ARTICLE OPEN

Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders

Sunny X. Tang[1,2,3]✉, Reno Kriz[4,8], Sunghye Cho[3,8], Suh Jung Park[2], Jenna Harowitz[2], Raquel E. Gur[2], Mahendra T. Bhati[2,5], Daniel H. Wolf[2], João Sedoc[6] and Mark Y. Liberman[3,7]

Computerized natural language processing (NLP) allows for objective and sensitive detection of speech disturbance, a hallmark of schizophrenia spectrum disorders (SSD). We explored several methods for characterizing speech changes in SSD ($n = 20$) compared to healthy control (HC) participants ($n = 11$) and approached linguistic phenotyping on three levels: individual words, parts-of-speech (POS), and sentence-level coherence. NLP features were compared with a clinical gold standard, the Scale for the Assessment of Thought, Language and Communication (TLC). We utilized Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art embedding algorithm incorporating bidirectional context. Through the POS approach, we found that SSD used more pronouns but fewer adverbs, adjectives, and determiners (e.g., "the," "a,"). Analysis of individual word usage was notable for more frequent use of first-person singular pronouns among individuals with SSD and first-person plural pronouns among HC. There was a striking increase in incomplete words among SSD. Sentence-level analysis using BERT reflected increased tangentiality among SSD with greater sentence embedding distances. The SSD sample had low speech disturbance on average and there was no difference in group means for TLC scores. However, NLP measures of language disturbance appear to be sensitive to these subclinical differences and showed greater ability to discriminate between HC and SSD than a model based on clinical ratings alone. These intriguing exploratory results from a small sample prompt further inquiry into NLP methods for characterizing language disturbance in SSD and suggest that NLP measures may yield clinically relevant and informative biomarkers.

npj Schizophrenia (2021)7:25 ; https://doi.org/10.1038/s41537-021-00154-3

INTRODUCTION

Language disturbance has long been recognized as a hallmark of psychosis, ranging from marked disorganization in threshold schizophrenia spectrum disorders (SSD) to less pronounced phenotypes among subthreshold psychosis-spectrum conditions like schizotypy, clinical high risk, and genetic risk for psychosis[1–5]. Speech is considered the observable surface phenomena that reveal, in part, the shrouded thoughts of the inner "mind." Thus, thought disorder in psychosis has often been equated with speech disturbance[6]. Novel digital phenotyping and computerized natural language processing (NLP) methods offer the opportunity to capture speech in ecologically valid settings and automatically quantify objective parameters that reflect underlying thought disturbance[7]. This would enable substantial advancements in assessment and monitoring of treatment response and could be used in research to probe more deeply into the psychosis disease process. However, a great deal remains to be understood about how to best leverage both established and state-of-the-art NLP tools to tease out clinically meaningful linguistic parameters that will have diagnostic and prognostic value.

A variety of NLP techniques have been used to characterize different language phenotypes in psychosis with varying degrees of success, mostly focusing on SSD[8]. Here, "schizophrenia spectrum disorders" (SSD) refers to threshold-level primary psychotic disorders[9,10]. Elvevåg et al.[11] first applied Latent Semantic Analysis to quantify decreased coherence in SSD speech and was able to predict human ratings of organizational structure, tangentiality, and content, as well as discriminating between SSD and control participants with 80–82% accuracy. Several subsequent studies have also modeled decreased linguistic cohesion and found significant group differences between SSD and control participants[12,13]. Decreased semantic coherence may predict conversion to psychotic disorders among young people at clinical risk for psychosis with 70–80% accuracy[5,14]. Other successful approaches have included graph analysis[15,16], quantifying semantic density[17], and automated metaphor detection[18]. In one study, NLP analysis of semantic cohesion accounted for 10% of variance in neurocognition beyond clinician ratings[19]. Another study used automated speech recognition and scoring on a smartphone verbal memory test to accurately predict human ratings[20]. Multiple levels of analysis including generic features like number of words per sentence, individual word identity, dictionary features reflecting word categories, and language model features like n-grams were incorporated in another study to discriminate between SSD and control speech with 74% accuracy[21].

However, other studies have demonstrated negative or conflicting results, suggesting that a great deal remains to be understood about applying NLP to evaluate speech disturbance in psychosis. Despite affective flattening being a common negative symptom in SSD, use of emotion words were nevertheless similar

[1]Zucker Hillside Hospital, Department of Psychiatry, 75-59 263rd St., Glen Oaks, NY, USA. [2]University of Pennsylvania, Department of Psychiatry, 3400 Spruce St, Gates Building, Philadelphia, PA, USA. [3]Linguistics Data Consortium, 3600 Market St, Suite 810, Philadelphia, PA, USA. [4]University of Pennsylvania, Department of Computer Science, 3330 Walnut St, Levine Hall, Philadelphia, PA, USA. [5]Stanford University, Department of Psychiatry and Neurosurgery, 401 Quarry Road, Stanford, CA, USA. [6]New York University, Department of Technology, Operations, and Statistics, 44 West Fourth Street, Kaufman Management Center, New York, NY, USA. [7]University of Pennsylvania, Department of Linguistics, 3401-C Walnut St, Suite 300, C Wing, Philadelphia, PA, USA. [8]These authors contributed equally: Reno Kriz, Sunghye Cho. ✉email: Stang3@northwell.edu

Published in partnership with the Schizophrenia International Research Society

between SSD and control participants[22]. Cohesion was not significantly reduced in another study comparing first-episode psychosis and healthy control (HC) participants[23]. Multiple levels of acoustic and linguistic analysis were shown to have a poor correlation with clinician ratings for negative symptoms[7]. In general, despite seemingly straightforward analogies between acoustic features like pitch variability and pause duration with negative symptoms like affective flattening and alogia, NLP measures have shown poor ability to track negative symptoms, including affective flattening and alogia[24]. Existing studies in psychiatry also use older NLP methods that have been supplanted by more advanced methods in the computational field. In particular, previous work on categorizing psychosis spectrum language samples[12,14] have leveraged non-contextual word embeddings such as GloVe and Word2Vec[25]. For example, these older NLP methods are unable to incorporate contextual cues that would distinguish between the word "banks" in "I rob banks" and "I walked along the banks."

