings combined with personalized patterns of keystrokes and mouse movements might indicate dementia. The slowing down or alteration of keystrokes and mouse movements can indicate signs of dementia, and this approach is currently undergoing trials.[15] Another suggested approach to diagnose dementia is evaluating idea density (low density = cognitive impairment). Idea density describes how efficiently a person conveys ideas without repeating phrases or producing vague responses.[16–19] Low grammatical complexity and negativity in writings can predict an eventual dementia diagnosis.[16–19] These earlier studies have been challenging to implement due to their reliance on the longitudinal collection of long-form writings augmented with clinical diagnostic supporting findings. Le et al.[17,20] have proposed a short-form text analysis and have patented a hierarchical model of text analysis. However, they have yet to create a working model of the proposed analysis.

Researchers have also studied a verbal test administered to a broad population outside of a clinical environment.[21–23] These verbal tests use word vector mapping in everyday conversations or conversations with interactive avatars and time series of conversational word repetition as methods for dementia detection.[21–23] While successful, these methods also require specialized efforts, longitudinal tracking (including a need for a personal baseline), and large-scale computing to successfully identify potential dementia-related cognitive issues. Text-based methods are more usable on broad populations than spoken analysis due to the relatively small samples needed (200 words minimum) and the wide availability of text samples such as social media posts. Finally, computer scientists have suggested using advanced data mining and text analysis to detect dementia.[24,25] However, these researchers still depend upon evaluating patient responses to the neuro-cognitive tests traditionally utilized in health care, even with their inaccuracy in diagnosis.

The promise of artificial intelligence (AI), machine learning, and natural language processing in detecting early changes in cognition is the possibility for broader population screening.[26,27] While there are limitations to the data contained in the medical record, the availability of personal writings in the form of blogs, transcriptions from videos posted on social media (TikTok, Facebook live), social media postings, emails, and other communication offer a potential source of more natural, less edited and biased communication to screen. Other methods of potential interest for implementation in electronic health/medical records (EHR/EMR) are additions of population initial screening via speech generation tasks (a scripted reading read aloud by patients) and recorded for AI analysis.[26,27] Even when screening in the clinical setting, early signs are easy to miss due to patients being able to use cognitive reserves to camouflage symptoms, making full population, baseline and ongoing screening an advantage for clinicians and a potential area for health informatics to be assist clinical care.[28]

Recognizing that early detection is vital and that most patients and their families will not ask for a diagnostic exam until impairment is significant, we propose a text-based model utilizing short-form personal writings for early dementia detection. Utilizing various text and data mining methods, we evaluate personal writings for their style, psychological affect, and personal demographic factors, focusing on those associated with an eventual diagnosis of dementia. This work seeks to identify a pattern of pre-dementia diagnosis cognitive issues in text files in early life when prevention efforts could delay dementia onset and advancement.

Methods

This study utilizes the personal writings (number) of public figures due to the high probability that their writing samples exist and were publicly available from different periods of their lives (Appendix A). Preferred data samples are un-edited (letters and diaries) rather than edited speeches or published works. As a control group, we also collected writings of public figures who did not die from dementia. Public figures with dementia and control group writings were matched based on gender and year of birth based on the belief that those of the same gender and generation would be more alike in shared-world experiences. Sample cases were in English and predominantly from the United States. Basic demographic data (dates of birth & death, occupation, highest level of education completed, dementia/no dementia diagnosis) were collected, and writing samples from three periods. The three periods include early life <= 29 years of age, middle life = 30 years of age to 11 years before the date of death, late life = 10 years before death.

Linguistic Inquiry and Word Count (LIWC) software program evaluation occurred once all the text file writing samples were collected. LIWC is a text analysis program that matches text samples to dictionaries marking each word in a sample with up 84 different variables (each word can fall into multiple categories) within 27 dimensions (17 linguistic and 10 relativity) and 44 categories (25 psychological constructs and 19 personal concerns).[29] "Cried" would be coded as a negative emotion, sadness, overall affect, and past tense verb in the analysis output.[29] These categories tested positively for internal consistency and validity.[30]

This descriptive study tests if text-based analysis detects differences between those with dementia and those without dementia. If there are noticeable differences, especially those aligned with expected changes caused by dementia over the life span, this model may provide a fruitful method of identifying early dementia effects in affect from writing samples. This early warning would need follow-up cognitive testing and clinical evaluation. However, it could ensure that if found, early-stage medications can begin prior to significant impairment or cognitive decline when they are most effective at halting disease progression.

