Research Paper

Characterizing and detecting delirium with clinical and computational measures of speech and language disturbance

Sunny X. Tang, MD; Yan Cong, PhD; Gwenyth Mercep, BS; Mutahira Bhatti, BS; Grace Serpe BA; Valeria Gromova, BS; Sarah Berretta, BA; Majnu John, PhD; Mark Y. Liberman, PhD; Liron Sinvani, MD

Background: Delirium is a critically underdiagnosed syndrome of altered mental status affecting more than 50% of older adults admitted to hospital. Few studies have incorporated speech and language disturbance in delirium detection. We sought to describe speech and language disturbances in delirium, and provide a proof of concept for detecting delirium using computational speech and language features. **Methods:** Participants underwent delirium assessment and completed language tasks. Speech and language disturbances were rated using standardized clinical scales. Recordings and transcripts were processed using an automated pipeline to extract acoustic and textual features. We used binomial, elastic net, machine learning models to predict delirium status. **Results:** We included 33 older adults admitted to hospital, of whom 10 met criteria for delirium. The group with delirium scored higher on total language disturbances and incoherence, and lower on category fluency. Both groups scored lower on category fluency than the normative population. Cognitive dysfunction as a continuous measure was correlated with higher total language disturbance, incoherence, loss of goal and lower category fluency. Including computational language features in the model predicting delirium status increased accuracy to 78%. **Limitations:** This was a proof-of-concept study with limited sample size, without a set-aside cross-validation sample. Subsequent studies are needed before establishing a generalizable model for detecting delirium. **Conclusion:** Language impairments were elevated among patients with delirium and may also be used to identify subthreshold cognitive disturbances. Computational speech and language features are promising as accurate, noninvasive and efficient biomarkers of delirium.

Introduction

Delirium is a syndrome characterized by an acute change in mental status, marked by inattention and global cognitive dysfunction.[1,2] As many as 50% of older adults (aged ≥ 75 yr) experience delirium in the hospital setting, which is associated with increased morbidity, mortality and resource utilization.[1,3–6] Delirium is a leading cause of iatrogenic complications — such as falls, incontinence and pressure ulcers — that lead to longer hospital stays and higher rates of hospital discharge to a skilled nursing facility and institutionalization.[1,5,7,8] In addition, delirium is associated with hospital mortality of 25%–33% and annual health care expenditures of up to $152 billion in the United States.[9,10] One of the most common and deleterious consequences of delirium is the development of long-term cognitive decline and new dementia.[1,2,11] Delirium is associated with a 12-fold increased risk for new-onset dementia.[12,13]

Currently, 75% of delirium in the hospital setting goes undiagnosed.[14–18] Barriers to diagnosis include varied clinical presentation,[1,2,19] underuse and inaccurate use of screening tools[20,21] and difficulty distinguishing delirium from preexisting cognitive impairment.[14,19,22] Failure to diagnose delirium prevents implementation of effective mitigation strategies (e.g., identification or treatment of underlying cause, safety precautions, nonpharmacologic management), which leads to poor clinical outcomes.[1,2,11,15,16] Although many molecular and imaging measures have been investigated as potential biomarkers for diagnosis of delirium, these approaches are invasive and inefficient, requiring blood draws and specialized equipment and expertise to interpret.[12,23–25] Therefore, a noninvasive, cost-effective and timely biomarker is urgently needed to improve detection of delirium.

Assessing speech and language disturbances could be informative for improving detection of delirium, but has not been adequately explored. Disturbance in language is among

Correspondence to: L. Sinvani, North Shore University Hospital, 300 Community Drive, Manhasset, NY, 11030; ldanay@northwell.edu

Submitted Feb. 15, 2023; Revised Apr. 13, 2023; Accepted Apr. 24, 2023

Cite as: *J Psychiatry Neurosci* 2023 July 4;48(4). doi: 10.1503/jpn.230026

the diagnostic criteria of delirium, as listed in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, but there is no guidance on the specific language domains that are affected or how these language deficits should be measured.[26] In addition, although a few delirium screening tools (e.g., Delirium Rating Scale–Revised–98) mention the evaluation of language, there are no standardized criteria for identifying or quantifying these impairments.[27] Disturbance in language is not required for the diagnosis of delirium, but it has potential to improve delirium detection. Only a handful of studies have systematically examined speech in this context, likely because of the difficulty of achieving standardized and reliable assessments in the hospital setting.[26,28–30] An assessment of 100 adult patients with delirium found that language impairment, as measured by the Delirium Rating Scale–Revised–98, was identified in more than half of participants.[28] In a 1994 study that compared 13 patients with an acute confusion state and 11 patients with probable Alzheimer disease, unrelated misnamings and perseverations were found more frequently among those with acute confusion state.[29] Lastly, in a case–control study of 45 patients (15 with delirium, 15 with dementia without delirium and 15 with no cognitive impairment), those with delirium scored lower on conversational and written conversational assessment and performed worse on written comprehension tests than patients with no cognitive impairment, which was likely related to visual misperceptions.[30]

In other neuropsychiatric disease areas, artificial intelligence and machine learning have been leveraged to identify clinically relevant speech biomarkers. Automated speech analysis uses computerized algorithms to extract information from the textual and acoustic elements of speech (i.e., what is said and how it is delivered). Acoustic signals pertaining to how speech sounds (e.g., pitch, voicing duration v. pausing) can be extracted directly from the audio recordings, typically sampled at 100 frames per second.[31] Speech can also be transcribed either automatically or using human annotators, then analyzed using natural language processing (NLP) methods.[32,33] For example, automated computer algorithms can label, then count, the use of different parts of speech, or identify and quantify the use of emotional words. The organization and flow of speech can also be quantified using graph analysis or by measuring the distances among word embeddings, which are multidimensional numerical representations of the meaning of each word.[34]

For example, machine learning models have used computational speech features to classify participants with Alzheimer disease versus healthy volunteers and also to predict onset of Alzheimer disease with accuracy and area under the curve greater than 0.8.[33,35–37] In schizophrenia, where speech disturbances have been described in detail and captured with a variety of standardized clinical scales,[38–40] NLP features were shown to be more sensitive to subtle variations in speech than gold-standard clinical measures.[41] However, objective quantification of speech and language disturbances in delirium has not been thoroughly explored.