In this study, we explored multiple NLP methods for characterizing speech changes in SSD and approached linguistic phenotyping on three levels: individual words, parts-of-speech (POS), and sentence-level measures of coherence. The NLP linguistic parameters were compared with a clinical gold standard, the Scale for the Assessment of Thought, Language and Communication (TLC)[26], in their ability to discriminate between speech samples from SSD and HC participants. To our knowledge, we are the first to apply Bidirectional Encoder Representations from Transformers (BERT) to a psychiatric sample; BERT is a state-of-the-art embedding architecture that is trained to incorporate bidirectional context and predict whether two sentences are adjacent[27]. This advance is significant because it allows us to look at larger units of discourse, and provides more accurate representation of the concepts being conveyed—treating discourse as an ordered series of information exchanges with a direction rather than an unordered collection of phrases. Moreover, BERT has been shown to be superior to other methods on several NLP tasks[27]. The samples were not enriched for overt thought or language disorder, in order for the speech and language phenotype to be more representative of the range present in SSD as a whole. Because this was an exploratory study and because no consistent set of key NLP predictors have emerged from prior work, we did not approach the analyses with a priori expectations of specific features that would be associated with SSD. Instead, we broadly hypothesized that speech from individuals with SSD would show abnormalities on each level of analysis: individual words, POS, and sentences. In addition, we expected that NLP would be better able to discriminate between SSD and HC language than the gold standard clinical rating scale for language disturbance. Of note, we encountered unexpected findings regarding incomplete words in SSD and potential methodological pitfalls, which are also detailed below.

RESULTS

Clinical ratings of language disorder

Individual TLC items are detailed in Supplemental Table 2. Clinically significant language disorder was present in four participants with SSD and none of the HC participants. Three of the SSD participants with TLC global score ≥2 were identified as outliers in this sample. However, there were no significant group differences in the TLC global, sum, or any individual item score (Table 1, Fig. 1). The largest effect sizes were in poverty of content of speech (Cohen's $d = 0.70$) and illogicality (Cohen's $d = 0.51$).

NLP measures for individual words and POS

SSD and HC employed distinct word usage patterns, with top SSD- and HC-associated words listed in Supplemental Table 3. Of note, top SSD-associated words included first-person singular ("I," etc.) and second-person pronouns ("you," etc.), the filler word "uh," and incomplete words. Top HC-associated words included first-person plural pronouns ("we," etc.), the filler word "um," and laughter. Supplemental Fig. 1 further details the propensity for participants with SSD to use the first-person singular over plural pronouns, and the filler word "uh" over "um." The use of incomplete words, on its own, was able to discriminate between SSD and HC groups with AUC = 0.88, accuracy = 90% (Supplemental Fig. 2A). With leave-one-out cross validation, word usage alone discriminated between SSD and HC with AUC = 0.80, accuracy = 76% (Supplemental Fig. 2B).

Mean POS counts per 100 words and group comparisons while covarying for demographic variables, education level, and cohort are reported in Table 2. Participants with SSD produced fewer adverbs ($p = 0.001$), adjectives ($p = 0.03$), and determiners ($p = 0.03$) but produced more pronouns ($p = 0.03$). The group differences in the other POS categories were not significant, including for interjections as a whole (despite differences for the specific filler words, "um" and "uh"). Excluding the three SSD outliers, the group differences in adverbs ($p < 0.001$), adjectives ($p = 0.04$), and determiners ($p = 0.03$) remained but the difference in pronouns became trend-level ($p = 0.09$).

Table 1. Sample Characteristics.

	HC	SSD	p value	Cohen's d
Sample				
n	11	20		
Cohort			0.10	
Cohort 1	5	15		
Cohort 2	6	5		
Age (mean years ± SD)	35.6 ± 5.8	36.5 ± 7.2	0.75	0.12
Sex (n, %)				
Female	7 (64%)	9 (45%)	0.32	
Male	4 (36%)	11 (55%)		
Race (n, %)			0.12	
African American	3 (30%)	13 (65%)		
Asian	0 (0%)	1 (5%)		
Caucasian	7 (70%)	6 (30%)		
Education level	15.8 ± 2.2	13.4 ± 2.5	0.01	−1.00
Recording Characteristics				
Recording duration (min)	11.6 ± 2.2	12.7 ± 4.5	0.48	0.29
Mean sentence length (words)	17.5 ± 3.1	14.4 ± 4.3	0.04	0.81
Word count	1748.8 ± 448.0	1782.3 ± 908.2	0.92	0.04
Language Measures				
TLC Global Score	0.0 ± 0.0	0.5 ± 1.0	0.13	0.56
TLC Total Score	0.9 ± 1.7	4.4 ± 9.2	0.10	0.46
Next-sentence predictability	0.96 ± 0.03	0.94 + 0.04	0.25	-0.44

TLC global score is an overall impression of speech and language disturbance based on standard anchors. TLC total score is summed using the published formula, Total = 2*(Sum of items 1–11) + (Sum of items 12–18). Next-Sentence Predictability derived from BERT, with 0 indicating low predictability and 1 indicating high predictability. Additional details about the SSD participants are provided in Supplemental Table 1.
Bolded – categories; *HC* healthy control participants, *SSD* participants with schizophrenia spectrum disorder, *TLC* Scale for the Assessment of Thought Learning and Communication (Andreasen[26]).