Dementia versus non-dementia writings

The dementia group names mainly came from a list maintained by the Alzheimer's association and were augmented through search engine queries "people who died of dementia" and "people who died of Alzheimer's disease", resulting in 404 entertainers (actor, writer, director, singer, and musicians), politicians, and sports figures (athletes and coaches).[31] Of these, 64 born in non-English speaking countries were excluded from the list since early writings were not English; most were U.S. born, followed by U.K., and Australia. Searches for primary unedited writings excluded even more figures from this list. It was common to find edited samples from authors and politicians, which can alter tone and wording. Sixty-two individuals with dementia writings over two or three periods (early writings were the most difficult to find) are included.

The matched cohort of figures had a cause of death unrelated to dementia. Public figures with similar education levels, same gender, same age (within 1–2 years), and when possible, the same profession (Table 1). This matching schema assumed that contemporaries had similar educational experiences and were products of a similar generational experience and global zeitgeist. Text has been collected for 62 deceased public figures who were diagnosed with dementia (early $n = 12$, middle $n = 16$, late $n = 59$) and 22 matched non-dementia cases (early $n = 9$, middle $n = 20$, late $n = 20$).

To answer the research question, is there a textual difference detectable with LIWC between those who develop dementia and the non-dementia group, we evaluated the 18 summary variables that represent 84 individual variable counts using SPSS v26. We tested for normality and found the Shapiro-Wilk test to be significance for the majority of the dependent variables. As a result, we tested our hypothesis using a series of non-parametric tests for independent samples. The first analysis matched dementia and without dementia groups with a single file for each individual and included life period; results shown in the tables. Secondary analysis included life periods (early, middle, and late), comparison of only letters (the least likely text to have been edited), as well as text length comparisons. Results of the secondary analyses are included only when significant.

Table 1. Demographics of dementia and no dementia cohorts.

Demographics	Dementia	No dementia
Gender		
Female	20	6
Male	42	16
Highest education attended		
Home school	0	1
Middle school	3	0
High school	11	6
Theatre/arts school	7	2
College	9	5
Bachelors	16	4
Masters	6	2
Law school	3	1
Post graduate	6	1
Unknown	1	0
Career		
Activist	1	0
Actor/director	19	9
Astronaut/military	1	1
Artist	4	0
Athlete/coach	2	1
Author/poet	13	3
Composer, singer	2	2
Entertainer	0	1
Evangelist	1	1
Journalist	2	0
Philanthropist	0	1
Politician	12	2
Professor	2	0
Scientist	3	1

Results

Composite metrics

Linguistic inquiry and word count includes composite measures built from the subscales proprietary to LIWC and thus not user definable. The composite measures are analytical thinking, clout, authentic, and emotional tone. As shown, the Shapiro-Wilk is significant for "Analytics," "Authentic, and "Emotional Tone." Specifically, those with dementia use significantly more analytical thinking words (formal, logical, hierarchical patterns) ($p = .004$), significantly lower authentic wording ($p = .020$), and significantly less emotional tone ($p = .003$) in their writing (Table 2). The emotional tone is a composite measure of negative and positive emotions (scaled so that lower rates are more negative in tone, especially below a sum of 50).[32] Authentic wording includes those that indicate feelings of humility and personal vulnerability.

Table 2. Dementia versus no dementia composite LIWC measures.

LIWC composite dimensions	Mean Word Usage		Tests of normality						Kruskal-Wallis
			Kolmogorov-Smirnov[a]			Shapiro-Wilk			
	Dementia	No dementia	Statistic	df	Sig.	Statistic	df	Sig.	Sig.
Analytic	65.82	53.98	0.078	136	0.041	0.964	136	0.001	0.004
Clout	72.27	71.07	0.074	136	0.067	0.965	136	0.001	0.777
Authentic	37.51	47.76	0.078	136	0.042	0.964	136	0.001	0.020
Emotional tone	70.56	84.65	0.237	136	0.000	0.813	136	0.000	0.003

[a]Lilliefors significance correction.

