Schizophrenia Bulletin vol. 49 suppl. 2 pp. S183–S195, 2023
https://doi.org/10.1093/schbul/sbac176

SUPPLEMENT ARTICLE

Language Analytics for Assessment of Mental Health Status and Functional Competency

Rohit Voleti[*,1], Stephanie M. Woolridge[2], Julie M. Liss[3,4], Melissa Milanovic[5], Gabriela Stegmann[3,4], Shira Hahn[3,4], Philip D. Harvey[6], Thomas L. Patterson[7], Christopher R. Bowie[2], and Visar Berisha[1,3,4]

[1]School of Electrical Computer, and Energy Engineering, Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ, USA[2]Department of Psychology, Queen's University, Kingston, ON, Canada[3]College of Health Solutions, Arizona State University, Phoenix, AZ, USA[4]Aural Analytics Inc., Scottsdale, AZ, USA[5]CBT for Psychosis Service at the Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada[6]Department of Psychiatry, University of Miami Miller School of Medicine, Miami, FL, USA[7]Department of Psychiatry, University of California, San Diego, La Jolla, CA USA

*To whom correspondence should be addressed; 3427 COOR Hall, 975 S. Myrtle Ave.Tempe, AZ 85287, USA; tel: +1(480) 727-6455, e-mail: rnvoleti@asu.edu

***Background and Hypothesis*: Automated language analysis is becoming an increasingly popular tool in clinical research involving individuals with mental health disorders. Previous work has largely focused on using high-dimensional language features to develop diagnostic and prognostic models, but less work has been done to use linguistic output to assess downstream functional outcomes, which is critically important for clinical care. In this work, we study the relationship between automated language composites and clinical variables that characterize mental health status and functional competency using predictive modeling. *Study Design*: Conversational transcripts were collected from a social skills assessment of individuals with schizophrenia ($n = 141$), bipolar disorder ($n = 140$), and healthy controls ($n = 22$). A set of composite language features based on a theoretical framework of speech production were extracted from each transcript and predictive models were trained. The prediction targets included clinical variables for assessment of mental health status and social and functional competency. All models were validated on a held-out test sample not accessible to the model designer. *Study Results*: Our models predicted the neurocognitive composite with Pearson correlation PCC = 0.674; PANSS-positive with PCC = 0.509; PANSS-negative with PCC = 0.767; social skills composite with PCC = 0.785; functional competency composite with PCC = 0.616. Language features related to volition, affect, semantic coherence, appropriateness of response, and lexical diversity were useful for prediction of clinical variables. *Conclusions*: Language samples provide useful information for the prediction of a variety of clinical variables that characterize mental health status and functional competency.**

Key words: natural language processing/social skills prediction/speech analysis/machine learning/schizophrenia/bipolar disorder

Introduction

Schizophrenia and bipolar disorder (BD) are chronic, severe mental illnesses that manifest early in life and persist throughout, presenting significant challenges to individuals, families, and healthcare providers. The symptoms associated with these conditions are a primary source of disability for affected individuals and can have a drastic detrimental impact on real-world functional outcomes, including attaining employment, forming personal relationships, and maintaining social connectedness.[1,2]

Speech and language abnormalities are included in criteria for diagnosing schizophrenia and BD and establishing symptom severity.[3] In schizophrenia, poverty of speech and disorganized or incoherent speech are common. In BD, depressive and manic mood states are associated with different speech and language symptoms.[3,4] Manic episodes are characterized by rapid pressured speech, increased verbosity, and flight of ideas.[5] Depressive episodes result in poverty of speech or increased pause times, similar to impairments associated with negative symptoms of schizophrenia.

Speech and language measures have promise as new biomarkers in digital health applications, with the potential

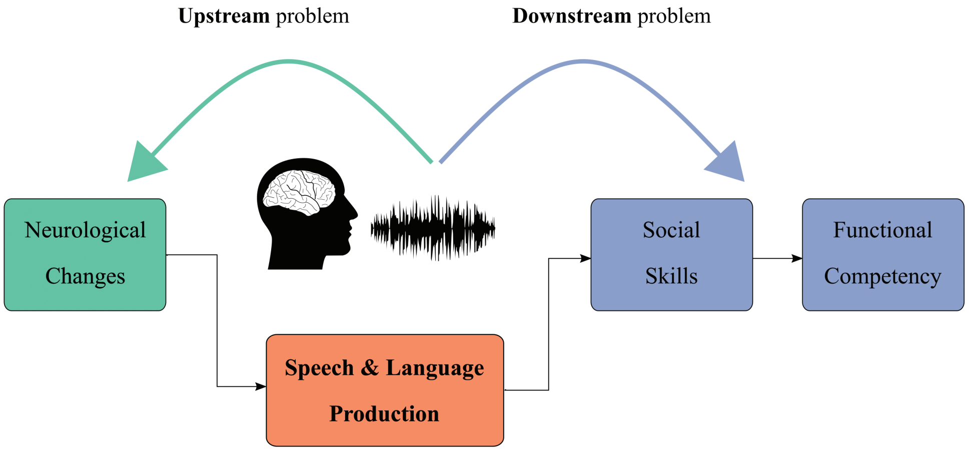


Fig. 1. Speech and language abnormalities are manifestations of the underlying upstream neurological function, and they impact the downstream functions of activities of daily living and participation (social and functional competency).