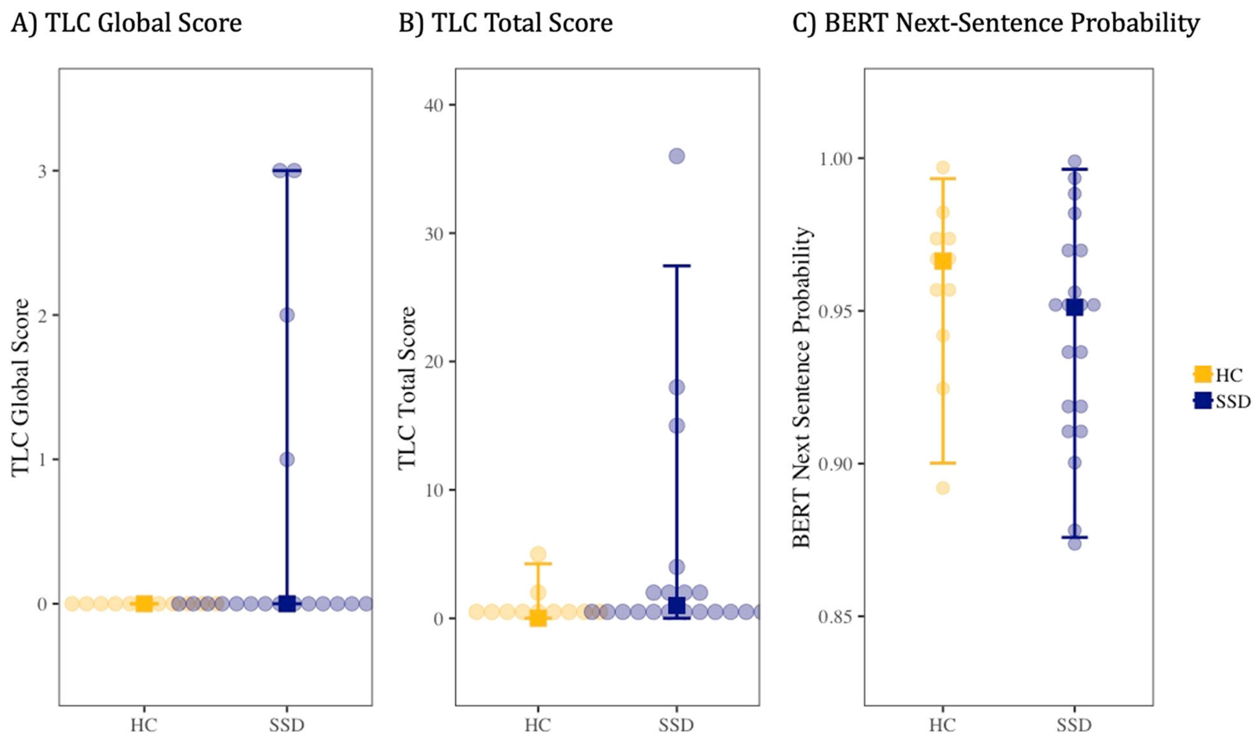

Fig. 1 Group effects on clinical language ratings and BERT next-sentence probability. Individual scores with group median and interquartile range are shown. There were no significant group differences for (**A**) TLC Global Score (Cohen's $d = 0.55$, $p = 0.13$), (**B**) TLC Total Score or Sum (Cohen's $d = 0.48$, $p = 0.10$), and (**C**) BERT Next-Sentence Probability (Cohen's $d = -0.44$, $p = 0.25$). The three SSD participants with TLC global scores ≥2 were identified as outliers.

Table 2. Parts-of-speech frequencies in SSD and HC.

	HC (*N* = 11)	SSD (*N* = 20)	*p*-value	Cohen's *d*
Adverb	10.65 (0.95)	8.11 (1.76)	0.001	1.66
Determiner	7.50 (0.96)	6.53 (1.25)	0.03	0.83
Adjective	7.10 (1.57)	6.19 (0.78)	0.03	0.82
Pronoun	11.77 (1.47)	13.41 (2.67)	0.03	−0.71
Preposition	8.84 (1.37)	7.97 (1.41)	0.08	0.62
Particle	2.65 (0.52)	2.35 (0.50)	0.08	0.59
Conjunction	5.33 (1.35)	4.61 (1.40)	0.29	0.53
Noun	13.16 (0.93)	13.67 (2.16)	0.57	−0.28
Interjection	6.07 (1.66)	6.35 (2.45)	0.75	−0.12
Verb	19.34 (1.74)	19.55 (2.60)	0.82	−0.09

HC healthy control participants, *SSD* participants with schizophrenia spectrum disorder, *Adverb* word that modifies an adjective or verb, *Determiner* determines the kind of reference for a noun, e.g. the, this, a, *Adjective* word that modifies a noun, *Pronoun* word that refers to the self or another noun mentioned elsewhere, e.g., I, she, them, *Preposition* word expressing relation to another clause, e.g. on, after, for, *Particle* function word providing meaning to associated words, e.g., **to** run, ate **up**, talk **over**, *Conjunction* connector word, e.g., and, but, because, *Noun* a person, place, thing, state, or quality, *Interjection* utterance expressing emotion, e.g., ouch, ugh, hey, *Verb* word expressing action, state, or relation.
P-values shown for ANCOVA tests, co-varying for education level, study cohort, and demographic variables (age, sex, and race).

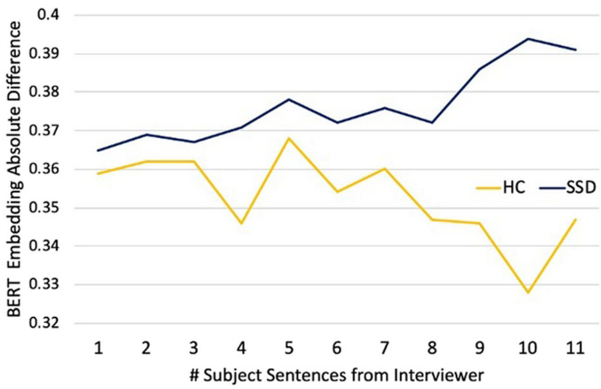

Fig. 2 Sentence embedding distance by interviewer-participant exchanges. The average BERT sentence embedding difference between the original interviewer prompt and the participant's response sentence, varying by distance from the original prompt. Responses from individuals with SSD began significantly farther from interviewer prompts relative to HC and traveled increasingly father away where those of HC did not. Fitting linear regressions to the trajectories, we find: SSD intercept = 0.260, 95% CI [0.257, 0.263]; SSD slope = 6.6e−4, 95% CI [2.6e−4, 1.1e−3]; HC intercept = 0.247, 95% CI [0.242–0.252]; HC slope = 1.5e−5, 95% CI [−6.2e−4, 6.5e−4]. The analysis was repeated excluding the 3 SSD outliers with high TLC scores and the results were consistent.