Additional analysis of the letters only samples, dementia ($n = 46$) and non-dementia ($n = 46$), using a Quade Non-Parametric ANCOVA, found that only the 'Analytic' dimension is significantly higher ($p = .024$) for people with dementia as compared to those with no dementia. When compared by the age range, the letters-only samples show that the 'Emotional Tone' dimension was significantly lower in middle age compared to the early age ($p = .047$) of the people who had dementia. No significant differences existed between the age ranges for the other composite dimensions.

Communication patterns

Individuals with dementia use functional communication words differently than those without dementia (Table 3). LIWC functional dimension includes function, total pronoun, (other) grammar, affect, and social sub-categories. Those with dementia use significantly less function words ($p = .039$) except total pronouns. Total pronouns include personal, 1st person singular and plural, 2nd person, 3rd person singular and plural, as well as impersonal pronouns. Dementia writings included significantly more usage of total pronouns ($p = .003$) but less "other grammar" words than those without dementia ($p = .002$). "Other grammar" words including regular verbs, adjectives, comparatives, interrogatives, numbers, and quantifiers are used less frequently in dementia writings. In addition, those with dementia show significantly less affect in their writing ($p = .000$) overall; words referring to positive and negative emotions including anxiety, anger, and sadness. There are no notable differences in social word usage (family, friends, female and male referents).

Communication with others includes informal speech and, when written, punctuation. Informal speech includes swearing, netspeak, assent, non-fluencies ("um," "hm," "er"), and fillers ("you know," "I mean," and other meaningless words) (Table 3). Netspeak refers to shorthand abbreviations used in text messaging such as "LOL" for "laugh out loud" and 'XOXO' for "hugs and kisses." Those with dementia use more informal words and punctuation, however the rate was not-significant (Table 3).

Secondary analysis of the letters-only samples for both dementia and no dementia using a Quade Non-Parametric ANCOVA found that "Grammar" ($p = .025$) and "Affect" ($p = .038$) are still significantly less in the "Dementia" group versus "No Dementia" group. When the letter test files were compared among group ages, "affect" was significantly lower among all age ranges ($p = .033$),

Table 3. Functional, informal speech, and punctuation scale LIWC.

| LIWC Function, language, and punctuation dimensions | Mean Word Usage | | Tests of normality | | | | | | Kruskal-Wallis |
| | | | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | | |
	Dementia	No dementia	Statistic	df	Sig.	Statistic	df	Sig.	Sig.
function	49.97	53.63	0.191	136	0.000	0.693	136	0.000	**0.039**
Total pronouns	25.50	17.00	0.061	136	.200[b]	0.977	136	0.022	**0.003**
Grammar	29.52	45.00	0.100	136	0.002	0.956	136	0.000	**0.002**
Affect	6.27	8.25	0.083	136	0.023	0.969	136	0.004	**0.000**
Social	10.84	11.86	0.048	136	.200[b]	0.988	136	0.301	0.263
Informal Language	1.08	0.54	0.318	136	0.000	0.460	136	0.000	0.074
All Punctuation	18.05	15.12	0.215	136	0.000	0.582	136	0.000	0.644

Function includes: total pronouns (personal pronouns, 1st person singular, 1st person plural, 2nd person, 3rd person singular, 3rd person plural, and impersonal pronouns); grammar other (regular verbs, adjectives, comparatives, interrogatives, numbers, and quantifiers): affect (positive emotion, negative emotions including anxiety, anger, sadness); social (family, friends, female referents, male referents); informal language (swear words, netspeak, assent, nonfluencies, fillers); all punctuation (periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses (pairs), other punctuation).
[a]Lilliefors significance correction.
[b]lower bound of the true significance.

especially when comparing the early to middle-life age ranges ($p = .028$). Analysis of the length of the text files showed no significant difference for the other measures.

Processes, drives, orientation, relativity, and concerns

Table 4 includes analysis of the dimensions for processing (cognition, perceptual, biological), core drives and needs, time orientation, relativity, and personal concerns. The only significant difference between dementia writings and non-dementia are words related to time orientation (past, present, and future focuses). Those with dementia use significantly fewer time-oriented words ($p = .016$).

Dementia and non-dementia writings show no differences in the cognitive (insight, cause, discrepancies, tentativeness, certainty, differentiation), perpetual (sight, hearing, touch) or biological (body, health/illness, sexuality, ingesting) process word usage. Nor are there significant differences between groups in usage of the core drives and needs, relativity, nor personal concerns wording. Core drives and needs words refer to (affiliation, achievement, power, reward focus, risk/prevention focus). Relativity in LIWC analysis includes terms related directly to relativity (motion, space, time). Personal concerns relate to work, leisure, home, money, religion, and death.