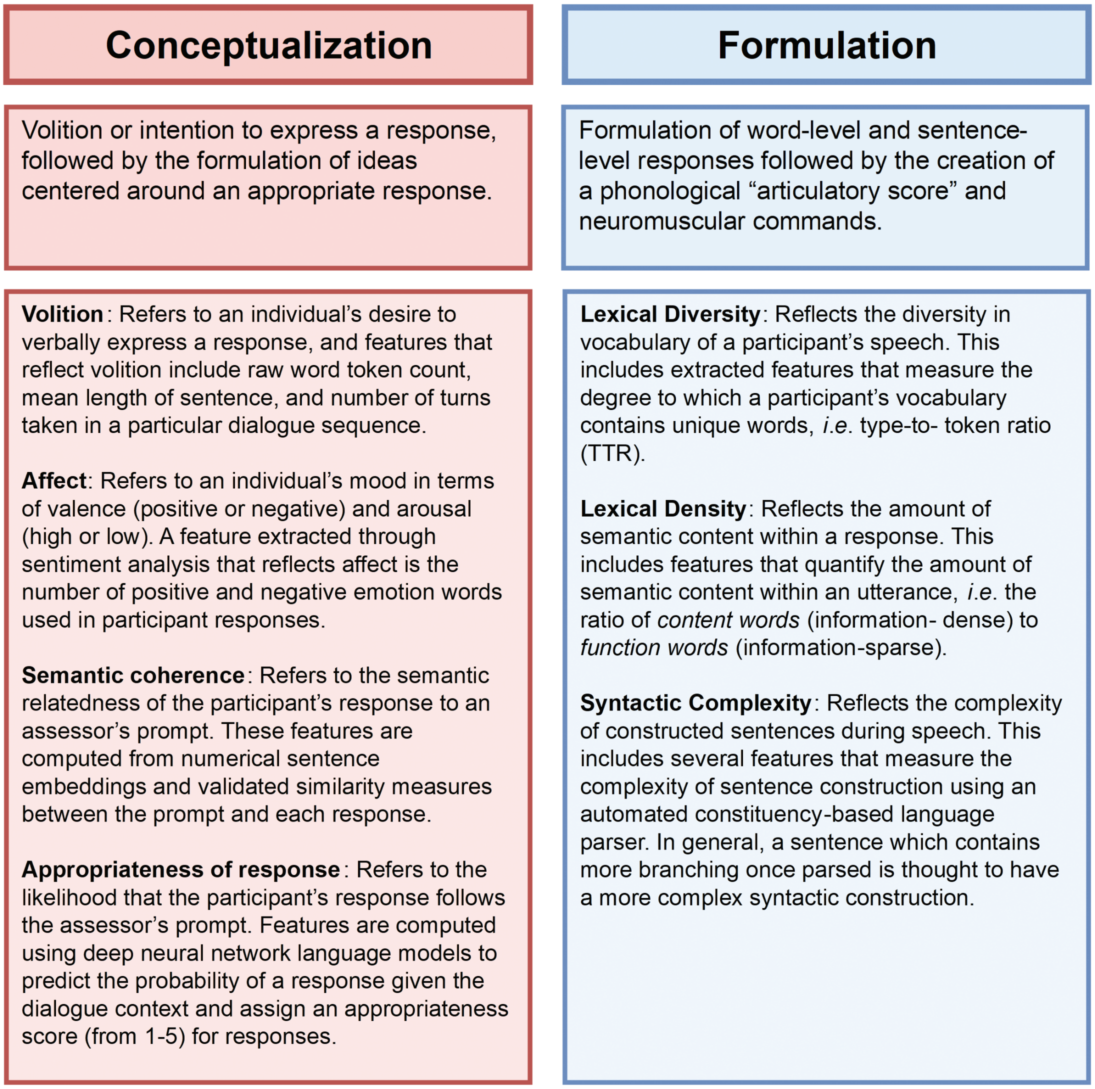
to extend clinical reach through remote assessment.[6–10] Figure 1 shows the central role that speech and language play in mental health disorders for ascertaining both the *upstream* information about presence, severity, and prognosis of the underlying neurological changes,[11,12] and the *downstream* impact of the symptoms on activities of daily living and quality of life. While a few studies have examined how speech and language relate to downstream functional competency,[2,11,13] most of the previous work in this area has focused on answering the upstream questions.

Several recent articles have reviewed the use of computational linguistic and speech processing methods for detection, assessment and identification of schizophrenia and BD.[6–10] A seminal study of language metrics to predict formal thought disorder (FTD)[14] compared healthy and FTD participants using latent semantic analysis (LSA)[15] to generate objective estimates of language similarity scores across samples elicited using a variety of tasks. Bedi et al.[16] and Corcoran et al.[17] also made use of LSA to predict the onset of psychosis in clinically high-risk youth. More recent work has also made use of neural word and sentence embeddings (ie, *word2vec*[18] and *GloVe*[19]) to assess similar types of coherence in speech from those with schizophrenia or BD.[20,21] A novel approach using neural word embeddings was recently proposed by Rezaii et al.[22] in which a vector unpacking approach was used to decompose an average sentence vector into its most significant components. They showed that low semantic density for given language elicitation tasks could reliably predict psychosis onset. Beyond semantics, other aspects of language have been computationally analyzed for individuals with schizophrenia and BD. Previous work has measured different features related to syntax,[23] conversational pragmatics,[24] linguistic complexity,[21] and ambiguous pronouns,[17] among others. These largely data-driven studies converge in identifying language metrics as useful upstream prognostic and diagnostic markers for schizophrenia and BD.

In contrast, few studies have explored the opportunity to use automated language metrics to objectively assess downstream problems, such as the impact of symptoms on social and functional competency.[25] Our work aims to use a set of language measures based on a theoretical model of speech production in service of this goal. The utility of upstream measures for earlier diagnosis and prognosis is self-evident; however, we posit that the downstream problem is equally important, and has been addressed to a much lesser extent. Illness recovery increasingly considers social participation and quality of life. Recently, there has been a concerted effort to develop digital therapeutics that target social competency in patients with schizophrenia.[26] To that end, objective proxies for constructs like social competency, which affect participation and quality of life, are critical in evaluating the real-world impact of interventions.

We expand on previous work by introducing a set of composite features guided by a theoretical model of speech and language production. We then use these features to develop robust models for addressing both the upstream and the downstream problems. Importantly, the models are evaluated on holdout test data that the machine learning model designer did not have access to during development. The composite representation we propose is based on a model proposed by Levelt[27,28] that characterizes spoken language production as a complex, multi-stage event consisting of three major stages:

1. *Conceptualization*: involves abstract idea formation and the intent or volition to communicate the idea.
2. *Formulation*: involves selection and sequencing of words and the precise linguistic construction of an utterance, along with a sensorimotor score for muscle activation.
3. *Articulation*: involves execution of this sensorimotor score by activation and coordination of speech musculature (i.e. respiratory, phonatory, articulatory, etc.)



Conceptualization

Volition or intention to express a response, followed by the formulation of ideas centered around an appropriate response.

Volition: Refers to an individual's desire to verbally express a response, and features that reflect volition include raw word token count, mean length of sentence, and number of turns taken in a particular dialogue sequence.

Affect: Refers to an individual's mood in terms of valence (positive or negative) and arousal (high or low). A feature extracted through sentiment analysis that reflects affect is the number of positive and negative emotion words used in participant responses.

Semantic coherence: Refers to the semantic relatedness of the participant's response to an assessor's prompt. These features are computed from numerical sentence embeddings and validated similarity measures between the prompt and each response.

Appropriateness of response: Refers to the likelihood that the participant's response follows the assessor's prompt. Features are computed using deep neural network language models to predict the probability of a response given the dialogue context and assign an appropriateness score (from 1-5) for responses.

Formulation

Formulation of word-level and sentence-level responses followed by the creation of a phonological "articulatory score" and neuromuscular commands.

Lexical Diversity: Reflects the diversity in vocabulary of a participant's speech. This includes extracted features that measure the degree to which a participant's vocabulary contains unique words, *i.e.* type-to- token ratio (TTR).

Lexical Density: Reflects the amount of semantic content within a response. This includes features that quantify the amount of semantic content within an utterance, *i.e.* the ratio of *content words* (information- dense) to *function words* (information-sparse).

Syntactic Complexity: Reflects the complexity of constructed sentences during speech. This includes several features that measure the complexity of sentence construction using an automated constituency-based language parser. In general, a sentence which contains more branching once parsed is thought to have a more complex syntactic construction.

Fig. 2. Two of the 3 stages of the speech production framework, a brief description of each stage (second row), and list of domains that characterize each stage (third row). We note that the "Articulation" stage is not included here because acoustic speech samples were not available for the transcripts studied.

This framework has previously been applied in work to assess depression, cognition, and thought disorders.[6,29] Figure 2 shows our instantiation of this framework focusing on Conceptualization and Formulation, along with a set of domains for each. We do not explore the articulation stage in this study because acoustic speech samples were not available for the transcripts studied herein. We posit that the theoretical framework provides a more interpretable means by which to interrogate the relationship between different aspects of speech production and clinical variables of interest than would a purely data-driven approach.