We sought to systematically characterize speech and language disturbance among older adults in hospital with and without delirium, to evaluate cognitive dysfunction in this population as a continuous measure in relation to speech and language disturbance and to explore whether computational speech features could be used to detect delirium.

Methods

Participants

We recruited older adults (aged ≥ 75 yr) from 3 medicine (non-telemetry) units at a quaternary academic hospital in the New York metropolitan area. This study was focused on delirium, so we excluded patients with documented history of dementia in the electronic health record (EHR), as well as other causes of cognitive dysfunction. We excluded patients with an acute intracranial event (e.g., cerebrovascular accident, transient ischemic attack, subdural hemorrhage, subarachnoid hemorrhage); those who were critically ill or actively dying; those admitted to the surgical service; those unable to participate because of lethargy; patients with a history of dysarthria, aphasia or traumatic brain injury; those with acute serious psychiatric illness (e.g., schizophrenia, bipolar disorder, treatment-resistant depression); and those with moderate or severe intellectual disability. Additional inclusion criteria were English fluency and ability to provide consent, as evaluated by passing the consent quiz.

Research assistants screened the EHR for eligible patients. The research assistant approached the nurse to verify inclusion and exclusion criteria for eligible patients. Once confirmed, the research assistant approached the patient and conducted a quiz to assess capacity for consent by confirming that they understood the informed consent and research procedures. For patients with capacity to consent for research and who agreed to participate, the research assistant obtained written consent and initiated the assessment.

Assessment measures

Demographics and clinical characteristics

We obtained demographics and clinical history primarily from the EHR, including age, sex, race, ethnicity, marital status, residence before admission, admitting diagnosis and Modified Early Warning Score (MEWS). This score — comprising systolic blood pressure, heart rate, respiratory rate, temperature and responsiveness — determines the degree of the patient's acute illness.[42] Patient interview–based clinical measures included baseline functional status (activities of daily living),[43] education level (personal and parental years of completed education, starting from first grade), presence of secondary languages; current or previous occupation.

Delirium assessment

Trained research assistants completed delirium assessments, and a single expert in delirium (L.S.) completed ratings to avoid issues with interrater reliability. These ratings were completed while blinded to the speech and language ratings and results. Delirium assessment measures included orientation (person, place and time), 3-item recall, the Confusion Assessment Method (CAM) long form[44] and the Richmond Agitation and Sedation Scale.[45] In addition to direct patient interview

Table 1: Computational speech and language features*

Modality	Measures	Final feature set
Speech quantity or dysfluencies	Partial words, repeated words, repeated segments, NLTK stop works, neologisms, speech errors (any), restarts, filled pauses, total words	Total speech errors (growing up narrative and paragraph reading); restarts (spending time narrative), filled pauses (family narrative), filled pauses (spending time narrative), total words (picture description), total words (category fluency)
Parts of speech	Adjectives, adpositions, adverbs, auxiliaries, coordinating conjunctions, determiners, interjections, nouns, numbers, particles, pronouns, proper nouns, punctuation, subordinating conjunctions, symbols, verbs	Adjectives (picture description), adverbs (open-ended tasks), determiners (family narrative and spending time narrative), interjections (picture description), numbers (family narrative), particles (open-ended tasks), symbols (open-ended tasks)
Tempo or pauses	Total speech duration, speech duration per turn (mean and SD), speech duration per segment (mean and SD), average turn latency, speaking rate (mean, SD, min, max), pause length (mean, SD, min, max), short pause length (mean, SD), pause length variability (mean, SD)	Speech duration (fluency tasks), average turn latency (paragraph reading), SD of speech duration per segment (about yourself narrative), variability in short pause length (growing up narrative), SD in pause variability (paragraph reading)
Voice quality or prosody	Mean pitch f_0 (mean, SD), pitch variability (mean, SD), mean jitter (mean, SD), jitter variability (mean, SD), mean shimmer (mean, SD), shimmer variability (mean, SD)	Mean of jitter variability (paragraph reading), SD of mean jitter (phonemic fluency)
Semantic coherence	Mean sentence embedding cosine distances using GloVe, LSA and Word2Vec techniques (mean, minimum, maximum); TF-IDF sentence embedding cosine distances using GloVe, LSA and Word2Vec techniques (mean)	Mean cosine distances among TF-IDF sentence embeddings from Word2Vec (all tasks), maximum cosine distances from mean sentence embeddings from GloVe (picture description)
Lexical characteristics	Age of acquisition, prevalence, semantic diversity, valence, arousal, dominance	Semantic diversity (fluency tasks), prevalence (spending time narrative)

f_0 = fundamental frequency; GloVe = Global Vectors for Word Representation; LSA = latent semantic analysis; NLTK = natural language tool kit; SD = standard deviation; TF-IDF = term frequency–inverse document frequency.
*Each measure was computed separately for the 4 narrative tasks, picture description, paragraph reading, and 2 fluency tasks, and as aggregate measures for the open-ended tasks, fluency tasks and across all tasks. Because little variability was expected, we omitted parts-of-speech and semantic coherence measures for the fluency and paragraph reading tasks, as well as lexical characteristics for the paragraph reading task. The final feature set (26 measures) is listed for each modality after excluding features without trend-level correlation with the total Confusion Assessment Method Severity Score ($p < 0.10$) and highly redundant measures.

and observation, the research assistant asked the bedside nurse to think of their assessment during their shift and answer a series of questions. Is there evidence of an acute change in mental status from the patient's baseline (acute onset)? Does the patient display abnormal behaviour that fluctuates during the day; that is, did it tend to come and go or increase or decrease in severity (fluctuating course)? Does the patient have any evidence of perceptual disturbances, such as hallucinations, illusions or misinterpretations (e.g., thinking something was moving when it was not) (perceptual disturbances)? Does the patient have evidence of disturbance of the sleep–wake cycle, such as excessive daytime sleepiness with insomnia at night (altered sleep–wake cycle)? The delirium expert based scores on the final CAM long form on the recording of the delirium assessment, the research assistant's documentation of the CAM long form and nursing responses. Each patient was classified as positive or negative for delirium based on the presence of feature 1 (acute onset or fluctuating course), feature 2 (inattention) and either feature 3 (disorganized thinking) or feature 4 (altered level of consciousness). In addition, we calculated a total severity score based on the CAM long form's severity scale (CAM-S).[46] For the purpose of this study, cognitive dysfunction was represented by the CAM-S score as a continuous dimensional construct.

