Computational Analysis of Spoken Language in Acute Psychosis and Mania

Jeffrey M. Girard[a,1], Alexandria K. Vail[b,1], Einat Liebenthal[c,d], Katrina Brown[c], Can Misel Kilciksiz[c], Luciana Pennant[c], Elizabeth Liebson[c,d], Dost Öngür[c,d], Louis-Philippe Morency[e], Justin T. Baker[c,d,*]

[a]*Department of Psychology, University of Kansas, Lawrence, Kansas, USA*
[b]*Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*
[c]*Division of Psychotic Disorders, McLean Hospital, Belmont, Massachusetts, USA*
[d]*Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA*
[e]*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*

Abstract

Objectives: This study aimed to (1) determine the feasibility of collecting behavioral data from participants hospitalized with acute psychosis and (2) begin to evaluate the clinical information that can be computationally derived from such data. **Methods:** Behavioral data was collected across 99 sessions from 38 participants recruited from an inpatient psychiatric unit. Each session started with a semi-structured interview modeled on a typical "clinical rounds" encounter and included administration of the Positive and Negative Syndrome Scale (PANSS). **Analysis:** We quantified aspects of participants' verbal behavior during the interview using lexical, coherence, and disfluency features. We then used two complementary approaches to explore our second objective. The first approach used predictive models to estimate participants' PANSS scores from their language features. Our second approach used inferential models to quantify the relationships between individual language features and symptom measures. **Results:** Our predictive models showed promise but lacked sufficient data to achieve clinically useful accuracy. Our inferential models identified statistically significant relationships between numerous language features and symptom domains. **Conclusion:** Our interview recording procedures were well-tolerated and produced adequate data for transcription and analysis. The results of our inferential modeling suggest that automatic measurements of expressive language contain signals highly relevant to the assessment of psychosis. These findings establish the potential of measuring language during a clinical interview in a naturalistic setting and generate specific hypotheses that can be tested in future studies. This, in turn, will lead to more accurate modeling and better understanding of the relationships between expressive language and psychosis.

Keywords: language, schizophrenia, bipolar disorder, positive symptoms, negative symptoms

1. Introduction

Mental health clinicians are trained to detect nuances in patient behavior and thought processes considered to index critical diagnostic and prognostic features of illness. However, a mental healthcare system that relies primarily on expert human judgment is bound to be costly, difficult to monitor for efficacy, and prone to the influence of subjective biases (which may contribute to disparities and inequalities in care (e.g., Neighbors et al., 2003; Singhal et al., 2016). Despite considerable advances in identifying genetic, molecular, and neuroimaging biomarkers of mental illness (e.g., Holmes and Patrick, 2018; Lakhan and Kramer, 2009; Prata et al., 2014; Saykin et al., 2010), quantitative means for tracking behavior and predicting the course of illness that are equivalent or superior to subjective methods are still unavailable. For real utility

*Corresponding author.
Email address: jtbaker@partners.org (Justin T. Baker)
[1]Equal contribution.

in clinical settings, there is an urgent need for quantitative benchmarks of illness that afford all the computational opportunities assigning a reliable numeric system to illness states can provide (even if the scientific understanding of their biological underpinnings is still incomplete).

In this study, we sought to quantify the verbal behavior of patients in a psychiatric inpatient unit during brief clinical encounters. The dyadic clinical encounter was designed to simulate an efficient mental status exam conducted during clinical rounds in a psychiatry unit. The study's primary objective was to determine how quantitative measures of expressive language can predict the clinician's assessment of symptom severity in patients with a psychotic disorder. The psychiatric inpatient unit constitutes an optimal environment for this type of experiment in that many patients present with acute symptoms that fluctuate throughout the hospitalization, and the clinical encounter occurs in a naturalistic yet controlled setting.

An individual's comprehension and expression of ideas and thoughts via language communication reflect to a great extent their cognitive, mental, and emotional state. In individuals with a psychotic disorder, language comprehension and expression may reflect aspects of both *positive symptoms* (e.g., delusions, conceptual disorganization, hallucinations, feelings of grandiosity, and feelings of excitement and hostility) and *negative symptoms* (e.g., emotional withdrawal, difficulty in abstract thinking, and disfluency of speech). Impairment in language expression, termed clinically as "thought disorder," has long been recognized as a hallmark of psychotic disorders (Andreasen and Grove, 1986). Thought disorder may present as positive symptoms of disorganized language (including derailment, tangentiality, incoherence, distraction, abnormal use of referential markers, and illogicality) and negative symptoms consisting of poverty of language (including lexicon, content, and syntax) and disfluent speech (including flat affect and pausing). Disorganized language has been found to correlate with other positive symptoms of psychosis such as delusions (Andreasen and Olsen, 1982; Docherty et al., 2003; Harrow and Marengo, 1986), whereas poverty of language lexicon and syntax has been associated with other negative symptoms and semantic processing abnormalities (Gooding et al., 2013; Nagels et al., 2016). The positive aspects of thought disorder have been associated with psychosis across clinical diagnoses, whereas the negative aspects of thought disorder have been associated more specifically with schizophrenia (Andreasen, 1979; Gooding et al., 2013; Yalincetin et al., 2017).