One of the challenges with operationalizing the framework is reliable measurement of the latent domains listed in figure 2. These are likely multidimensional constructs that have yet to be operationally defined in the literature. Briefly, our measurement model consists of three parts: extraction of a set of low-level features that have been used in previous work, mapping of these features to the individual Levelt stages, and denoising of these features using principal components analysis (PCA).[30–32] The denoising step is critical as there is converging evidence that out-of-the-box speech and language features may be highly variable, prone to confounding, and can exhibit poor test–retest reliability.[25,33] Furthermore, machine learning models built on top of these features sometimes exhibit poor external validity.[34] The Levelt model serves as a theoretical guide for grouping the less reliable low-level features that aim to represent similar constructs into composites. In the section that follows, we provide a high-level overview of the methods. The supplementary material provides a more detailed description of the methods.

Methods

Data Used for Model Development and Evaluation

Data from a total of 281 participants with a clinical diagnosis of either schizophrenia/schizoaffective (Sz/Sza) disorder (n = 140) or BD (n = 141) and 22 healthy controls were used in this study. The demographics of the participants in this study are seen in table 1. Participants underwent extensive clinical evaluations that consisted of neurocognitive batteries, symptom ratings, social, and functional assessments, including the Social Skills Performance Assessment (SSPA),[35] the Specific Level of Functioning (SLOF)[36] scale, the Positive and Negative Syndrome Scale (PANSS),[37] and a neurocognitive composite[2] from standardized z-scores from eight well-known neurocognitive batteries.

Language Samples: Language samples were elicited via the SSPA,[35] a role-playing task that can serve as a measurement of skills related to social competence. Participants are asked to act out the following three "scenes" with a clinical assessor:

- *Scene 1 (practice)*: plan a weekend activity with a friend (~1 min).
- *Scene 2 (scored)*: introduce a new neighbor to your neighborhood (~3 min).
- *Scene 3 (scored)*: negotiate with a difficult landlord to fix a leak in your apartment (~ 3 min).

Scenes 2 and 3 are individually scored on a scale from 1 to 5 across a variety of dimensions, such as overall interest/disinterest, affect, negotiation ability, and fluency. An overall score for each scene is computed by averaging the scores across each dimension for each scene.

The SSPA was administered by doctoral-level psychologists and coded by trained research assistants who were not aware of study aims, group membership, or hypotheses. The samples were manually transcribed by two research assistants and discrepancies were reviewed and corrected. The assistants were naïve to study design, group membership, or hypotheses.

Development/test split: Prior to the development of any models, the schizophrenia group and bipolar group were randomly split into two sets individually, a development set and a test set. The split was performed multiple times until the range of SSPA scores was approximately matched between the development and test set. The healthy controls were randomly split once as they did not have much variation in the SSPA. This was done to ensure we had sufficient variability in the SSPA samples in both the development and test set. The development set was used by the algorithm developer (Stegmann) to develop the model. Importantly, the algorithm developer did not have access to the test set at any point during model development. Once the models were fixed, they were shared with the first author, who evaluated the performance of the model on the test sets.

Language Feature Composites and Model Development. A detailed description of the computed features and model fitting is provided in the supplementary material. Briefly, we sorted our computed features into the seven domains across the two stages listed in figure 2, to create composite variables for assessment of schizophrenia and BD. For

Table 1. Demographic information for each of the cohorts

	Sample Size (Gender)	Age	Years of Education
Training			
Sz/Sza	N = 98 (37 F, 61 M)	μ: 51.27 σ^2: 10.10 R: 25–75	μ: 14.43 σ^2: 2.65 R: 6–20
BD	N = 98 (51 F, 47 M)	μ: 47.45 σ^2: 13.23 R: 18–80	μ: 16.08 σ^2: 2.20 R: 11–20
Control	N = 11 (3 F, 7 M, 1 undisclosed)	μ: 38.4 σ^2: 10.42 R: 23–52	μ: 16.40 σ^2: 1.96 R: 13–18
Out-of-sample			
Sz/Sza	N = 43 (18 F, 24 M, 1 undisclosed)	μ: 50.26 σ^2: 10.83 R: 23–78	μ: 13.73 σ^2: 2.76 R: 8–18
BD	N = 42 (18 F, 24 M)	μ: 50.57 σ^2: 11.83 R: 21–75	μ: 16.29 σ^2: 1.78 R: 12–20
Control	N = 11 (8 F, 3 undisclosed)	μ: 43.63 σ^2: 10.90 R: 24–57	μ: 16.75 σ^2: 1.49 R: 14–18

Note: Not all information was available for each participant; therefore, we report the sample size for each variable separately.

each of the seven categories, we applied principal component analysis (PCA)[30] to produce composite features that contain most of the information.

We began with a set of 43 low-level features that spanned the seven domains. The number of principal components (PCs) used to represent each domain was chosen such that they contain at least 85% of the variance of all variables within that domain. As a result, we obtained 2 PCs for volition, 4 PCs for affect, 2 PCs for lexical diversity, 2 PCs for lexical density, 1 PC for syntactic complexity, 6 PCs for semantic similarity, and 4 PCs for appropriateness of response (a total of 21 features). These were provided to the model designer along with the raw count of word tokens (W) to use for model development. Recent analysis has shown that many computational measures for assessment in psychosis are highly correlated with the number of words spoken.[25] Therefore, we controlled for W in each of the models and assessed the additional value provided by the more complex language measures.

This feature set was used to develop several downstream and upstream prediction models: The models are:

- Prediction of average SSPA score.
- Prediction of functional competency SLOF scores (3 subscales and overall functional competency).
- Prediction of the neurocognitive composite score.
- Prediction of symptom ratings on the PANSS scale (positive symptoms mean and negative symptoms mean).
- Diagnostic group classification.

For all predictive analyses, linear regression models were developed and optimized using leave-one-out cross-validation on only the training samples; the best performing model was selected, fixed, and subsequently evaluated on the test samples. Note that healthy control target scores are only available for the SSPA prediction model. For all other analyses, we only considered the Sz/Sza and BD samples.

Results

A table of descriptive statistics for all relevant outcome measures is shown in the top part of table 2. The cross-validation results on the development set and the out-of-sample results on the test sets are reported separately..

A summary of the cross-validation (for model development) and holdout test performance for SSPA, SLOF, PANSS, and neurocognition regression models is shown in figure 3 and table 3. A summary of the diagnostic classification models (*Clinical vs. Control* and *BPD vs. Sz/Sza*) is shown in figure 4 and table 4.

For all regression models in figure 3, we show a visual representation of model performance by plotting the *Predicted vs. True* values. In table 3, we also provide the Pearson correlation coefficient (PCC) and the mean square error (MSE) between the predicted and true values, computed using cross-validation and on the out-of-sample test set.