Language assessments

We collected audio recordings for open-ended prompts (4 narratives and 1 picture description task), a paragraph reading and fluency tasks. Human annotators completed verbatim transcriptions using EUDICO Linguistic Annotator (ELAN),[47,48] which included determination of intuitive utterance boundaries based on pauses and syntax, as well as marking dysfluencies such as incomplete words, restarts, repetitions. A single expert rater (S.X.T.) gave each participant subjective ratings for speech and language disturbances to avoid issues with interrater reliability. These ratings were completed based on recorded speech while blinded to the delirium status of the participant. Ratings were drawn from the Scale for the Assessment of Thought Language and Communication (TLC),[38] as well as 2 complementary items from the Scale for the Assessment of Negative Symptoms (Table 1).[49] These scales were developed for assessment of speech disturbance related to psychiatric disorders and cover a broad range of symptoms, from underproductive to disorganized or superfluous speech. We chose these measures because they have good psychometric properties and for the breadth of speech- and language-related symptoms captured. To our knowledge, there is no validated scale for assessing different dimensions of language disturbance in delirium. Further details on the language assessments are included in Appendix 1, available at https://www.jpn.ca/lookup/doi/10.1503/jpn.230026/tab-related-content.

Data analysis

Computational speech and language features

We processed recordings and transcripts separately for each task using an automated pipeline to extract acoustic (prosody and voice quality, speaking tempo, pauses) and textual features (semantic coherence, dysfluencies and speech errors,

lexical characteristics, parts of speech, speech quantity). Details on the automated pipeline are described in Appendix 1. We initially generated 73 features for each task, as well as aggregate measures for the open-ended responses, fluency tasks and all tasks, for a total of 684 speech features (Table 1). We excluded acoustic features on a task-by-task basis when recording quality was poor (primarily because of high background noise). For the prediction modelling, we imputed missing data with permuted mean matching using the mice package in R;[50] for group comparisons, we omitted missing data. To lower the likelihood of overfitting and to reduce the feature space, we first selected only features that showed at least a trend-level correlation with the total CAM-S score ($p < 0.10$), then visually inspected correlation plots to remove highly redundant features. The final feature set included 26 measures.

Statistical analysis and machine learning

We compared groups using analysis of variance for continuous variables (age, MEWS and ratings for the CAM long form and TLC) and the χ^2 test for categorical variables (sex, race, ethnicity). We standardized category fluency totals with respect to normative data for older adults based on age group and education level.[51] We used 1-sided t tests to compare fluency totals from this sample to population norms. We calculated Spearman coefficients to assess the correlations for clinical language measures with delirium and illness severity measures. To account for age, sex and illness severity in the correlations between clinical speech and language ratings with cognitive dysfunction, we used multiple linear regressions with 1 covariate at a time. Hypothesis testing was based on 2-tailed tests with an α of 0.05.

We used binomial, elastic net regression models to predict delirium status (positive v. negative for delirium) with 10-fold internal cross-validation (i.e., training the model on 90% of the data and predicting results in the remaining 10%, 10 times). We determined average accuracy and κ statistics for the 10-fold cross-validations. A final model was constructed for the overall data set, and was used to report feature loadings and the confusion matrix for the full sample. We did not create a set-aside test set because of the limited sample size and because the aim of this study was to produce a proof of concept that computational language features are promising biomarkers for delirium, not to create a generalizable, predictive classifier. There are no standard power analyses for this type of analysis, but given the sample size, we caution against overinterpreting the specific predictors or model outcomes. We chose the elastic net regression method because it is suitable for small samples with relatively larger numbers of predictors, and it is able to accommodate collinear predictors and generates a reduced set of predictors with numerical loadings. We used the caret package of R to conduct model training and testing.[52]

Ethics approval

All eligible participants provided informed consent, and study procedures were approved by the institutional review board at the Feinstein Institutes for Medical Research (21-0568-NSUH).

Results

We included 33 participants, of whom 10 met criteria for delirium and 23 did not (Table 2). Because all participants had to be capable of providing consent for this study, none of them had marked cognitive dysfunction, and none had been identified by their clinical treatment team as having delirium. Participants were admitted for a range of medical concerns, and 8 patients were prescribed either opiates or benzodiazepines during their admission. Appendix 1, Table 1, provides details on admitting diagnoses and use of medications in greater detail, but given the sample size, a more detailed analysis of underlying diagnoses and medications was not possible.

As expected, the group that was positive for delirium had a significantly higher total CAM-S score ($p < 0.001$, Cohen $d = 4.32$). There were no significant differences in age, sex, race, ethnicity or MEWS score between the groups. Figure 1A illustrates group differences for total TLC score and 8 individual TLC items that had an overall prevalence of greater than 20% in the sample. The delirium group scored significantly higher on total TLC score ($p = 0.05$, Cohen $d = 0.81$) and the TLC item for incoherence ($p = 0.001$, Cohen $d = 1.41$). In addition, there was a trend for higher loss of goal, defined as wandering off topic and failing to follow a chain of thought to its conclusion ($p = 0.05$, Cohen $d = 0.80$), and higher global language disturbance ($p = 0.07$, Cohen $d = 0.73$). Participants with delirium also scored significantly lower on the category fluency task than those without delirium ($p = 0.02$, Cohen $d = –0.97$), and both groups scored lower than the normative population ($p < 0.001$, Cohen $d = –1.22$ for those without delirium; $p < 0.001$, Cohen $d = –2.20$ for those with delirium) (Figure 1B).

Clinical ratings for speech and language disturbance and associations with cognitive dysfunction

Cognitive dysfunction as a continuous measure (CAM-S score) was significantly correlated with several clinical measures of speech and language disturbance. As a dimensional measure of cognitive dysfunction and delirium severity, higher total CAM-S score was correlated with higher total TLC score ($r = 0.41$, $p = 0.02$), higher incoherence ($r = 0.58$, $p < 0.001$), higher loss of goal ($r = 0.36$, $p = 0.04$) and lower category fluency ($r = –0.41$, $p = 0.02$) (Figure 1C). Each of these relationships remain significant even when covarying for age, sex, education and MEWS score, with the exception of loss of goal covaried with education ($p = 0.052$). Appendix 1, Figure 1, further details the dimensional relationship between cognitive dysfunction severity (CAM-S score) and clinical measures of language disturbance. These correlations appear to be specific to cognitive dysfunction and were not present for general illness severity as there were no significant correlations between speech and language measures and the MEWS score.