Sentence-level NLP results

In a linear regression model covarying for demographic variables, education level, and cohort, group failed to predict mean BERT next-sentence probability scores (Table 1; Beta coefficient = 0.01, $p = 0.28$). Weighting by normalized sentence length and exclusion of very brief sentences (less than five words) did not alter the null results. Figure 2 shows the average BERT embedding difference between the original interviewer prompt and the participant's response sentence, varying by distance from the original prompt. Modeling the trajectories for embedding distance, we found the intercept for SSD (0.260, 95% CI [0.257, 0.263]) was significantly higher than of HC (0.247, 95% CI [0.242–0.252]), suggesting that

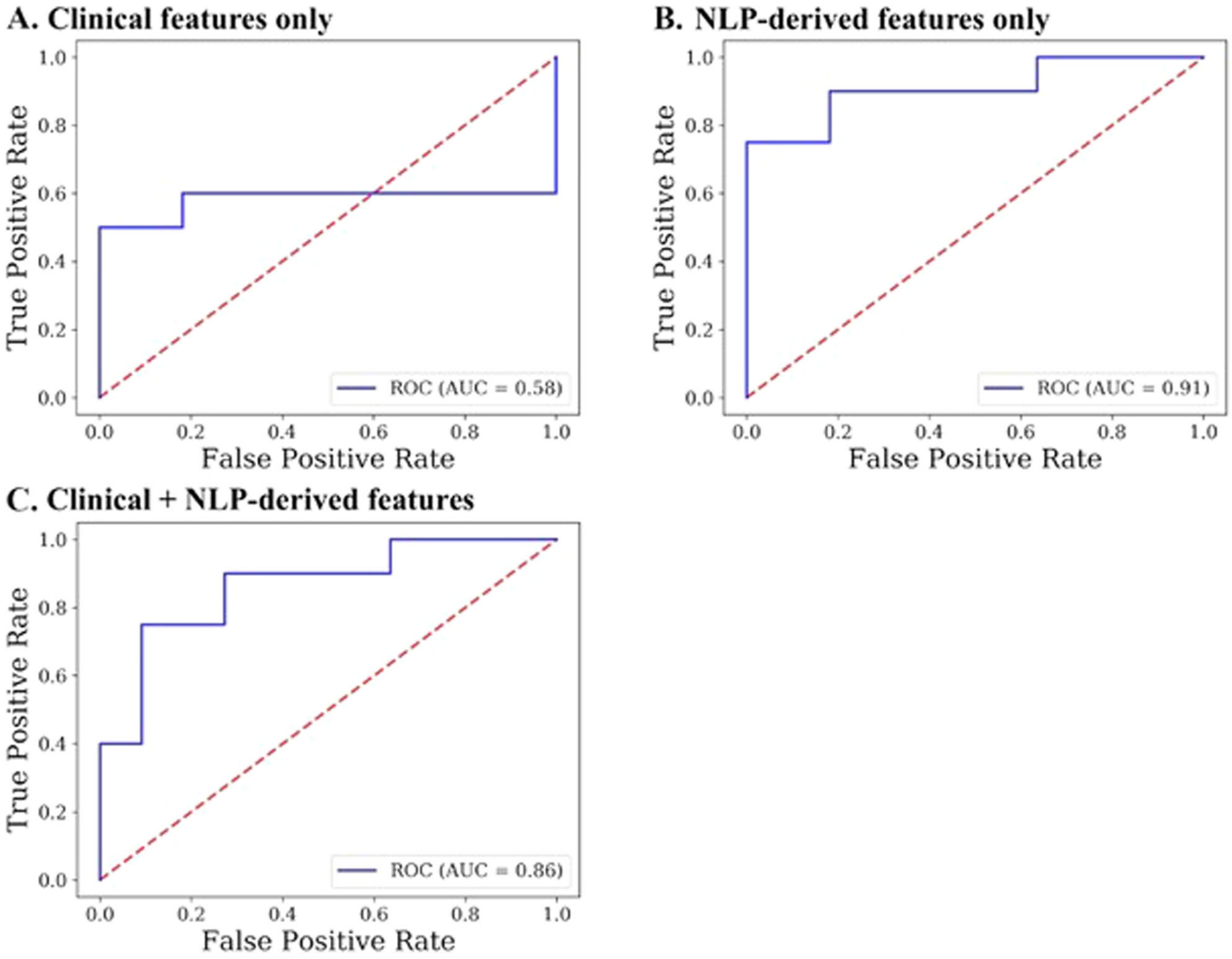

Fig. 3 Discrimination between SSD and HC group status. Naive Bayes models with leave-one-out cross-validation were constructed for (**A**) Clinical features alone, from the Scale for the Assessment of Thought Language and Communication (TLC), (**B**) Natural language processing-derived linguistic features alone, and (**C**) A combination of TLC and NLP-derived features. The NLP-only model performed better than the clinical-only model (accuracy 87% compared to 68%) and was similar to the model incorporating both NLP and clinical linguistic features (accuracy 81%).

responses were more unrelated to interviewer prompts. Furthermore, the slope of the SSD embedding trajectory was significantly above zero (6.6e−4, 95% CI [2.6e−4, 1.1e−3]) but the slope for HC was not significantly different from zero (1.5e−5, 95% CI [−6.2e−4, 6.5e−4]), suggesting that SSD sentences diverged significantly from interviewer prompts while HC sentences did not. Omitting the 3 outliers from the analysis did not alter this pattern.

Comparing clinical and NLP language measures

Figure 3 shows receiver operative characteristics for naive Bayes models discriminating between SSD and HC groups based on language measures. Using leave-one-out cross-validation, area under the curve (AUC) for clinical measures alone (0.58, Accuracy = 68%) was substantially lower than for the model based on NLP measures alone (AUC = 0.91, Accuracy = 87%). Adding clinical measures to the NLP model did not improve discriminating ability (AUC = 0.86, Accuracy = 81%). The inclusion of education level as an additional predictor did not substantially change the accuracy of these models—NLP alone (Accuracy = 87%) and NLP + TLC measures (Accuracy = 77%) continued to outperform the model based on TLC measures alone (Accuracy = 68%). We also confirmed these results using fivefold cross validation and found a similar pattern, and higher accuracy. There was insufficient variability in clinical ratings to explore the correlation between TLC and NLP measures.