Similarly, for the classification models, we show a visual representation of model performance on the out-of-sample data. In addition, we describe model performance with average precision, recall, the $F1$-score for correctly predicting each class (weighted by the support of that class), and the area-under-curve (AUC) for the receiver operating characteristic (ROC) curve to evaluate the performance of the classifier. It is clear in figure 4a that the clinical and healthy control classification model performs very well even on new unseen transcripts, but that the Sz/Sza vs. BD classification problem in figure 4b is more difficult. Still, the tabulated results across both tables 3 and 4 demonstrate that the learned models perform similarly well using cross-validation as they do on the holdout data.

In the supplementary material, we provide extended data analyses for interpretation of the relationship between the composites and the neuropsychological variables. In that analysis, we constrain our composite definitions to one dimension so that we can evaluate the directionality of the features relative to the neuropsychological variables we predict. For all models in both sets of analyses we controlled for raw word count (W) to interpret the additional predictive value of the PCs associated with these features.

Discussion

While there is converging evidence that speech and language analytics can play an important role in computational psychiatry, a well-accepted measurement model for clinical speech applications has yet to be defined. The approach adopted in most studies has been inclusion of large numbers of low-level features into machine learning models; however, recent evidence suggests that low-level features are highly variable and that models built with such features have poor external validity.[25,33,34] This is not surprising given that training robust machine learning models requires massive data sets, whereas in psychiatry research the sample sizes are relatively small. In fact, recent work has shown that the ML paradigm can result in overoptimistic models, especially in digital health applications where clinical data is sparsely available and problem complexity is high (eg, large numbers of features are needed)[38] That is, even when best practices are followed, the models exhibit seemingly good performance during model development and testing but exhibit drastically reduced performance when deployed. The reduction in performance can be explained by “blind spots” in the training data. If there are regions in feature space that are not well-represented during training (ie, blind spots), we never observe the model’s performance with data from those regions. If data from those regions are encountered post deployment, we will not know how the model will perform.

One way to mitigate the impact of blind spots is to use theory to guide model development (eg, see Section titled

Table 2. The Top Half of This Table Shows Participant Statistics for Downstream Assessments of Social and Functional Competency. Healthy Control Participants Were Only Evaluated on the SSPA Task. The Bottom Half of the Table Shows Statistics for Clinical Upstream Assessments of Neurocognition and Symptom Ratings

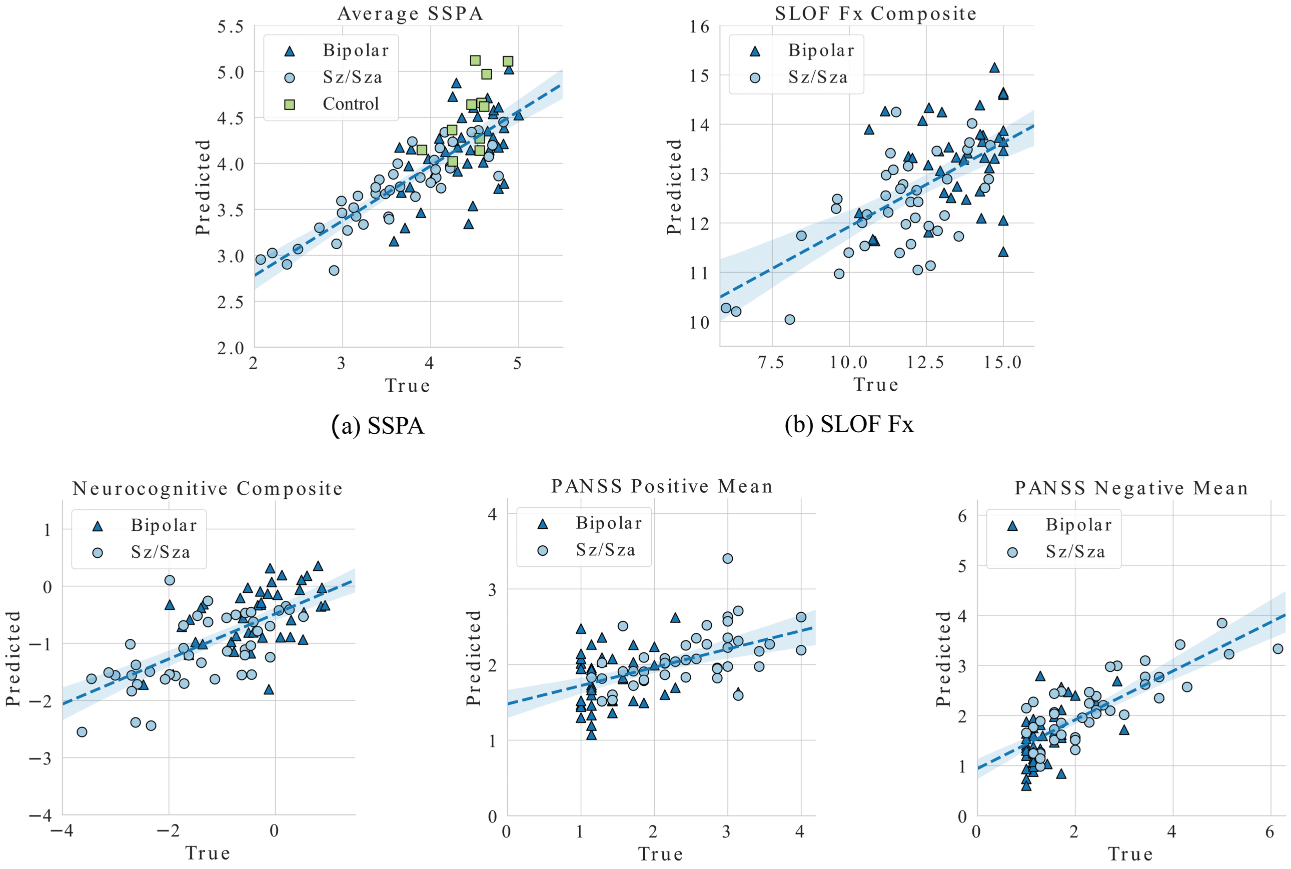
		Sz/Sza		BD		Control	
		Training	Out-of-Sample	Training	Out-of-Sample	Training	Out-of-Sample
Downstream							
SSPA Avg.	*n*	97	43	98	42	11	11
	μ	3.79	3.61	4.42	4.37	4.48	4.47
	σ^2	0.73	0.70	0.39	0.40	0.24	0.26
	R	1.11–5.00	2.07–4.83	3.58–5.00	3.58–5.00	4.14–4.88	3.90–4.88
SLOF							
Interpersonal	*n*	98	42	97	42	—	—
	μ	3.93	3.95	4.49	4.44	—	—
	σ^2	0.85	0.94	0.67	0.66	—	—
	R	1.57–5.00	1.29–5.00	2.14–5.00	2.57–5.00	—	—
Activities	*n*	96	42	97	42	—	—
	μ	4.44	4.31	4.82	4.82	—	—
	σ^2	0.64	0.64	0.28	0.30	—	—
	R	1.73–5.00	2.55–5.00	3.45–5.00	3.50–5.00	—	—
Work	*n*	93	41	98	41	—	—
	μ	3.56	3.34	4.37	4.26	—	—
	σ^2	1.03	0.87	0.84	0.82	—	—
	R	1.40-5.00	1.83–5.00	1.67–5.00	2.33–5.00	—	—
Fx composite	*n*	92	41	97	41	—	—
	μ	11.94	11.57	13.67	13.50	—	—
	σ^2	2.04	1.98	1.55	1.38	—	—
	R	5.68–15.00	6.00–15.00	8.55–15.00	10.32–15.00	—	—
Upstream							
Neurocog. composite	*n*	97	43	98	42	—	—
	μ	−1.13	–1.43	–0.34	–0.42	—	—
	σ^2	1.00	1.10	0.85	0.87	—	—
	R	(−3.27) to (+0.76)	(−3.63) to (+0.53)	(−2.40) to (+1.41)	(−2.48) to (+0.93)	—	—
PANSS							
Pos. symptoms mean	*n*	98	43	98	42	—	—
	μ	2.27	2.44	1.53	1.40	—	—
	σ^2	0.87	0.76	0.59	0.47	—	—
	R	1.00–4.85	1.14–4.00	1.00–3.29	1.00–3.14	—	—
Neg. symptoms mean	*n*	98	43	98	42	—	—
	μ	2.38	2.41	1.26	1.39	—	—
	σ^2	1.12	1.22	0.37	0.48	—	—
	R	1.00–5.86	1.00–6.14	1.00–2.43	1.00–3.00	—	—