Table 2: Participant characteristics and clinical ratings for speech disturbance

Variable	No. (%) of patients*			*p* value	Cohen *d*
	Without delirium *n* = 23	With delirium *n* = 10	Total *n* = 33		
Age, yr, mean ± SD	82.7 ± 5.9	84.1 ± 6.3	83.1 ± 5.9	0.53	0.25
Sex, female	15 (65)	4 (40)	19 (58)	0.18	
Race				0.06	
Black	4 (17)	0 (0)	4 (12)		
Other	1 (4)	3 (30)	4 (12)		
White	18 (78)	7 (70)	25 (76)		
Hispanic ethnicity	0 (0)	1 (10)	1 (3)	0.13	
MEWS, mean ± SD	4.3 ± 0.9	4.0 ± 0.7	4.2 ± 0.9	0.42	–0.32
CAM severity score, mean ± SD	0.9 ± 1.1	5.3 ± 1.0	2.2 ± 2.3	< 0.001	4.32
Speech ratings, mean ± SD					
SANS 6 — Lack of vocal inflection	0.87 ± 1.06	0.80 ± 1.23	0.85 ± 1.09	0.87	–0.06
SANS 11 — Increased latency	0.22 ± 0.60	0.20 ± 0.63	0.21 ± 0.60	0.94	–0.03
TLC 1 — Poverty of speech	0.30 ± 0.70	0.70 ± 0.82	0.42 ± 0.75	0.17	0.55
TLC 2 — Poverty of content of speech	0.43 ± 0.79	0.80 ± 0.79	0.55 ± 0.79	0.23	0.48
TLC 3 — Pressure of speech	0.17 ± 0.39	0.40 ± 0.84	0.24 ± 0.56	0.29	0.42
TLC 4 — Distractible speech	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	–	–
TLC 5 — Tangentiality	0.57 ± 0.99	1.00 ± 1.15	0.70 ± 1.05	0.28	0.43
TLC 6 — Derailment	0.09 ± 0.42	0.40 ± 0.70	0.18 ± 0.53	0.12	0.63
TLC 7 — Incoherence	0.13 ± 0.34	1.00 ± 1.05	0.39 ± 0.75	0.001	1.41
TLC 8 — Illogicality	0.09 ± 0.29	0.10 ± 0.32	0.09 ± 0.29	0.91	0.05
TLC 9 — Clanging	0.09 ± 0.29	0.00 ± 0.00	0.06 ± 0.24	0.35	–0.37
TLC 10 — Neologisms	0.13 ± 0.34	0.40 ± 0.70	0.21 ± 0.48	0.15	0.58
TLC 11 — Word approximations	0.09 ± 0.29	0.00 ± 0.00	0.06 ± 0.24	0.35	–0.37
TLC 12 — Circumstantiality	0.48 ± 0.85	0.80 ± 1.03	0.58 ± 0.90	0.36	0.37
TLC 13 — Loss of goal	0.13 ± 0.34	0.50 ± 0.71	0.24 ± 0.50	0.05	0.80
TLC 14 — Perseveration	0.57 ± 0.79	0.60 ± 0.84	0.58 ± 0.79	0.91	0.04
TLC 15 — Echolalia	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	–	–
TLC 16 — Blocking	0.22 ± 0.60	0.30 ± 0.67	0.24 ± 0.61	0.73	0.14
TLC 17 — Stilted speech	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	–	–
TLC 18 — Self-reference	0.35 ± 0.71	0.60 ± 0.97	0.42 ± 0.79	0.41	0.33
TLC global score	0.65 ± 0.78	1.20 ± 0.79	0.82 ± 0.81	0.07	0.73
TLC total score	5.91 ± 6.93	12.40 ± 10.73	7.88 ± 8.64	0.05	0.81
Fluency tasks, mean ± SD					
Category total (animals)	11.43 ± 5.41	6.90 ± 2.96	10.06 ± 5.20	0.02	–0.97
Phonemic total (f-letter)	6.96 ± 4.34	5.30 ± 4.57	6.45 ± 4.41	0.33	–0.80

CAM = Confusion Assessment Method Long Form; MEWS = Modified Early Warning Score; SANS = Scale for the Assessment of Negative Symptoms; SD = standard deviation; TLC = Scale for the Assessment of Thought Language and Communication.
*Unless indicated otherwise.

Using computational speech and language features to detect delirium diagnosis

Delirium status was predicted with demographics alone, demographics and clinical speech ratings, demographics and computational speech features, or all of these. Model parameters and performance are detailed in Table 3. Overall, the model with demographics and computational speech features performed best, classifying delirium status with an average accuracy of 78% (10-fold cross-validation, κ 0.4, area under the curve 0.90). In the final model, all 23 participants without delirium were correctly identified, as were 8 of the 10 participants with delirium. With regard to feature loadings, the presence of delirium was most highly predicted by speech errors during paragraph reading, transcribed symbols (including punctuation, restarts and incomplete words) in the open-ended narratives, use of determiners in the family narrative task and semantic diversity (ambiguity) of words given during the fluency tasks. Absence of delirium was most highly predicted by filled pauses during the family narrative task (e.g., "um,"