DISCUSSION

In this exploratory study, we applied NLP to multiple levels of linguistic analysis comparing speech among SSD and HC participants. Individuals with SSD used significantly more pronouns and made more speech errors/partial words but used fewer adverbs and determiners. Analysis of individual word usage was notable for more frequent use of first-person singular pronouns among individuals with SSD and more frequent use of first-person plural pronouns among HC. The filler “uh” was highly associated with SSD, while “um” was associated with HC. There was also a striking increase in incomplete words among the SSD group. Sentence-level analysis using BERT reflected potentially increased tangentiality among SSD; sentence embedding distance increased for participants with SSD more than for HC. Participants with SSD generally had low language disorder and group means were not distinguishable from HC based on blinded clinical TLC ratings. It is unclear whether the specific linguistic group differences reported here represent clinically relevant, generalizable features of language disturbance in SSD. However, on the whole, NLP measures of language disturbance appear to be sensitive to these subclinical differences as NLP-based models had a greater ability to discriminate between HC and SSD than a model based on clinical ratings—whether or not education level was included as an additional predictor.

Differences in pronoun usage have been noted by several other studies. Buck et al.[28] also found increased pronoun use overall and increased first-person singular pronoun use among individuals with SSD when compared to HC. While the differences were not significant, they found similar trends with increased use of second-person pronouns among SSD and increased use of first-person plural pronouns among HC[28]. Increased use of first-person pronouns was also found among written text samples from individuals with SSD compared to HC[29]. Another study found significantly increased use of second-person pronouns among individuals with genetic risk for schizophrenia who later developed schizophrenia, compared to those who did not[30]. These differences in pronoun use appear to be context specific. Analysis of Twitter language found increased use of first-person singular and plural pronouns among posts from individuals who self-identified as having schizophrenia[31]. When compared to individuals with depression, those with SSD used fewer first-person singular pronouns[32]. Further delineation of differences in pronoun use is necessary as this is among the more highly replicated objective language disturbances in SSD and may reflect informative underlying differences in self-concept and/or social and metacognition[28,32].

"Um" and "uh" are frequently referred to as filled pauses, but may also be considered interjections signaling a preparedness problem on the part of the speaker and an impending delay[33,34]. "Um" takes a longer time to pronounce and usually signals a longer delay than "uh"[33]. The overall occurrence of filled pauses in speech from individuals with schizophrenia was previously found to be similar to that of neurotypical controls, despite greater occurrence of unfilled pauses or blocking prior to initiation of a new clause[34,35]. Our finding of greater "uh" occurrence among SSD and greater "um" occurrence among HC has not been previously reported, to our knowledge. In fact, these filler words are often excluded from analysis[12]. However, interestingly, a similar difference in filler word usage has been described among children with autism spectrum disorder with individuals with ASD using a higher proportion of "uh" compared to "um"[36]. While any connection between these findings is purely speculative at this point, it does raise the fascinating possibility that shared changes in brain circuitry may lead to a preference for "uh" over "um" among individuals with SSD and ASD. It remains to be determined whether the greater prevalence of "uh" spoken by SSD corresponds to a greater prevalence of short pauses, compared to long pauses, or whether this difference in "um"/"uh" usage is a symptom of impaired or inaccurate signaling by speakers with SSD, dissociated from pause duration.

Similarly, we are the first to report an increase in incomplete spoken words among individuals with schizophrenia. While decreased fluency has been noted in general, to our knowledge, the phenomenon of stopping mid-word has not been discussed in the literature in relation to SSD despite our finding that this one feature discriminates between SSD and HC with AUC 0.88. It is likely that this finding has not been replicated because these partial words were often considered meaningless and therefore not transcribed or discarded from the analysis. In live conversations, there is also a cognitive propensity not to notice these disfluencies[37]. If this finding is replicated, the use of incomplete words may provide a powerful objective marker of language disturbance and language circuity abnormality among individuals with SSD.

In this study, NLP measures alone were better able to discriminate between SSD and HC than clinical TLC ratings. The prediction model achieved 87% accuracy with leave-one-out cross validation. This accuracy is similar to results from other computational linguistics analyses, which have been reported at 69%[14,16], 69–75%[38], 83%[39], 84%[18], and 87–93%[12]. Advances in this field may contribute substantially to our understanding of the neurobiological underpinnings of schizophrenia. NLP measures have been tied to specific cognitive deficits in schizophrenia-like attention and social cognitive impairments[28,40,41]. Outside of psychiatry, NLP measures from healthy adults predicted stress response-related gene expression with better accuracy than self-reports[42]. To better understand the pathological mechanisms of psychiatric disorders and to develop targeted treatments, quantitative objective biomarkers of changes in brain circuitry are sorely needed.

There are several limitations to this study. The sample is small and heterogeneous with respect to study conditions between cohorts, with a higher n of SSD participants in Cohort 1 than Cohort 2, accounting for 48% of the total sample. However, results did not differ when we covaried for cohort. We were powered for differences with large group effects and are likely to miss small and medium effects. Due to the exploratory nature of this investigation, we did not correct for multiple comparisons, and therefore some findings may be the result of "overfitting" and may not be generalizable. The low number of individuals with clinically evident thought disorder limited our ability to assess correlations between NLP and TLC measures, but also demonstrates the ability of NLP measures to detect subclinical language disturbances in SSD. As a field, careful attention should be paid to technical quirks which can bias results, like our experience with automated sentence parsing resulting in falsely decreased coherence in the SSD group (as measured by BERT next-sentence probability). Other groups have had similar experiences[12]. Some of these errors result from the fact that most NLP tools have been trained on written texts, not speech. There is a need for a large shared corpus of clinical speech samples to advance NLP methods in psychiatry research. Since BERT is specifically trained on sentence-to-sentence prediction, this was the most straightforward and natural way of applying BERT to the coherence task and within-sentence coherence was not examined in this study. Furthermore, prior studies suggest that medication effects may drive some language abnormalities in SSD, particularly impaired phonation, as well as decreased quantity and diversity of words spoken[43,44]. Future efforts should assess how NLP measures correlate with specific dimensions of psychosis symptoms and medications used, for which consistent data were not available in these cohorts.