Note: SSPA, Social Skills Performance Assessment; Sz/Sza, schizophrenia/schizoaffective; BD, bipolar disorder; SLOF, Specific Level of Functioning; PANSS, Positive and Negative Syndrome Scale. Not all information was available for each participant; therefore, we report the sample size for each variable separately.

"Feature Engineering" in[38]). Reducing feature dimension using a theory-guided approach reduces the likelihood that features are coincidentally correlated with the outcome. Further, it allows for individual interrogation and optimization of features in a clinically-interpretable way. In this paper we used the Levelt model, focusing on the stages of Conceptualization and Formulation (see figure 2), as a guide to operationally define domains of speech production for inclusion as features into predictive models. We used a variety of NLP techniques combined with principal components analysis for feature denoising to compute several composites that characterize each of these domains and are known to be impaired in schizophrenia and BD, including volition, affect, semantic coherence and appropriateness of thought, lexical diversity, lexical density, and syntactic complexity. Since these constructs are complex and multidimensional, the composites themselves are multidimensional. In the discussion that follows, we provide an overview of the relationship between the language domain composites and the measures of mental health status and social and functional capacity.

Volition

By controlling for the raw word count in the model, we introduce collinearity between the volition variables and the control variable (since the word count is also used in the volition composite). This makes it difficult to isolate the contribution of volition to the models individually. Nevertheless, we found that the PCs associated with volition



(a) SSPA

(b) SLOF Fx

(c) Neurocognitive

(d) PANSS +

(e) PANSS -

Fig. 3. Visual representations of the regression prediction results for the (a) Average SSPA score, (b) SLOF Fx, (c) neurocognitive composite score, (d) PANSS positive symptoms average, and (e) PANSS negative symptoms average. Detailed statistics can be found in Table 3.

were particularly useful in our assessment of PANSS positive symptoms. Positive symptoms are associated with overactive expression (eg, hallucinations, delusions) and would therefore be directly impacted by increased or decreased volition.[3] Previous work has confirmed that the impact on social competency outcomes can be mediated by positive symptom severity.[2] Consistent with these findings, we found that the PCs associated with volition were also useful in predicting our SSPA score outcomes. It is surprising that volition did not appear as important for PANSS negative symptoms; however, this is likely due to the collinearity issue mentioned above. The raw word count captures much of the variation in that predictive model and renders the volition composite irrelevant.

Affect

Schizophrenia and BD have overlapping features, including cognitive and mood symptoms. Individuals with schizophrenia often exhibit poverty of speech and reduced affective experience and expression. In BD, individuals experience a wide range of emotions and moods. During a manic episode they may be unusually upbeat and even exhibit euphoria, and during a depressive episode they may express extreme sadness, hopelessness, worthlessness, or guilt. All these emotional expressions contrast with what we expect with healthy individuals.

The SSPA task itself is not ideal for a natural expression of emotions, as the participants are required to perform a specific exercise in which they are role-playing for a short amount of time. Still, we did expect to see differences in emotional processing for individuals in each group based on these measures, as the two scored scenes (new neighbor and landlord conversations) are intended to contain very different emotional content.

Emotional processing is often thought of separately from cognition, but several researchers have argued that they are more directly linked in both BD and schizophrenia. In BD, neurocognitive and emotional deficits are known to have impacts on downstream social and functional outcomes and are closely linked.[39] For schizophrenia, Anticevic and Corlett[40] argue that cognition plays a critical role in the maintenance of emotional information, and it is thought that neurocognitive deficits are partially responsible for

Table 3. Results of regression prediction models. Note that the SSPA Average results are reported separately using all samples with superscript (a) and for only clinical samples with superscript (b). All other predictions are only available for clinical (Sz/Sza + BD) samples. We report the Pearson correlation coefficient (PCC) and the mean squared error (MSE).

Regression		PCC	MSE
Downstream			
SSPA Avg.	Cross-validation	0.787[a]	0.178[a]
		0.791[b]	0.185[b]
	Out-of-sample	0.785[a]	0.171[a]
		p <0.0001[a]	0.182[b]
		0.789[b]	
		p <0.0001[b]	
SLOF			
Interpersonal	Cross-validation	0.473	0.511
	Out-of-sample	0.569	0.493
		p <0.0001	
Activities	Cross-validation	0.647	0.160
	Out-of-sample	0.572	0.211
		p <0.0001	
Work	Cross-validation	0.535	0.734
	Out-of-sample	0.351	0.830
		p <0.01	
Fx Composite	Cross-validation	0.608	2.507
	Out-of-sample	0.616	2.422
		p <0.0001	
Upstream			
Neurocognitive Composite	Cross-validation	0.621	0.623
	Out-of-sample	0.674	0.682
		p <0.0001	
PANSS			
Positive Symptoms Mean	Cross-validation	0.497	0.515
	Out-of-sample	0.509	0.492
		p <0.0001	
Negative Symptoms Mean	Cross-validation	0.718	0.487
	Out-of-sample	0.767	0.476
		p <0.0001	

[a]: *clinical + healthy control*, [b]: *clinical only (Sz/Sza + BD)*

the emotional disassociation that affected individuals exhibit. To this end, we found that affect played an important role in computing the neurocognitive composite score for our upstream regression model. Similarly, we found that the downstream impacts of affect are also apparent in our model predicting overall SSPA score. As argued by Bowie et al.,[2] social competency (measured by the SSPA) is directly correlated with neurocognitive measures.

Semantic Coherence & Appropriateness of Response

Semantically incoherent speech is observed as a common occurrence for many individuals with schizophrenia (associated with Formal Thought Disorder), and it is occasionally observed for those with BD.[3] Disorganized and incoherent speech has been previously cited as an early predictor of an oncoming psychotic episode[16,17] and as a useful feature for classifying between healthy controls and those with psychosis.[21,24] In our work, the language samples that were collected for the SSPA task are conversational in nature. This allows us to take a closer look at the semantic coherence of provided responses and the appropriateness of those responses in context.