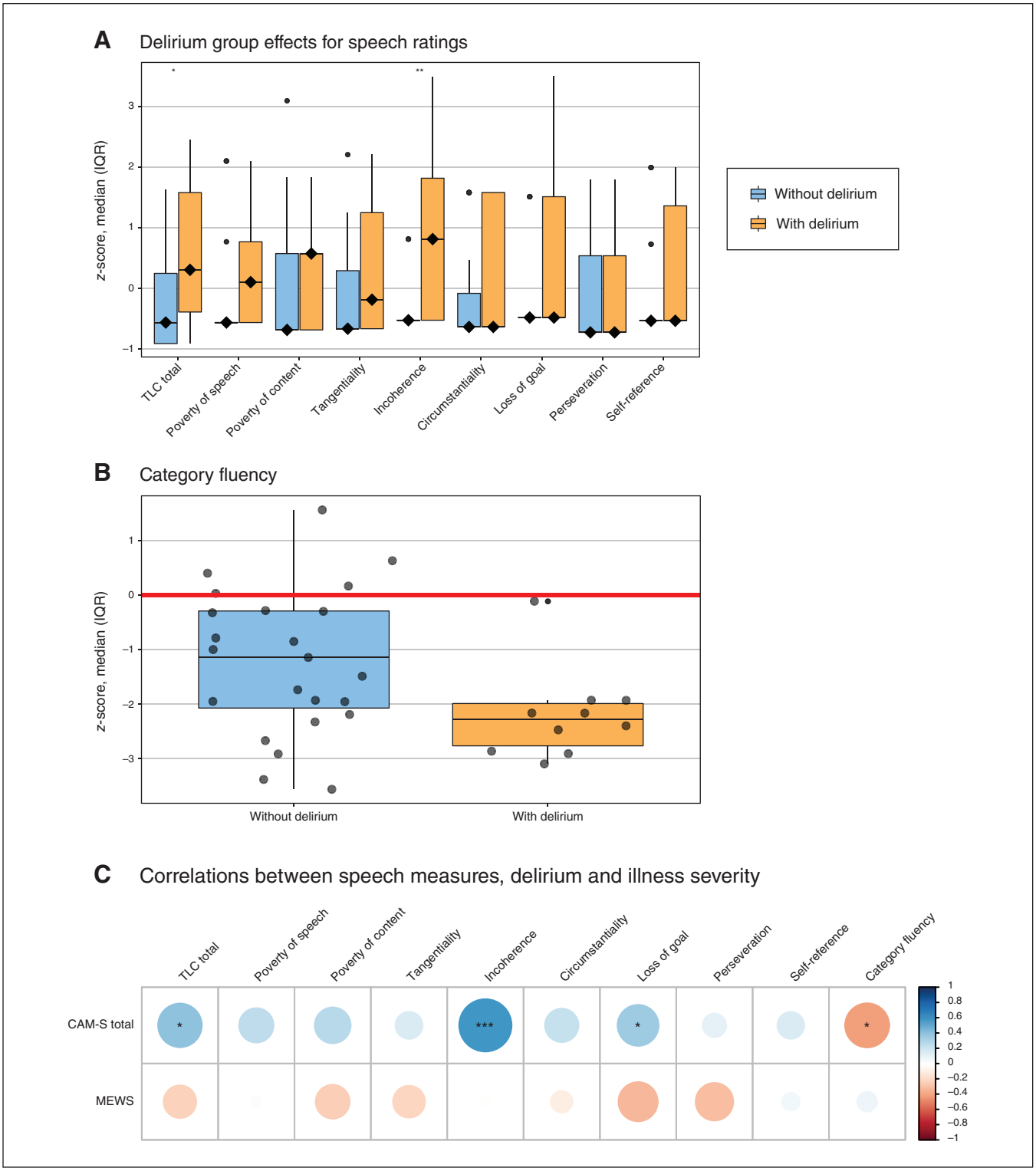

Figure 1: Relationships between delirium and clinical speech measures. (A) Group effects for delirium status and clinical ratings of speech disturbance on the Scale for the Assessment of Thought Language and Communication (TLC), including total score and individual items that had a total prevalence greater than 20% in this sample; TLC scores are normalized within the current sample. (B) Group effects for category fluency scores (number of animals listed in 1 min). Fluency scores were normalized with respect to population norms for the participants' age group and educational level. (C) Spearman correlations between speech measures, delirium severity (total Confusion Assessment Method [CAM] Severity Score [CAM-S]) and illness severity (Modified Early Warning Score [MEWS]). IQR = interquartile range. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 3: Prediction of delirium status*

Predictor set	Accuracy, %	κ	Positive predictors		Negative predictors		Confusion matrix			
			Variable	Coefficient	Variable	Coefficient	TP	FP	TN	FN
Demographic variables only	69	0	Male sex	0.39	MEWS	–0.23	1	0	23	9
			Age	0.27	Education level	–0.20				
					White race	–0.12				
Demographic variables and clinical speech ratings	78	0.3	Male sex	0.05			3	0	23	7
			Incoherence	0.73						
Demographic variables and computational speech features	78	0.4	Total speech errors in paragraph reading	0.72	Filled pauses in family narrative	–0.27	8	0	23	2
			Symbols in open-ended narratives	0.46	Adverbs in open-ended tasks	–0.27				
			Determiners in family narrative	0.39	Numbers in family narrative	–0.23				
			Semantic diversity in fluency tasks	0.38	Adjectives in picture description	–0.19				
			Average turn latency in paragraph reading	0.19	Variance in mean vocal jitter in fluency tasks	–0.12				
			Lexical prevalence in spending time narrative	0.16	Variance in speech duration in about yourself narrative	–0.09				
			Variation in short pause duration	0.09	Determiners in spending time narrative	–0.03				
			Male sex	0.09	Maximum semantic similarity of glove embedding in picture description	–0.03				
			Speech duration in fluency tasks	0.02						
			Particles in open-ended narratives	0.005						
			Interjections in picture description	0.004						
			Age	0.003						
			Restarts in spending time narrative	0.002						
Demographic variables, clinical language ratings and computational speech features	75	0.25	Total speech errors in paragraph reading	0.27	Adverbs in open-ended narratives	–0.15	7	0	23	3
			Incoherence	0.21	Filled pauses in family narrative	–0.12				
			Semantic diversity in fluency tasks	0.20	Variance in mean vocal jitter in fluency tasks	–0.10				
			Determiners in family narrative	0.19	TF-IDF semantic similarity with Word2Vec embeddings for all tasks	–0.08				
			Symbols in open-ended narratives	0.19	Variance in speech duration in about yourself narrative	–0.08				
			Loss of goal	0.14	Numbers in family narrative	–0.07				
			Lexical prevalence in spending time narrative	0.08	Word approximations	–0.06				
			Average turn latency in paragraph reading	0.08	Determiners in spending time narrative	–0.06				
			Variation in short pause duration	0.08	Maximum semantic similarity of glove embedding in picture description	–0.06				
			Male sex	0.07	Adjectives in picture description	–0.05				
			Derailment	0.04	Clanging	–0.04				
			Global TLC score	0.04	Filled pauses in spending time narrative	–0.02				
			Particles in open-ended narratives	0.02	Total words in picture description	–0.02				
			Variance in SD of pause lengths in paragraph reading	0.02	MEWS	–0.01				
			Speech duration in fluency tasks	0.002						
			Age	0.001						