The potential value of quantitative analysis of natural spoken language has become increasingly apparent in recent years, with progress in developing automated natural language processing and machine learning methods (Hitczenko et al., 2020). Most quantitative studies of spoken language in psychotic disorders to date have focused on identifying differences from healthy controls or other types of mental disorders at the group level. A meta-analysis of studies using semantic space models such as latent semantic analysis to represent word and sentence meanings and assess language disorganization mathematically suggested that this class of methods could effectively distinguish between individuals with a psychotic disorder and healthy controls (de Boer et al., 2018). The disorganization scores obtained with latent semantic analysis in some of these studies were correlated with the manual scores (Corcoran et al., 2018; Elvevåg et al., 2007). Metaphor-identification and sentiment-analysis algorithms were used to generate features for a classifier automatically, and these features were shown to effectively distinguish between individuals with first-episode psychosis and healthy controls (Gutiérrez et al., 2017). A meta-analysis of studies of non-verbal vocal expression found that some quantitative measures of speech disfluency (specifically, pauses in speech) could distinguish individuals with schizophrenia from healthy controls (Cohen et al., 2014).

Only a handful of studies have used quantitative analysis of spoken language to index the severity of thought disorder and predict psychosis risk within a group of individuals with a psychotic disorder (Corcoran et al., 2020). In a study of speech transcripts in which different aspects of thought disorder were manually coded, an elevated rate of illogical expressions and poverty of content were shown to predict later transition to psychosis in at-risk adolescents (Bearden et al., 2011). Latent semantic analysis and part-of-speech tagging were used to evaluate discourse coherence and syntactic complexity (Bedi et al., 2015; Corcoran et al., 2018; Elvevåg et al., 2007), and vector unpacking was used to evaluate poverty of language content (Rezaii et al., 2019), to predict psychosis onset among high-risk individuals. Preliminary work from our group further indicated that automated analysis of expressive language during psychiatric interviews is a promising method for quantifying psychopathology *in situ*, on an individual basis, and longitudinally, as required for tracking the trajectory of the disease in clinical settings (Vail et al., 2018).

Here, we characterized language expression during the dyadic encounters using measures of lexical in-

formation, sentence coherence, and spoken disfluency to test for potential associations with positive and negative symptoms of psychotic disorders captured by the clinical rating scales.

2. Material and Methods

The study's primary goal was to identify associations between language measures derived from participant utterances during brief semi-structured interviews and positive and negative symptom scores derived from comprehensive clinical interviews. After describing the study procedures, we detail our analytic plan, which used a multilevel modeling approach to identify between-person differences and changes over repeated assessments within the same individual, using predictive and inferential models.

2.1. Study Procedures

Study participants were adult individuals admitted to inpatient clinical services at McLean Hospital. This population was selected to maximize the likelihood of studying individuals spanning the symptom severity range on one or more symptom inventories and heighten the probability of observing clinically meaningful severity changes within individuals over a brief period (i.e., three or fewer weeks). Hospitalized individuals were eligible if they were (a) under a voluntary legal status and (b) had received a diagnosis (per electronic health record) of either a primary psychotic disorder (i.e., schizophrenia, schizoaffective disorder, or psychotic disorder not otherwise specified) or a psychotic condition secondary to an affective disorder (e.g., bipolar disorder with psychotic features).

All screened individuals were first assessed by their treating physician for the capacity to be approached for the study, after which research staff met with eligible individuals to explain the study and obtain informed consent. Participants answered a short survey to document their understanding of the study, how the data would be used, and whether it would affect their treatment. The survey was designed to protect individuals with profound mental disturbance from study participation in cases where someone may not understand essential information related to potential risks and benefits. The protocol was reviewed and approved by the Institutional Review Board (IRB) of both Partners Healthcare, where the data was collected, and Carnegie Mellon University, where identified samples were shared for analysis. All identifiable data were encrypted at rest and in transit at both sites, and efforts were made to limit the exposure of potentially sensitive information throughout data collection and analysis.