Indeed, our models revealed that features from the appropriateness and semantic coherence domains were useful in separating impaired individuals (Sz/Sza or BD) from healthy control participants. We also found that semantic coherence was especially useful in discriminating between individuals with Sz/Sza and BD, in line with results from previous work.[21] Positive symptom severity is also known to differ between those with each condition,[41] which is evident from our previous observation that positive symptom severity can be predicted with these variables.

For the downstream models, appropriateness was important for predicting the overall average SSPA and SLOF-Activities scores; semantic coherence variables were significant in the prediction of the other two SLOF subscales (work skills and interpersonal relationships) as well as the overall SLOF functional score (SLOF-Fx). In

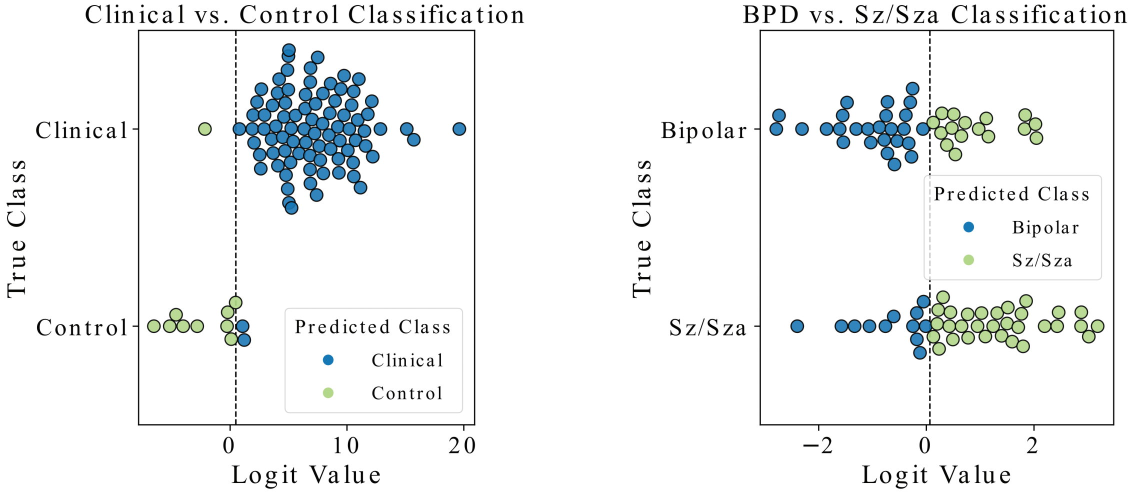


Fig. 4. Visual representation of the two classification prediction experiments conducted in this study. The first plot shows the out-of-sample test set results for the clinical vs healthy control classification, and the second plot shows the same for the BPD vs Sz/Sza classification. Associated statistics are reported in Table 4.

Table 4. Diagnostic group classification results with confusion matrices for the Clinical vs Control classification model and the BD vs. Sz/Sza classification model. Note: The precision and recall statistics reported here are a weighted average of the precision and recall for each class in the binary classification problem to account for class imbalance.

Clinical vs. Control					
Cross Validation			*Out-of-Sample*		
	Clin. Predicted	*Cont. Predicted*		*Clin. Predicted*	*Cont. Predicted*
Clin. True	193	3	*Clin. True*	84	1
Cont. True	3	8	*Cont. True*	2	9
Precision = 0.971, Recall = 0.971 F1 = 0.971, AUC = 0.856			Precision = 0.968, Recall = 0.969 F1 = 0.968, AUC = 0.903		
BPD vs Sz/Sza					
Cross Validation			*Out-of-Sample*		
	BPD Predicted	*Sz/Sza Predicted*		*BPD Predicted*	*Sz/Sza Predicted*
BPD True	74	24	*BPD True*	26	16
Sz/Sza True	29	69	*Sz/Sza True*	12	31
Precision = 0.730, Recall = 0.730 F1 = 0.729, AUC = 0.730			Precision = 0.672, Recall = 0.671 F1 = 0.670, AUC = 0.670		

the case of the SSPA score prediction, appropriateness of response measures were arguably among the most important features used in the regression model. In previous work, Bowie et al.[2] examined how adaptive and social competency measures (such as those measured by the SSPA) are good predictors of downstream functional assessments measured by the SLOF scale. Therefore, it is reasonable to expect to find that these same features may play a role in determining functional competency outcomes in our models predicting the SLOF subscale assessments. Bowie et al. found that the interpersonal relationships and work skills SLOF subscales showed direct correlation with social competency measures from the SSPA scale. They also used a separate set of adaptive competency measures and showed their strong relationship to the SLOF-Activities subscale; the adaptive competency test consists of the *UCSD Performance-based Skills Assessment* (UPSA-B)[42] that evaluates several functional skills in communication and financial literacy. They found that there was no clear relationship between the SSPA and the activities subscale for SLOF. However, in our work, we found that our measures of appropriateness of response in a social context were important components of the SLOF-Activities subscale prediction. From previous work,[35,36] we see that there is a strong negative correlation between PANSS positive symptoms and performance on the UPSA-B evaluation. Since appropriateness was useful in measuring the severity of positive symptoms, we posit that variables from this domain serve as a proxy for positive symptom severity in predicting SLOF-Activities outcomes.

Lexical Diversity

Some declines in lexical diversity (ie, unique vocabulary usage) are observed at late stages of aging,[43] but are more significantly impacted when cognitive deficits are present. In many previous studies, lexical diversity has been a direct indicator of a decline in cognitive ability for those with dementia,[44,45] chronic traumatic encephalopathy (CTE),[46] and our previous work on Sz/Sza and BD.[21] In this study, we provide further evidence on the importance of lexical diversity as the variables associated with this domain were a critical component of our model predicting the neurocognitive composite score.

The neurocognitive deficits that are measured by these variables have a known impact on downstream outcomes. As we noticed with appropriateness of response and semantic coherence, we saw a significant correlation between lexical diversity measures and downstream SSPA task performance. It is possible that this impact on the social competency outcomes is mediated by the positive symptom severity measured by the PANSS scale, which is also correlated with lexical diversity in our models. This is consistent with previous work by Bowie et al.[2] and our previous work with SSPA language samples.[21]

Feature Domains Not Used in Our Prediction Models

The final two feature domains, *lexical density* and *syntactic complexity* were not subsequently chosen by the independent algorithm designer for any of our predictions. Both were included in the domain set since previous work demonstrated their ability to measure important outcomes for individuals with cognitive and thought disorders.[6,47] Here, we provide some insight as to why these feature domains were not found to be as significant in our study.

Lexical density is defined as a measure of "information packaging" in each utterance. Syntactic complexity assesses the complexity of the sentence structures formed in speech. Previous studies have found such measures to be useful in assessing cognitive deficits associated with mild cognitive impairment (MCI),[48] dementia,[49–51] primary progressive aphasia (PPA),[52] and several others. For this reason, we anticipated these measures to be potentially useful in our study looking at upstream and downstream outcomes in Sz/Sza and BD; however, this was not the case. The conversational nature of our transcripts likely plays a significant role in determining the utility of these variables. Many participant responses in the SSPA elicitation task are quite short in nature (ie, "yes" or "okay") and do not lend themselves well to measures of lexical density or syntactic complexity, which are more insightful with increased verbal output. Most previous work using these variables was performed with language samples that were spoken or written with much more natural verbal output.