FN = false negative; FP = false positive; MEWS = Modified Early Warning Score; SD = standard deviation; TF-IDF = term frequency–inverse document frequency; TLC = Scale for the Assessment of Thought Language and Communication; TN = true negative; TP = true positive.
*Binomial, elastic net regression models were trained with 10-fold cross-validation. Average accuracy and κ statistics are shown for the 10-fold cross-validations (predicting 10% of the data using a model trained on the remaining 90%, repeated 10 times). Confusion matrix refers to classification results using the final model on the whole sample.

"uh"), use of adverbs in open-ended narratives, use of adjectives during the picture description, and variance in jitter (fluctuations in voice amplitude) during fluency tasks.

Discussion

There is an urgent need for a noninvasive and efficient diagnostic biomarker to detect delirium, a common, deleterious and underdiagnosed syndrome that affects more 50% of older adults admitted so hospital. Our study aimed to provide a proof of concept for using computational speech and language features as diagnostic biomarkers by systematically characterizing the degree and nature of speech and language disturbances in delirium from a clinical perspective, evaluating cognitive dysfunction as a continuous measure and, finally, predicting delirium using computational speech and language features. We found that delirium status and a dimensional measure of cognitive dysfunction were related to several domains of speech and language disturbances, assessed through standardized clinical ratings. In addition, we found that computational speech and language features contribute to the detection of delirium diagnosis among older adults in the hospital setting.

As reflected by comprehensive standardized clinical ratings, the group with delirium was significantly more incoherent, had higher total symptoms of language disturbances, and scored lower on a category fluency task than the group without delirium. There was also a trend toward greater tendency to wander off-topic. These results were generally consistent with those of previous studies that detected several areas of language disturbance in delirium.[28–30] Green and colleagues[30] compared 3 cohorts of 15 patients with delirium, dementia or no cognitive impairment — one of the only systematic studies of domains of language disturbance among patients with delirium — and found that participants with delirium were more likely to speak on irrelevant material, which is similar to our finding of a trend toward increased loss of goal. Other items related to going off topic were also rated higher in the group with delirium, but did not reach statistical significance (e.g., tangentiality, circumstantiality, derailment). Green and colleagues[30] also found that participants with delirium spoke fewer words in a picture description task, which may be related to our finding of significantly lower scores among those with delirium on the category fluency task. In our sample, poverty of speech was also rated higher on average in the group with delirium. Although we also found higher incoherence in this group, Green and colleagues[30] did not find significant differences in related measures of grammatical and semantic irregularities. The variations that exist in these findings are very likely related to the limited sample sizes of both studies. Altogether, patients with delirium reliably show clinically detectable language disturbances. These disturbances are likely to present in multiple domains, including going off topic or deviating from the logical focus of conversation.

As a continuous dimensional measure of cognitive dysfunction, we also found significant correlations between the CAM-S score and total language disturbances, incoherence, loss of goal, and lower scores on the category fluency task. This effect appears to be specific to the CAM-S score and not overall illness severity because no such correlations existed when comparing the MEWS with clinical ratings of speech and language disturbance. Moreover, we found that, not only did the group with delirium score lower on category fluency than the group without delirium, but both groups scored lower than published norms for age- and education-matched healthy controls. These findings are consistent with previous literature describing subclinical cognitive dysfunction that does not meet criteria for delirium but nevertheless results in clinically detectable changes in cognition, speech and language.[53] To our knowledge, there has not been any description of speech and language disturbance in subclinical cognitive dysfunction among older adults admitted to hospital. The functional and clinical outcomes for such changes should be further explored. In particular, these findings suggest that speech and language features could detect subthreshold cognitive dysfunction, which may have long-term consequences for post-hospital functioning.

Automated speech analysis is a promising direction of investigation because these methods have the potential to generate objective and sensitive speech biomarkers that can be collected in a noninvasive, cost-effective and timely manner. A fully automated process is achievable, whereby patients can speak into a simple recorder or smartphone and computerized algorithms can analyze acoustic and textual characteristics to produce objective, digital biomarkers at the point of care. We found that speech samples could be feasibly collected at the bedside for older adults in hospital. The recordings were then processed using an automated pipeline that measured the sound and content of patient speech. When incorporated into machine learning algorithms, these computational speech and language features (combined with demographic variables) improved the accuracy of delirium detection, compared with demographic variables and general illness severity measures alone or demographics and clinical language ratings. Values of the κ statistic were fair and were modestly higher when the computational speech features were added to the clinical ratings (0.4 v. 0.3).[54] The current model should only be considered a proof of concept and needs optimization and validation before clinical implementation. The generalizability of our findings needs to be confirmed with a larger sample and across multiple sites. However, our results do suggest that computational speech and language features have the potential to serve as accurate and efficient biomarkers for detecting delirium.

The use of computational speech analysis in the diagnosis of delirium has practical clinical applications. Of interest, none of the patients had delirium documented in their EHR, which means that the machine learning algorithms were able to detect cases of delirium that were not evident to clinical treatment teams. In the real-world, delirium is missed in more than 75% of cases; even when standardized screening tools are used, their sensitivity at the bedside is less than 50%.[55] Failure to diagnose delirium prevents implementation

of effective mitigation strategies (i.e., identification or treatment of underlying cause, safety precautions and non-pharmacologic management), which leads to poor clinical outcomes. Therefore, computational speech and language biomarkers could be used for timely detection of delirium and the implementation of management strategies to improve patient outcomes.