Analysis of letters only and age range showed no significant difference for the composite measures.

Discussion

Linguistic inquiry and word count has been used to identify underlying personality characteristics and effects of behavioral health diagnosis and social relationships in research utilizing text analysis,

Table 4. Drives, relativity, and processes (cognitive, perpetual, bodily).

LIWC Processes, drives, time orientation, relativity, and personal concerns dimensions	Mean word usage		Tests of normality						Kruskal-Wallis
			Kolmogorov-Smirnov[a]			Shapiro-Wilk			
	Dementia	No dementia	Statistic	df	Sig.	Statistic	df	Sig.	Sig.
Cognitive processes	9.65	10.95	0.082	136	0.025	0.981	136	0.059	0.118
Perceptual processes	2.58	2.16	0.112	136	0.000	0.916	136	0.000	0.299
Biological processes	1.28	1.35	0.179	136	0.000	0.816	136	0.000	0.235
Drives	6.69	7.05	0.084	136	0.021	0.965	136	0.002	0.742
Time orientation	5.13	6.02	0.047	136	0.200[b]	0.972	136	0.007	0.016
Relativity	12.53	13.31	0.052	136	0.200[b]	0.983	136	0.081	0.548
Personal concerns	18.46	10.73	0.108	136	0.001	0.934	136	0.000	0.583

Processes include: Cognitive (insight, cause, discrepancies, tentativeness, certainty, differentiation); Perpetual (seeing, hearing, feeling); Biological (body, health/illness, sexuality, ingesting); Core Dirves and Needs (affiliation, achievement, power, reward focus, risk/prevention focus); Time Orientation (past focus, present focus, future focus); Relativity (motion, space, time) Personal Concerns (work, leisure, home, money, religion, death).
[a]Lilliefors significance correction.
[b]lower bound of the true significance.

word categorization, and counts.[29] LIWC's ability to detect small changes that are less memory dependent but more affect and temporal such as thinking and focus, make LIWC a potentially powerful tool for screening dementia-related changes before significant, noticeable impairment. LIWC validation and psychometric testing show that in the normal aging process people in general become less self-centered, more current and future time focused, and that verbal complexity maintains or increases with age.[29] Broad usage of the "Mini-Cog" to detect cognitive changes common from dementia may only detect significant and late-stage disease after the time when the only current dementia treatment (medications) are less effective; they are "too little, too late."[3] This project attempts to use personal writings throughout a lifetime to see if there is a textual indicator of changes in emotion, wording, and grammar as an early indicator of upcoming cognitive issues such as dementia.

Those with dementia use more analytical words with less authentic and emotional tone wording. Text samples in the dementia group were 23% profession-related writings and 4% for those without dementia. Profession-related writings are more likely to be edited, more formal perhaps leading to the more analytic wording. Analytical wording is related to formal, logic-based, and hierarchical thinking. Formality and professionalism do not, however, explain the differences in emotional tone, affect, or authenticity. When letters only were compared for the two samples, dementia writings were still significantly different including less affect, emotional tone, and grammar. Dementia writings include more logic, hierarchical wording along with less words in the composite authentic and emotional tone scales. These words indicate that dementia writings are more guarded, socially distanced, and more neutral with less emotion. Lesser use of functional words by those with dementia includes less use of other grammar and affect words, but more total pronouns words overall. When dementia writings were compared for age ranges, affect was the only significant scale and then only between early to late life samples. As dementia advances communication and language losses increase causing difficulties in making semantic and conceptual distinctions; this

progression begins with difficulty in naming, then grammatical, and finally comprehension as the disease progresses.[33] Our results mirror these findings, with those experiencing dementia relying on simplified pronouns and less grammar and affect wording, complex semantic and conceptual distinctions that amplify over time. Dementia texts include fewer time-orientation words as well. Taken together, these more negative and formalized communications written by those with dementia correspond with a less personal, time-oriented, and personalized narrative style of writing and confirm that there is something different about the communication for those with dementia than that expected simply by the aging process.[29]