Limitations of This Work

There are several limitations associated with this work. First, the speech elicitation task is optimized for assessing social skills and, as a result, is likely not optimal for assessing other upstream or downstream variables. Follow-on work should be conducted with a more diverse set of language samples (eg, some that are cognitively more taxing or that tap into sensorimotor control) such that we can fully understand the potential of using these computational variables in our model development. Similarly, the existing analysis did not use any variables from the articulation stage of speech production as the acoustic signal was unavailable for analysis. Several studies have shown that there is important clinical information that can be measured from speech acoustics[6,53]; as such, future studies should consider these variables in the analyses also. It is likely that acoustic metrics such as speaking rate, prosodic variation, and articulatory precision will further strengthen the models' predictive value.

An additional limitation of this work is that we did not consider whether the BD individuals are in a clinical (manic or depressed) or non-clinical phase. It is likely that BD individuals will appear more like Sz/Sza when they are in a clinical phase and this could explain the overlap between the two classes in figure 4b; however, this requires additional exploration in future prospective studies.

Future work can improve upon the foundation laid here in several ways. First, our control sample was smaller compared to our clinical sample and younger in age. A larger matched sample could improve the robustness of our model design. Second, the language samples were only of SSPA conversations and were not optimized for measurement of features requiring long and complex narratives. Lastly, the measurement model proposed herein provides a step forward in the development of an interpretable set of clinically important features; however additional work remains. Although the clinical groups were matched on age and gender, the Sz/Sza group had fewer years of education. We posit that individual norming of the language features using large-scale corpora from the general population will further help to reduce feature variability, place less emphasis on matching cohorts on demographics, and improve the quality of models built on these features by accounting for age/gender/demographic-related changes to the language domains. Finally, the Levelt model was used as a theoretical guide to combine the low-level features. However, this model also contains feedback components that are not directly observable during data acquisition. As a result, we did not attempt to directly model the feedback component during feature learning in this work. Future work can focus on modeling the feedback component; this would likely require a fundamentally new elicitation task where the input stimulus is actively modified (eg, via perturbation).

Conclusion

Language parameters measured from conversations elicited from the SSPA protocol[35] allowed us to predict several measures of mental health status and social and functional competency. The best model performance was obtained for the regression models that predicted average SSPA performance. This was expected since the SSPA transcripts were the source of our language samples, however, reasonable predictive value also was shown for measures of neurocognition, symptom ratings, and functional competency tasks. In addition, classification of individuals into their respective diagnostic groups was also possible from the SSPA language samples, even though the SSPA is not intended as a clinical diagnostic tool. Most importantly, on every regression and classification model that was developed, the model performance generalized well to transcripts that were never seen during training by an independent biostatistician; this bolsters confidence in the external validity of the models.

Supplementary Material

Supplementary material is available at https://academic.oup.com/schizophreniabulletin/.

Funding

This work was supported by the National Institute of Health (R01DC006859 to J.M.L. and V.B., R01MH079784 to Ann Pulver), the Canadian Institutes of Health Research (Operating Grant 343639 to C.R.B.), and a gift from the Boehringer Ingelheim Foundation (to V.B. and J.M.L.).

Acknowledgments

The authors would like to acknowledge the psychologists and research assistants responsible for obtaining and transcribing the conversations used in our study, as well as all study participants.

Conflicts of Interest

V.B. and J.M.L. are co-founders and have equity in Aural Analytics. C.R.B. has received grant support in the past five years from Pfizer, Lundbeck, and Takeda; royalties from Oxford University Press; and in-kind research user licenses from Scientific Brain Training. Harvey has received consulting fees or travel reimbursements from Alkermes, Bio Excel, Boehringer Ingelheim, Karuna Pharma, Merck Pharma, Minerva Pharma, SK Pharma, and Sunovion Pharma during the past year. He receives royalties from the Brief Assessment of Cognition in Schizophrenia (Owned by WCG Verasci, Inc. and contained in the MCCB). He is the chief scientific officer of i-Function, Inc. The remaining authors declare no conflicts of interest.

References

1. Green MF. What are the functional consequences of neurocognitive deficits in schizophrenia? *AJP* 1996;153(3):321–330. doi:10.1176/ajp.153.3.321.
2. Bowie CR, Depp C, McGrath JA, *et al*. Prediction of real-world functional disability in chronic mental disorders: a comparison of schizophrenia and bipolar disorder. *Am J Psychiat.* 2010;167(9):1116–1124.
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. 5th ed. Washington, D.C: American Psychiatric Publishing; 2013.
4. Weiner L, Doignon-Camus N, Bertschy G, Giersch A. Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures. *Sci Rep.* 2019;9(1):14282. doi:10.1038/s41598-019-50818-5.
5. Hoffman RE, Stopek S, Andreasen NC. A comparative study of manic vs schizophrenic speech disorganization. *Arch Gen Psychiat.*. 1986;43(9):831–838. doi:10.1001/archpsyc.1986.01800090017003.
6. Voleti R, Liss JM, Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J Sel Top Signal Process.* 2020;14(2):282–298. doi:10.1109/JSTSP.2019.2952087.
7. Cecchi GA, Gurev V, Heisig SJ, Norel R, Rish I, Schrecke SR. Computing the structure of language for neuropsychiatric evaluation. *IBM J Res Dev.* 2017;61(2/3):1:1–1:10. doi:10.1147/JRD.2017.2648478.
8. Corcoran CM, Cecchi G. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biol Psychiatry: Cogn Neurosci Neuroimaging*. Published online June 2020;5:770–779. doi:10.1016/j.bpsc.2020.06.004
9. Raugh IM, James SH, Gonzalez CM, *et al*. Digital phenotyping adherence, feasibility, and tolerability in outpatients with schizophrenia. *J Psychiatr Res.* 2021;138:436–443. doi:10.1016/j.jpsychires.2021.04.022.
10. Cohen AS, Cox CR, Le TP, *et al*. Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia. *npj Schizophr.* 2020;6(1):26. doi:10.1038/s41537-020-00115-2.
11. Palaniyappan L, Mota NB, Oowise S, *et al*. Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog Neuropsychopharmacol Biol Psychiat.* 2019;88:112–120. doi:10.1016/j.pnpbp.2018.07.007.
12. de Boer JN, van Hoogdalem M, Mandl RCW, *et al*. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *npj Schizophr.* 2020;6(1):10. doi:10.1038/s41537-020-0099-3.
13. Bowie CR, Gupta M, Holshausen K. Disconnected and underproductive speech in schizophrenia: Unique

relationships across multiple indicators of social functioning. *Schizophr Res.* 2011;131(1-3):152–156. doi:10.1016/j.schres.2011.04.014.

14. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1-3):304–316. doi:10.1016/j.schres.2007.03.001.
15. Landauer TK, Dumais ST. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol Rev*. 1997;104(2):211–240. doi:10.1037/0033-295X.104.2.211
16. Bedi G, Carrillo F, Cecchi GA, *et al*. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* 2015;1:15030.
17. Corcoran CM, Carrillo F, Fernández-Slezak D, *et al*. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiat.* 2018;17(1):67–75. doi:10.1002/wps.20491.
18. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Poster presented at: 1st International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May, 2013.
19. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Moschitti A, Pang B, Daelemans W, eds. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October, 2014. Association for Computational Linguistics; 2014:1532–1543. doi:10.3115/v1/D14-1162.
20. Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing schizophrenia. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LA, 5 June, 2018; 2018:136–146. https://aclanthology.org/W18-0615/
21. Voleti R, Woolridge S, Liss JM, Milanovic M, Bowie CR, Berisha V. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. *Proc. Interspeech* 2019;2019:1433–1437. doi:10.21437/Interspeech.2019-2960.
22. Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* 2019;5(1):9. doi:10.1038/s41537-019-0077-9.
23. Mota NB, Vasconcelos NAP, Lemos N, *et al*. Speech graphs provide a quantitative measure of thought disorder in psychosis. Solé RV ed. *PLoS One.* 2012;7(4):e34928. doi:10.1371/journal.pone.0034928.
24. Kayi ES, Diab M, Pauselli L, Compton M, Coppersmith G. Predictive linguistic features of schizophrenia. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017), Vancouver, BC, Canada, 3–4 August, 2017; 2017:241–250. https://aclanthology.org/S17-1028/
25. Hitczenko K, Cowan H, Mittal V, Goldrick M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, Mexico City, Mexico, 11 June, 2021. Association for Computational Linguistics; 2021:129–150. https://aclanthology.org/2021.clpsych-1.16
26. Firth J, Torous J Smartphone apps for schizophrenia: a systematic review. *JMIR mHealth uHealth*. 2015;3(4):e102. doi:10.2196/mhealth.4930
27. Levelt WJM. Models of word production. *Trends Cogn Sci.* 1999;3(6):223–232. doi:10.1016/S1364-6613(99)01319-4.
28. Levelt WJM. Producing spoken language: a blueprint of the speaker. In: Brown CM, Hagoort P, eds. *The Neurocognition of Language*. Oxford University Press, USA; 1999:83–122.
29. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 2015;71:10–49. doi:10.1016/j.specom.2015.03.004.
30. Pearson K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin Philos Mag J Sci.* 1901;2(11):559–572. doi:10.1080/14786440109462720.
31. Deledalle CA, Salmon J, Dalalyan A. Image denoising with patch based PCA: local versus global. In: Proceedings of the British Machine Vision Conference 2011, Dundee, Scotland, UK, August 29–September 2, 2011. British Machine Vision Association; 2011:25.1–25.10. http://www.bmva.org/bmvc/2011/proceedings/paper25/index.html
32. Takiguchi T, Ariki Y. PCA-Based speech enhancement for distorted speech recognition. *J Multimed.* 2007;2(5):13–18.
33. Stegmann GM, Hahn S, Liss J, *et al*. Repeatability of commonly used speech and language features for clinical applications. *Digit Biomark* 2020;4(3):109–122. doi:10.1159/000511671.
34. Rusz J, Švihlík J, Krýže P, Novotný M, Tykalová T. Reproducibility of voice analysis with machine learning. *Mov Disord.* 2021;36(5):1282–1283. doi:10.1002/mds.28604.
35. Patterson TL, Moscona S, McKibbin CL, Davidson K, Jeste DV. Social skills performance assessment among older patients with schizophrenia. *Schizophr Res.* 2001;48(2-3):351–360. doi:10.1016/S0920-9964(00)00109-2.
36. Schneider LC, Struening EL. SLOF: a behavioral rating scale for assessing the mentally ill. *Soc Work Res Abstr.* 1983;19(3):9–21. doi:10.1093/swra/19.3.9.
37. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull.* 1987;13(2):261–276. doi:10.1093/schbul/13.2.261.
38. Berisha V, Krantsevich C, Hahn PR, *et al*. Digital medicine and the curse of dimensionality. *NPJ Digit Med.* 2021;4(1):153. doi:10.1038/s41746-021-00521-5.
39. Lima IMM, Peckham AD, Johnson SL. Cognitive deficits in bipolar disorders: Implications for emotion. *Clin Psychol Rev.* 2018;59:126–136. doi:10.1016/j.cpr.2017.11.006.
40. Anticevic A, Corlett PR. Cognition-emotion dysinteraction in schizophrenia. *Front Psychol.* 2012;3:392. doi:10.3389/fpsyg.2012.00392.
41. Mancuso SG, Morgan VA, Mitchell PB, Berk M, Young A, Castle DJ. A comparison of schizophrenia, schizoaffective disorder, and bipolar disorder: results from the Second Australian national psychosis survey. *J Affect Disord.* 2015;172:30–37. doi:10.1016/j.jad.2014.09.035.
42. Mausbach BT, Harvey PD, Pulver AE, *et al*. Relationship of the Brief UCSD Performance-based Skills Assessment (UPSA-B) to multiple indicators of functioning in people with schizophrenia and bipolar disorder. *Bipolar Disord.* 2010;12(1):45–55. doi:10.1111/j.1399-5618.2009.00787.x.
43. Wright HH, ed. *Cognition, Language and Aging*. Amsterdam, The Netherlands: John Benjamins Publishing Company; 2016.
44. Snowdon DA, Kemper SJ, Mortimer JA, Greiner LH, Wekstein DR, Markesbery WR. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: findings from the nun study. *JAMA* 1996;275(7):528–532. doi:10.1001/jama.1996.03530310034029.
45. Berisha V, Wang S, LaCross A, Liss J. Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case

study comparing the press conferences of presidents Ronald Reagan and George Herbert Walker Bush. *J Alzheimer's Dis.* 2015;45(3):959–963. doi:10.3233/JAD-142763.

46. Berisha V, Wang S, LaCross A, Liss J, Garcia-Filion P. Longitudinal changes in linguistic complexity among professional football players. *Brain Lang.* 2017;169:57–63. doi:10.1016/j.bandl.2017.02.003.
47. Çokal D, Sevilla G, Jones WS, *et al*. The language profile of formal thought disorder. *NPJ Schizophr.* 2018;4(1):18. doi:10.1038/s41537-018-0061-9.
48. Roark B, Mitchell M, Hosom JP, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans Audio Speech Lang Process.* 2011;19(7):2081–2090. doi:10.1109/TASL.2011.2112351.
49. Bucks RS, Singh S, Cuerden JM, Wilcock GK. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: evaluation of an objective technique for analysing lexical performance. *Aphasiology* 2000;14(1):71–91. doi:10.1080/026870300401603.
50. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimer's Dis.* 2015;49(2):407–422. doi:10.3233/JAD-150520.
51. Bertola L, Mota NB, Copelli M, *et al*. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front Aging Neurosci.* 2014;6:185:1–10. doi:10.3389/fnagi.2014.00185.
52. Fraser KC, Meltzer JA, Graham NL, *et al*. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex.* 2012;55:43–60. doi:10.1016/j.cortex.2012.12.006.
53. Stegmann GM, Hahn S, Liss J, *et al*. Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *NPJ Digit Med* 2020;3(1):132. doi:10.1038/s41746-020-00335-x.