Limitations

This was a proof-of-concept study, leaving many additional questions to be answered. At this stage, findings regarding individual features cannot be generalized with confidence. Given that language depends on education, cultural background and other sociological factors, and may also vary with underlying disease, articulatory functions, time points of examination and sedating medications, future studies in more specific and larger samples are needed, with more detailed clinical characterization to determine the most sensitive and specific parameters to be applied as biomarkers in delirium. Therefore, the next step is to determine whether the results can be replicated in a larger sample, across sites and time points, and whether there are important variations for people of different sexes, genders, races and socioeconomic backgrounds, and from different geographic regions. From a machine learning perspective, separate set-aside training and testing samples need to be incorporated. Future studies should also benchmark the speech and language disturbances against more formal, performance-based cognitive testing. Additional computation features related to idea density and speech organization can also be introduced to improve the accuracy of delirium detection. Although we excluded participants with a documented history of dementia before hospital admission, some may have been experiencing cognitive decline before admission. As delirium presents at a higher rate in dementia, and is associated with accelerated cognitive decline among patients with pre-existing cognitive impairment, future studies using speech and language disturbance must assess pre-admission cognitive status and the interaction between dementia and delirium. Lastly, as this was a pilot study with limited resources, only patients who were able to provide consent were included, which likely led to the inclusion of predominantly mild delirium (as seen by the CAM-S scores). Although this limits the generalizability of our findings to all patients with delirium, patients with more severe delirium presentations are likely to have even more pronounced speech and language disturbance, not less. Therefore, we expect that our findings likely underestimate the prevalence and severity of speech and language disturbances in delirium. Future studies need to include participants with a wide range of delirium severity. The use of single raters for the TLC and CAM-S allowed us to circumvent issues of interrater reliability, and each of the raters are experts in their field; however, it is unclear how well the ratings would generalize across sites. Therefore, future studies should also include multiple sites and a more formal process to assess the reliability of ratings.

Conclusion

We applied standardized clinical rating scales for speech and language disturbance and found significant language disturbances among participants with delirium. These findings were in line with previous studies, but we applied a more systematic and detailed approach to evaluating individual speech- and language-related symptoms. Using automated speech analysis to generate computational speech and language features, we were able to detect the presence of delirium with greater accuracy than using demographics and general illness severity alone or in combination with clinical language disturbance ratings. Although molecular and imaging measures have been investigated as potential biomarkers for delirium diagnosis, these approaches are invasive and inefficient, requiring blood draws and specialized equipment and expertise to interpret. We find that computational speech and language biomarkers may be promising as accurate, non-invasive and efficient biomarkers for detecting delirium.

Affiliations: From the Institute of Behavioral Science, Feinstein Institutes for Medical Research, Northwell Health, Glen Oaks, New York (Tang, Cong, Serpe, Berretta, John); the Institute of Health System Science, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York (Mercep, Bhatti, Gromova, Sinvani); Department of Linguistics, University of Pennsylvania, Philadelphia, PA (Liberman).

Competing interests: Sunny Tang is a consultant for North Shore Therapeutics and Winterlight Labs, has received funding from Winterlight Labs and holds equity in North Shore Therapeutics and Psyrin. She reports an honorarium from Ohio State University and travel support from the Schizophrenia International Research Society. She sits on the program committee of the DISCOURSE in Psychosis Consortium and on the advisory boards for Psyrin and North Shore Therapeutics. Sarah Berretta reports travel support from the DISCOURSE in Psychosis Consortium. No other competing interests declared.

Acknowledgement: The authors thank Sunghye Cho from the Linguistic Data Consortium for her valuable input.

Funding: Sunny Tang received funding from the Young Investigator Grant from the Brain and Behavior Research Foundation (K23 MH130750-01) and the National Institutes of Health (K23 MH130750-01). Suny Tang and Liron Sinvani both received funding from the Barbara Hrbek Zucker Award for Emerging Scientists from the Feinstein Institutes for Medical Research.

References

1. Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383:911-22.
2. Wilson JE, Mart MF, Cunningham C, et al. Delirium. *Nat Rev Dis Primers* 2020;6:90.
3. Schor JD. Risk factors for delirium in hospitalized elderly. *JAMA* 1992;267:827-31.
4. Ahmed S, Leurent B, Sampson EL. Risk factors for incident delirium among older people in acute hospital medical units: a systematic review and meta-analysis. *Age Ageing* 2014;43:326-33.