Prior research has shown dementia-related effects on written communication, including a marked increase in negativity with low idea density and low grammatical complexity.[16–19] Our findings support those findings, with individuals diagnosed with dementia using more pronouns overall. In cases where writers struggle with time orientation (especially present and future) and may struggle to recall names, use of less grammar and proper names make communication easier. A writer without cognitive decline is more likely to use a variety of words and phrases to refer to others in correspondence to ensure flow and to avoid repetition. As predicted by others' research, these word choices, frequencies, and grammatical differences indicate a more general use of written language with less emotion and less complexity.[16–19,22,29] Current cognitive testing looks for dementia signs such as lost object naming, less active verbs, word repetition, and the use of non-fluencies.[3,4] Our findings support these frequencies in text (more pronouns instead of object/subject naming, fewer active verbs).

As people move into more significant impairment due to dementia, limiting their movement due to safety concerns is common. This repeated place, schedule, and predictable days minimize confusion and stress.[34] Advancing dementia also reduces person/object/memory recall and awareness, leading to isolation, potentially leading to a negative emotional state and less personal communication.[35] As the disease progresses and the person's world contracts, the most vivid memories are often from the distant past stored in long-term memory.[2]

This work's limitations include the reliance on personal writings or various lengths and people of differing education levels. Middle and late-life samples are more likely to have been reviewed for some of these public figures by personal assistants, secretaries, or family members. For this reason, whenever possible personal correspondence was utilized and not more formal job-related letters. The samples were of different lengths and different purposes. When writing to close family and friends or in unpublished diaries, there is a less formal tone. While length of text is variable LIWC analysis samples studied were at least 50 words of more in length (98.5% for dementia writings; 100% for non-dementia writings).[30,32] In the secondary analysis, greater word count or length of text lead to more significant differences and findings. Educational attainment could impact these findings as education influences vocabulary, grammar, and style.[36,37]

Conclusion

Linguistic inquiry and word count was able to identify significant word choices and affect from written communications. There were notable differences between those who had dementia and those without dementia. The differences mirror the expected patterns where people with dementia show more analytic language, less authentic wording, less emotional tone, fewer function words, more total pronouns, less grammar, less affection, and less time orientation than those without dementia. Usage of the brief, unedited written word is a potential tool for identifying early shifts potentially indicative of early dementia symptoms. Future studies need to look at the use of blogs, tweets, text

messages to see if there is a minimum necessary text needed and if current communication patterns are mappable to suspected dementia-related wording shifts.

ORCID iDs

Beni Asllani https://orcid.org/0000-0001-5425-6763
Deborah M Mullen https://orcid.org/0000-0001-6345-2889

Supplemental Material

Supplemental material for this article is available online.

References

1. WHO. *Dementia*. Geneva, Switzerland: World Health Organization, 2020. https://www.who.int/en/news-room/fact-sheets/detail/Dementia
2. Alzheimer's Foundation. *What-is-Dementia*. Geneva, Switzerland: World Health Organization, 2020. https://www.alz.org/alzheimers-Dementia/what-is-Dementia2020
3. Schroeder RW, Martin PK and Walling A. Neuropsychological evaluations in adults. *Am Fam Physician* 2019; 99(2): 101–108.
4. Appell J, Kertesz A and Fisman M. A study of language functioning in Alzheimer patients. *Brain Lang* 1982; 17(1): 73–91.
5. Hamilton HE. *Conversations with an Alzheimer's patient: an interactional sociolinguistic study.* Cambridge: Cambridge University Press, 2005.
6. Tsoi KK, Chan JY, Hirai HW, et al. Cognitive tests to detect dementia: a systematic review and meta-analysis. *JAMA Intern Med* 2015; 175(9): 1450–1458.
7. Ienca M, Vayena E and Blasimme A. Big data and dementia: charting the route ahead for research, ethics, and policy. *Front Med* 2018; 5: 13.
8. Roantree M, O'Donoghue J, O'Kelly N, et al. Mapping longitudinal studies to risk factors in an ontology for dementia. *Health Informatics J* 2016; 22(2): 414–426.
9. Bullard J, Murde R, Yu Q, et al. *Inference from structred and unstructured electronic medical data for early dementia detection*. Rochester Institute of Technology - RIT Scholar Works, Student and Faculty Works, 2015. www.scholarworks.rit.edu
10. Moreira LB and Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Comput Methods Programs Biomed* 2018; 165: 139–149.