To our knowledge, this is the first study to apply BERT or any state-of-the-art embedding method that incorporates bidirectional context. Given the limited sample size in this exploratory study, findings should not be considered definitive or generalizable. However, our results nevertheless prompt further inquiry into NLP methods for characterizing language disturbance in SSD and suggest that NLP measures as a whole may yield clinically relevant and informative biomarkers.

METHODS

Sample

Two cohorts are included and are described in Table 1 with further details on the SSD group given in Supplemental Table 1. Both cohorts were recruited at the University of Pennsylvania and underwent written informed consent for the research procedures. Study procedures were approved by the Institutional Review Board at the University of Pennsylvania. SSD participants were stable outpatients and underwent semi-structured diagnostic interviews and consensus diagnostic case conferences which conferred a diagnosis of schizophrenia or schizoaffective disorder under DSM IV criteria[45]. Individuals with diagnosed intellectual disabilities were excluded from both cohorts. There was no selection or enrichment of the sample for participants with thought or language disorder—i.e., we did not preferentially include individuals with clinically evident symptoms of speech disorganization. HC participants underwent the same diagnostic interviews and were determined to be free of major psychiatric disorders. Both cohorts provided recorded speech samples derived from open-ended interviewing. Cohort 1 was asked to talk about themselves. Cohort 2 was asked to recount positive and neutral memories. Cohort 1 recordings were significantly shorter than those of Cohort 2 when considering total participant and interviewer speech and pause duration (10.9 ± 1.7 vs. 15.5 ± 5.4 min, $p = 0.002$), but the difference in participant word count was not significant (1661 ± 567 vs. 1970 ± 1161 words, $p = 0.32$). There were no significant differences between the cohorts in participant age, sex, education level, mean sentence length or number of sentences spoken.

Comparing the SSD and HC groups, SSD participants had significantly lower mean education level than HC. This difference was accounted for in the statistical analyses. Language samples from SSD and HC were of similar duration and total word count, but HC used longer sentences than SSD.

Recordings were transcribed verbatim by human annotators for NLP analysis, including noting non-verbal vocalizations like laughter and disfluencies like "um." Transcriber v.1.5.2[46] was used to transcribe audio samples from Cohort 2, while samples from Cohort 1 were transcribed using standard word processing software. To protect participant privacy, all personal or identifying references to an individual or location's name were replaced with "[name]" and all dates replaced with "[date]."

Clinical assessments

Recordings and transcripts were reviewed by a blinded psychiatrist with significant experience with schizophrenia research and treatment (SXT). The Scale for the Assessment of Thought, Language and Communication (TLC) was used to rate participant speech on 18 items, as well as a global measure of language disorder (ranging from 0-Absent to 4-Extreme) and a summation score calculated as per the published formula detailed in Supplemental Table 2[26]. The TLC was chosen because it is freely available, highly cited, and includes a wide span of clinically evident linguistic features. The modal global severity score was 0 for both HC and SSD, reflecting generally mild or absent language disorder in the SSD group, likely because they were composed of relatively high-functioning individuals who are motivated to participate in research.

NLP analyses for individual words and POS

Individual words were identified and utilization was compared between SSD and HC language samples. To capture larger trends, all pronouns of the same type were considered together. For example, "I" "my" and "me" were all counted as first-person singular pronouns. All incomplete words where the participant began to say a word but stopped before finishing were noted in aggregate. Odds ratios were calculated for words spoken by individuals with SSD relative to HC. The odds ratios were log-transformed and underwent weighting based on informative Dirichlet prior[47], which takes into account the expected frequency of each word in a random text and selects for words that are more "unique" to these documents.

POS in participant speech were automatically tagged for all tokenized, individual words with spaCy, using their basic model ('en_core_web_sm') for English[48]. Token counts for each POS category were calculated per 100 words and compared between the SSD and HC groups.

NLP sentence-level analyses using BERT

A major recent innovation in NLP has been the creation of large-scale pre-trained language models[27,49,50]. The most relevant pre-trained model for this work is known as Bidirectional Encoder Representations from Transformers, or BERT[27]. Unlike previous non-contextual embedding models, BERT allows for richer word representations, which takes the context surrounding a word into account. Trained on the English Wikipedia (2.5 billion words) and the Google BooksCorpus (800 million words), BERT operates with two general language understanding tasks: (1) Masked language modeling, where 15% of words from an input text are removed at random, and BERT predicts these words, and (2) Next-sentence prediction, where given two sentences, BERT predicts the likelihood of the second sentence following directly after the first.

We take advantage of the directionality incorporated into BERT next-sentence prediction to explore sentence-level coherence. For example, the BERT next-sentence probability for the below sentence pair is 1.0; this means the model is extremely confident that sentence 2 directly follows sentence 1. On the other hand, if the order of the sentences is reversed, the next-sentence probability is instead 0.002.

"Um, what do you think about current political issues like the energy crisis?"

"They're destroying too many cattle and oil just to make soap."

Excerpted from an example of language disturbance in the TLC Scale[26]

In this work, we employed two BERT-based methods to compare sentence-level differences between SSD and HC language samples, one based on next-sentence predictability, and the other based on BERT embeddings. In the next-sentence predictability approach, we performed sentence splitting on the interviews by using NLTK[51], which split sentences at intuitive punctuation points identified by transcribers. Sentence pairs were extracted from each interview, where the second sentence of each pair was uttered by the interview subject. From there, we directly leveraged BERT to predict the next sentence probability of each pair. Weighting the next-sentence probabilities by sentence length was attempted but did not alter the results. Note that initially, we used SpaCy to automatically split interview turns into sentences[48]. However, upon closer inspection, we discovered that this was incorrectly segmenting sentences at incomplete, filler, and repeated words. These segments (e.g., "I sto- // stopped at the store") were more frequent in SSD and were assigned low BERT next-sentence predictability scores and biased initial results toward a significant difference in mean next-sentence predictability. The results presented here instead used sentences segmented intuitively by transcribers, which considered filler words and repeated phrases as a part of the same sentence.