5. Siddiqi N, House AO, Holmes JD. Occurrence and outcome of delirium in medical in-patients: a systematic literature review. *Age Ageing* 2006;35:350-64.
6. Hapca S, Guthrie B, Cvoro V, et al. Mortality in people with dementia, delirium, and unspecified cognitive impairment in the general hospital: prospective cohort study of 6,724 patients with 2 years follow-up. *Clin Epidemiol* 2018;10:1743-53.
7. Rosgen BK, Krewulak KD, Stelfox HT, et al. The association of delirium severity with patient and health system outcomes in hospitalised patients: a systematic review. *Age Ageing* 2020;49:549-57.
8. O'Keeffe S, Lavan J. The prognostic significance of delirium in older hospital patients. *J Am Geriatr Soc* 1997;45:174-8.
9. Kim S, Holsinger T. Delirium in elderly patients and the risk of post-discharge mortality, institutionalization, and dementia: a meta-analysis. In: Tampi RR, Tampi DJ, Young JJ, Balasubramaniam M, Joshi P, eds. *Essential Reviews in Geriatric Psychiatry*. Springer International Publishing; 2022:133-6.
10. Leslie DL, Inouye SK. The importance of delirium: economic and societal costs. *J Am Geriatr Soc* 2011;59:S241-3.
11. Mattison MLP. Delirium. *Ann Intern Med* 2020;173:ITC49-64.
12. Wang S, Lindroth H, Chan C, et al. A systematic review of delirium biomarkers and their alignment with the NIA-AA research framework. *J Am Geriatr Soc* 2021;69:255-63.
13. Pereira JV, Aung Thein MZ, Nitchingham A, et al. Delirium in older adults is associated with development of new dementia: a systematic review and meta-analysis. *Int J Geriatr Psychiatry* 2021; 36:993-1003.
14. Ritter SRF, Cardoso AF, Lins MMP, et al. Underdiagnosis of delirium in the elderly in acute care hospital settings: lessons not learned: delirium diagnosis: lessons not learned. *Psychogeriatrics* 2018;18:268-75.
15. Teodorczuk A, Reynish E, Milisen K. Improving recognition of delirium in clinical practice: a call for action. *BMC Geriatr* 2012;12:55.
16. Fick D, Foreman M. Consequences of not recognizing delirium superimposed on dementia in hospitalized elderly individuals. *J Gerontol Nurs* 2000;26:30-40.
17. Kales HC, Kamholz BA, Visnic SG, et al. Recorded delirium in a national sample of elderly inpatients: potential implications for recognition. *J Geriatr Psychiatry Neurol* 2003;16:32-8.
18. Moss SJ, Hee Lee C, Doig CJ, et al. Delirium diagnosis without a gold standard: Evaluating diagnostic accuracy of combined delirium assessment tools. *PLoS One* 2022;17:e0267110.
19. Sinvani L, Kozikowski A, Pekmezaris R, et al. Delirium: a survey of healthcare professionals' knowledge, beliefs, and practices. *J Am Geriatr Soc* 2016;64:e297-303.
20. Sinvani L, Hajizadeh N. The trouble with delirium—pitfalls of measurement in critical illness. *Crit Care Med* 2019;47:e381.
21. Sinvani L, Kozikowski A, Patel V, et al. Nonadherence to geriatric-focused practices in older intensive care unit survivors. *Am J Crit Care* 2018;27:354-61.
22. Nitchingham A, Caplan GA. Current challenges in the recognition and management of delirium superimposed on dementia. *Neuropsychiatr Dis Treat* 2021;17:1341-52.
23. Schmitt EM, Marcantonio ER, Alsop DC, et al. Novel risk markers and long-term outcomes of delirium: the Successful Aging after Elective Surgery (SAGES) study design and methods. *J Am Med Dir Assoc* 2012;13:818.e1-10.
24. Toft K, Tontsch J, Abdelhamid S, et al. Serum biomarkers of delirium in the elderly: a narrative review. *Ann Intensive Care* 2019;9:76.
25. Amgarth-Duff I, Hosie A, Caplan G, et al. Toward best practice methods for delirium biomarker studies: an international modified Delphi study. *Int J Geriatr Psychiatry* 2020;35:737-48.
26. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, fifth edition*. Arlington (VA): American Psychiatric Association Publishing; 2013.
27. Trzepacz PT, Mittal D, Torres R, et al. Validation of the Delirium Rating Scale-Revised-98: comparison with the Delirium Rating Scale and the Cognitive Test for Delirium. *J Neuropsychiatry Clin Neurosci* 2001;13:229-42.
28. Meagher DJ, Moran M, Raju B, et al. Phenomenology of delirium: Assessment of 100 adult cases using standardised measures. *Br J Psychiatry* 2007;190:135-41.
29. Wallesch CW, Hundsalz A. Language function in delirium: a comparison of single word processing in acute confusional states and probable alzheimer's disease. *Brain Lang* 1994;46:592-606.
30. Green S, Reivonen S, Rutter LM, et al. Investigating speech and language impairments in delirium: a preliminary case-control study. *PLoS One* 2018;13:e0207527.
31. Parola A, Simonsen A, Bliksted V, et al. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. *Schizophr Res* 2020;216:24-40.
32. Slegers A, Filiou RP, Montembeault M, et al. Connected speech features from picture description in Alzheimer's disease: a systematic review. *J Alzheimers Dis 2018*;65:519-42.
33. Liu N, Luo K, Yuan Z, et al. A transfer learning method for detecting Alzheimer's disease based on speech and natural language processing. *Front Public Health* 2022;10:772592.
34. Grand G, Blank IA, Pereira F, et al. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nat Hum Behav* 2022;6:975-87.
35. Clarke N, Foltz P, Garrard P. How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease. *Cortex* 2020;129:446-63.
36. Lindsay H, Tröger J, König A. Language impairment in Alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Front Aging Neurosci* 2021;13:642033.
37. Roshanzamir A, Aghajan H, Soleymani Baghshah M. Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech. *BMC Med Inform Decis Mak* 2021;21:92.
38. Andreasen NC. Scale for the assessment of Thought, Language, and Communication (TLC). *Schizophr Bull* 1986;12:473-82.
39. Liddle PF, Ngan ETC, Caissie SL, et al. Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *Br J Psychiatry* 2002;181:326-30.
40. Kircher T, Krug A, Stratmann M, et al. A rating scale for the assessment of objective and subjective formal Thought and Language Disorder (TALD). *Schizophr Res* 2014;160:216-21.
41. Tang SX, Kriz R, Cho S, et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophr* 2021;7:25.
42. Subbe CP, Kruger M, Rutherford P, et al. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001;94:521-6.
43. Katz S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *J Am Geriatr Soc* 1983;31:721-7.
44. Inouye SK. Clarifying confusion: the confusion assessment method: a new method for detection of delirium. *Ann Intern Med* 1990;113:941.
45. Sessler CN, Gosnell MS, Grap MJ, et al. The Richmond Agitation–Sedation Scale: validity and reliability in adult intensive care unit patients. *Am J Respir Crit Care Med* 2002;166:1338-44.
46. Inouye SK, Kosar CM, Tommet D, et al. The CAM-S: development and validation of a new scoring system for delirium severity in 2 cohorts. *Ann Intern Med* 2014;160:526.
47. Wittenburg P, Brugman H, Russel A, et al. ELAN : a professional framework for multimodality research. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* 2006:1556-9. Available: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_60436 (accessed 2022 Sept. 1).
48. ELAN. Available: https://archive.mpi.nl/tla/elan (accessed 2022 June 2).
49. Andreasen NC. Scale for the Assessment of Negative Symptoms (SANS). University of Iowa; 1984.
50. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011;45:1-67.
51. Tombaugh TN, Kozak J, Rees L. Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Clin Neurpsychology* 1999;14:167-77.
52. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008;28:1-26.
53. Inouye SK, Zhang Y, Han L, et al. Recoverable cognitive dysfunction at hospital admission in older persons during acute illness. *J Gen Intern Med* 2006;21:1276-81.
54. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
55. van Eijk MM, van den Boogaard M, van Marum RJ, et al. Routine use of the confusion assessment method for the intensive care unit: a multicenter study. *Am J Respir Crit Care Med* 2011;184:340-4.