11. Bull CN, Asfiandy D, Gledson A, et al. *Combining data mining and text mining for detection of early stage dementia: the SAMS framework*. UK: School of Computing and Communications, Lancaster University, 2016. http://ucrel.lancaster.ac.uk/sams/

12. Licher S, Leening MJ, Yilmaz P, et al. Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. *Am J Psychiatry* 2019; 176(7): 543–551.

13. Bansal D, Chhikara R, Khanna K, et al. Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia computer science* 2018; 132: 1497–1502.

14. Ernecoff NC, Wessell KL, Gabriel S, et al. A novel screening method to identify late-stage dementia patients for palliative care research and practice. *J Pain Symptom Manage* 2018; 55(4): 1152–1158.

15. Couth S, Stringer G, Leroi I, et al. Which computer-use behaviours are most indicative of cognitive decline? Insights from an expert reference group. *Health Informatics J* 2019; 25(3): 1053–1064.

16. Jarrold W, Peintner B, Wilkins D, et al. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, Baltimore, MD, June 2014.

17. Le X, Lancashire I, Hirst G, et al. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Lit Ling Comput* 2011; 26(4): 435–461.

18. Pakhomov S, Chacon D, Wicklund M, et al. Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behav Res Methods* 2011; 43(1): 136–144.

19. Snowdon DA, Kemper SJ, Mortimer JA, et al. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *JAMA* 1996; 275(7): 528–532.

20. Hirst GJ, Jokel R, Lancashire DI, et al. *Method and system of longitudinal detection of dementia through lexical and syntactic changes in writing*. Washington, DC: Google Patents, 2016.

21. Mirheidari B, Blackburn DJ and Walker T. Detecting signs of dementia using word vector representations. In: Blackburn D, Walker T, Venneri A, et al. (eds) Interspeech 2018 Conference September 2-6, 2018, held in Hyderabad, India, 2018.

22. Shinkawa K and Yamada Y. Word repetition in separate conversations for detecting dementia: a preliminary evaluation on data of regular monitoring service. *AMIA Jt Summits Transl Sci Proc* 2018; 2017: 206–215.

23. Tanaka H, Adachi H, Ukita N, et al. Detecting dementia through interactive computer avatars. *IEEE J Transl Eng Health Med* 2017; 5: 1–11.

24. Maroco J, Silva D, Rodrigues A, et al. Data mining methods in the prediction of dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes* 2011; 4(1): 299–314.

25. SR BS, Sheshadri H and Joshi S. A review on the method of diagnosing Alzheimer's disease using data mining. *Int J Eng Res* 2014; 3(3).

26. de la Fuente Garcia S, Ritchie C and Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *J Alzheimers Dis* 2020; 78: 1547–1574; Preprint.

27. Voleti R, Liss JM and Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J Sel Top Signal Process* 2020; 14(2): 282–298.

28. Meng X and D'arcy C. Education and dementia in the context of the cognitive reserve hypothesis: a systematic review with meta-analyses and qualitative analyses. *PLoS One* 2012; 7(6): e38268.

29. Tausczik YR and Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol* 2010; 29(1): 24–54.

30. Pennebaker JW, Boyd RL, Jordan K, et al. *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin, 2015.

31. Alz Foundation. *Famous people with dementia*. New York, NY: Alz Foundation, 2022. Available from: https://brevardalz.org/alzheimers-info/famous-people-with-dementia/

32. Kempler D and Goral M. Language and dementia: neuropsychological aspects. *Annu Rev Appl Linguist* 2008; 28: 73–90.

33. Gaugler JE, Jutkowitz E, Shippee TP, et al. Consistency of dementia caregiver intervention classification: an evidence-based synthesis. *Int Psychogeriatr* 2017; 29(1): 19–30.

34. Albinsson L and Strang P. Existential concerns of families of late-stage dementia patients: questions of freedom, choices, isolation, death, and meaning. *J Palliat Med* 2003; 6(2): 225–235.

35. Wilson S, Thorne A, Stephens M, et al. English vocabulary size in adults and the link with educational attainment. *Language in Focus* 2016; 2(2): 44–69.

36. Graham S. The sciences of reading and writing must become more fully integrated. *Read Res Q* 2020; 55: S35–S44.

37. Graham S. Changing how writing is taught. *Rev Res Educ* 2019; 43(1): 277–303.