In the embedding approach, we broke down the interviews into full dialogue turns (exchanges beginning with an interviewer prompt and followed by any number of participant sentences until the interviewer speaks again for the next turn). We generated a single embedding e_i for the entire Interviewer dialogue turn t_i; we did this by generating BERT embeddings for each word in t, and computed the mean embedding across all words. From there, we generated a sentence-level BERT embedding e_{sj} for each sentence s_j in the subsequent subject turn and calculated the mean difference between e_i and e_{sj}. The intuition behind this is that if tangentiality or derailment is present, participant responses are likely to move further away from the initial interview prompt compared to coherent exchanges. Simple linear regression models were fit on sentence-wise embedding distance to compare the slopes and intercepts of the embedding trajectories for the two groups. As BERT is specifically trained on sentence-to-sentence prediction, our analyses were done on the sentence-level and within-sentence incoherence was not analyzed.

Statistical analysis

Normality was assessed with the Shapiro-Wilkes test and significant departure from normality was found for TLC global score ($p < 0.001$) and TLC sum ($p < 0.001$). Three outliers in the SSD group with high TLC scores were identified with the standard boxplot method. Group differences for these measures were compared with the Wilcoxon Rank Sum test with two-sided *alpha* = 0.05. Other measures were also tested for normality and we found they met the requirements for parametric tests. We built ANCOVA models for group comparisons of POS, covarying for education level, cohort, and demographic variables (age, sex, and race). All effect sizes were measured with Cohen's *d*. BERT next-sentence probability was calculated for each participant by taking the mean of the BERT probability scores for every sentence uttered by the participant, directly from the BERT output. To evaluate group effects, linear models were constructed with mean BERT next-sentence probability as the outcome variable and group as predictor while covarying for cohort, education level, and demographic variables (age, sex, education level). Due to the exploratory nature of these analyses, we did not correct for multiple comparisons.

In order to test the discriminating ability of the features, we trained three different naive Bayes models from the scikit-learn package[52], varying the number of features. The naive Bayes model was chosen because predictors were assumed to be independent, as is common in statistical modeling, and we assumed that we know something about the distributions for the outcomes independent of the predictors. One model was trained with TLC ratings only, and another model was trained with linguistic features only (log-odds ratio, BERT scores, POS counts per 100 words, and number of incomplete words). The other model was trained with the linguistics features and the TLC sum ratings. In all models, we imputed missing values with the simple imputer function, and standardized the features with the StandardScaler function for effective learning. To assess for the effect of education level on model accuracy, we additionally ran these models while also including education level as a predictor. Leave-one-out cross-validation was chosen to minimize bias and used to evaluate the general performance of the models and report AUC and the mean accuracy scores in the results.

DATA AVAILABILITY

The datasets analyzed during the current study are not publicly available due to participant privacy and security concerns, including HIPAA regulations. Qualified researchers may contact the corresponding author for access.

CODE AVAILABILITY

The code used in this study is publicly available at https://github.com/rekriz11/nlp_schizophrenia. Files with the extension ".Rmd" were run in R v.4.0.3, files with extension ".py" were run in Python 3. Please cite this study in any use or adaptation of this code.

Received: 10 September 2020; Accepted: 26 March 2021;
Published online: 14 May 2021

REFERENCES

1. Andreasen, N. Thought, language and communication disorders. *Arch. Gen. Psychiatry* **36**, 1315–1321 (1979).
2. Bleuler, E. *Dementia Praecox or the Group of Schizophrenias*. (International Universities Press, 1950).
3. Minor, K. S., Cohen, A. S., Weber, C. R. & Brown, L. A. The relationship between atypical semantic activation and odd speech in schizotypy across emotionally evocative conditions. *Schizophr. Res.* **126**, 144–149 (2011).
4. Manschreck, T. C., Merrill, A. M., Jabbar, G., Chun, J. & DeLisi, L. E. Frequency of normative word associations in the speech of individuals at familial high-risk for schizophrenia. *Schizophr. Res.* **140**, 99–103 (2012).
5. Corcoran, C. M. et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17**, 67–75 (2018).
6. Andreasen, N. C. & Grove, W. M. Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophr. Bull.* **12**, 348–359 (1986).
7. Cohen, A. S. et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr. Res.* https://doi.org/10.1016/j.schres.2020.03.043 (2020).
8. Ratana, R., Sharifzadeh, H., Krishnan, J. & Pang, S. A Comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. *Front. Psychiatry* **10**, 1–15 (2019).
9. Chan, R. C. H., Mak, W. W. S., Chio, F. H. N. & Tong, A. C. Y. Flourishing with psychosis: a prospective examination on the interactions between clinical, functional, and personal recovery processes on well-being among individuals with schizophrenia spectrum disorders. *Schizophr. Bull.* **44**, 778–786 (2018).
10. Taylor, J. H. et al. Characteristics of youth with reported family history of psychosis spectrum symptoms in the Philadelphia Neurodevelopmental Cohort. *Schizophr. Res.* **216**, 104–110 (2020).
11. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr. Res.* **93**, 304–316 (2007).
12. Iter, D., Yoon, J. & Jurafsky, D. Automatic detection of incoherent speech for diagnosing schizophrenia. *Proc 5th Work. Comput. Linguist. Clin. Psychol*. https://doi.org/10.18653/v1/w18-0615 (2018).
13. Willits, J. A., Rubin, T., Jones, M. N., Minor, K. S. & Lysaker, P. H. Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia. *Schizophr. Res.* **197**, 365–369 (2018).
14. Bedi, G. et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* **1**, 15030 (2015).
15. Mota, N. B. et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE* **7**, 1–9 (2012).
16. Mota, N. B., Copelli, M. & Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophr.* **3**, 1–10 (2017).
17. Rezaii, N., Walker, E. & Wolff, P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* **5**, 9 (2019).
18. Gutiérrez, E. D., Corlett, P. R., Corcoran, C. M. & Cecchi, G. A. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. *EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc*. 2923–2930 https://doi.org/10.18653/v1/d17-1316 (2017).
19. Minor, K. S., Willits, J. A., Marggraf, M. P., Jones, M. N. & Lysaker, P. H. Measuring disorganized speech in schizophrenia: Automated analysis explains variance in cognitive deficits beyond clinician-rated scales. *Psychol. Med.* **49**, 440–448 (2019).
20. Holmlund, T. B. et al. Applying speech technologies to assess verbal memory in patients with serious mental illness. *npj Digit. Med.* **3**, 33 (2020).
21. Hong, K. et al. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry Res.* **225**, 40–49 (2015).
22. St-Hilaire, A., Cohen, A. S. & Docherty, N. M. Emotion word use in the conversational speech of schizophrenia patients. *Cogn. Neuropsychiatry* **13**, 343–356 (2008).
23. Mackinley, M., Chan, J., Ke, H., Dempster, K. & Palaniyappan, L. Linguistic determinants of formal thought disorder in first episode psychosis. *Early Interv. Psychiatry* https://doi.org/10.1111/eip.12948 (2020).
24. Parola, A., Simonsen, A., Bliksted, V. & Fusaroli, R. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. *Schizophr. Res.* **216**, 24–40 (2020).
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Proc. of the 26th International Conference on Neural Information Processing Systems* 3111–3119 (2013).
26. Andreasen, N. C. Scale for the assessment of thought, language, and communication (TLC). *Schizophr. Bull.* **12**, 473–482 (1986).
27. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2019).
28. Buck, B., Minor, K. S. & Lysaker, P. H. Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Compr. Psychiatry* **58**, 138–145 (2015).
29. Maatz, A. Use of the first-person pronoun in schizophrenia. *Br. J. Psychiatry* **205**, 409 (2014).
30. Watson, A. R. et al. Use of second-person pronouns and schizophrenia. *Br. J. Psychiatry* **200**, 342–343 (2012).
31. Birnbaum, M. L. et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *npj Schizophr.* **5**, 1–9 (2019).
32. Fineberg, S. K. et al. Word use in first-person accounts of schizophrenia. *Br. J. Psychiatry* **206**, 32–38 (2015).
33. Clark, H. H. & Fox Tree, J. E. Using uh and um in spontaneous speaking. *Cognition* **84**, 73–111 (2002).
34. Çokal, D. et al. Disturbing the rhythm of thought: speech pausing patterns in schizophrenia, with and without formal thought disorder. *PLoS ONE* **14**, 1–14 (2019).
35. Bersudsky, Y., Fine, J., Gorjaltsan, I., Chen, O. & Walters, J. Schizophrenia and second language acquisition. Prog. *Neuro-Psychopharmacol. Biol. Psychiatry* **29**, 535–542 (2005).
36. Parish-Morris, J. et al. Exploring Autism Spectrum Disorders Using HLT. 74–84 https://doi.org/10.18653/v1/w16-0308 (2016).
37. Zayats, V., Tran, T., Wright, R., Mansfield, C. & Ostendorf, M. Disfluencies and human speech transcription errors. in *Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH* 3088–3092 https://doi.org/10.21437/Interspeech.2019-3134 (2019).
38. Xu, S. et al. Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* 225–228 https://doi.org/10.1109/EMBC.2019.8857071 (2019).
39. Strous, R. D. et al. Automated characterization and identification of schizophrenia in writing. *J. Nerv. Ment. Dis.* **197**, 585–588 (2009).
40. Docherty, N. M. et al. Effects of social cognitive impairment on speech disorder in schizophrenia. *Schizophr. Bull.* **39**, 608–616 (2013).
41. Docherty, N. M., Strauss, M. E., Dinzeo, T. J. & St-Hilaire, A. The cognitive origins of specific types of schizophrenic speech disturbances. *Am. J. Psychiatry* **163**, 2111–2118 (2006).
42. Mehl, M. R., Raison, C. L., Pace, T. W. W., Arevalo, J. M. G. & Cole, S. W. Natural language indicators of differential gene regulation in the human immune system. *Proc. Natl Acad. Sci. USA* **114**, 12554–12559 (2017).
43. de Boer, J. N., Voppel, A. E., Brederoo, S. G., Wijnen, F. N. K. & Sommer, I. E. C. Language disturbances in schizophrenia: the relation with antipsychotic medication. *npj Schizophr.* **6**, 24 (2020).
44. Sinha, P., Vandana, V. P., Lewis, N. V., Jayaram, M. & Enderby, P. Evaluating the effect of risperidone on speech: a cross-sectional study. *Asian J. Psychiatr.* **15**, 51–55 (2015).
45. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders DSM-IV. (1994).

46. Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. Transcriber: development and use of a tool for assisting in speech corpora production. *Speech Commun.* **33**, 1–2 (2000).
47. Monroe, B. L., Colaresi, M. P. & Quinn, K. M. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Polit. Anal.* **16**, 372–403 (2008).
48. Honnibal, M. & Johnson, M. An Improved Non-monotonic Transition System for Dependency Parsing. in *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* 1373–1378 (Association for Computational Linguistics, 2015).
49. Peters, M. E. et al. Deep contextualized word representations. in *NAACL HLT 2018* 1, 2227–2237 (Association for Computational Linguistics (ACL), 2018).
50. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding with unsupervised learning. Technical Report, Open AI (2018).
51. Loper, E. & Bird, S. NLTK: the Natural Language Toolkit. in *Proc. ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* 1, 63–70 (Association for Computational Linguistics (ACL), 2002).
52. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

ACKNOWLEDGEMENTS

We thank the participants for their contributions. Funding sources include: Early Career Research Award, American Society of Clinical Psychopharmacology (S.X.T.); T32 MH019112 (S.X.T.); Microsoft Research Dissertation Grant (J.S.); DARPA grant HR0011-15-C-0115 (the LORELEI program), MH-62103, and M01RR0040.

AUTHOR CONTRIBUTIONS

S.X.T., R.K., S.C., R.E.G., M.T.B., D.H.W., J.S., and M.Y.L. contributed to the project design. J.H. was a key contributor to data collection. Data analysis was conducted by S.X.T., R.K., S.C., S.J.P., J.S., and M.Y.L. All listed authors contributed to manuscript preparation. R.K. and S.C. contributed equally to this work.

COMPETING INTERESTS

S.X.T. holds equity to North Shore Therapeutics, was a consultant for Neurocrine Biosciences, and received research funding from Winterlight Labs. The other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41537-021-00154-3.

Correspondence and requests for materials should be addressed to S.X.T.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.