2.1.1. Clinical Assessments

Throughout their hospitalization, enrolled participants were interviewed one or more times by study doctors (MDs), all Board certified psychiatrists, using a semi-structured interview. Each interview consisted of a set of 13 questions (Table 1) selected by the study doctors to capture elements from a mental status examination that psychiatrists perform during typical "clinical rounds" as part of their daily assessment of each hospitalized individual. As compared with the more comprehensive symptom assessments performed in research settings, which can take upwards of 45 minutes per study visit, we reasoned that this brief, semi-structured encounter might contain most or all of the relevant signals necessary to estimate key indices of symptom severity.

After the MD visit, research staff followed the brief encounter with the administration of a more comprehensive assessment of symptom severity, in order to facilitate comparison between conventional "gold-standard" metrics of symptom severity and the information that can be gleaned from our brief, semi-structured encounter. Specifically, participants were interviewed to provide the information needed to score the Positive and Negative Syndrome Scale (PANSS; Kay et al., 1987), a commonly used instrument for quantifying severity across multiple symptom domains experienced by individuals with schizophrenia and other psychotic disorders. Also administered (but not considered here) were the Montgomery-Åsberg Depression Rating Scale (MADRS; Montgomery and Åsberg, 1979), the Young Mania Rating Scale (YMRS; Young et al., 1978), and the Brief Psychiatric Rating Scale (BPRS; Overall and Gorham, 1962). While the published instruments are often framed to gauge symptoms and experiences over fixed intervals (e.g., "over the past two weeks"), we explicitly trained staff to inquire and score these instruments based on symptoms and

Table 1: Questions from the semi-structured clinical interviews

#	Scripted Question
Q01	What brought you into the hospital?
Q02	Has anything been on your mind recently?
Q03	What has the team been helping you with?
Q04	How's the team doing in helping you?
Q05	What are your goals while you are here?
Q06	How have people been treating you?
Q07	How is the food?
Q08	How has your mood been?
Q09	How is your thinking?
Q10	How is your energy?
Q11	How is your sleep?
Q12	How is your self-confidence?
Q13	Are there any changes you have observed in yourself?

experience since the previous evaluation, which could vary depending on their hospital course and pattern of study participation.

PANSS scores were assessed by trained clinical raters who reviewed each comprehensive assessment's responses to estimate the item-level severity scores. Inter-rater agreement was assessed biannually and required to exceed 0.7 on Krippendorff's alpha. Item-level scores are summed to indicate indices of "positive" and "negative" symptom severity, as well as a "general" severity index that was scored but not analyzed here. The PANSS Positive Total ("PANSS Positive") score provides an estimate of the severity of "positive" symptoms, which refer to an excess or distortion of normal functions (i.e., delusions, conceptual disorganization, hallucinations, excitement, grandiosity, suspiciousness, and hostility). The PANSS Negative Total ("PANSS Negative") score provides an estimate of the severity of "negative" symptoms, which represent a restriction or loss of normal functions (i.e., blunted affect, emotional withdrawal, poor rapport, passive social withdrawal, difficulty in abstract thinking, lack of spontaneity and flow of conversation, and stereotyped thinking). The PANSS General Total ("PANSS General") score provides an estimate of the severity of other symptoms (e.g., depression, anxiety, and impulsiveness), which was considered too broad and heterogeneous to test our approach.

2.1.2. Audio-Visual Recordings

We recorded all brief encounters using multiple synchronized acquisition systems to derive high-quality behavioral signals related to symptom severity. High-definition (1080p) webcams (Logitech c920 webcam) were placed on the table between clinician and participant and focused on the face and upper torso to optimize facial landmark detection. Both individuals (participant and clinician) wore a high-quality cardioid headset microphone (Sennheiser HSP4, Germany) to optimize audio quality and separation of the two audio streams from the two participants ("speaker diarization"), which improves transcription accuracy and facilitates individualized voice feature analysis. For a subset of sessions, additional cameras (Microsoft Kinect v1.0) were also positioned on a shelf above each participant to obtain gestural and postural signals. Signals from the two microphones and 2-4 cameras were recorded using a custom software package ("MultiSense Recorder," CMU) that creates synchronized data streams. Here, we consider only linguistic features derived from the diarized audio.

2.1.3. Transcription and Data Processing

Audio recordings were professionally transcribed using a medical transcription service (TranscribeMe, Inc.), which produced timestamped and speaker-annotated transcripts with the redaction of personally identifying information such as names and locations. Before language analysis, transcripts were further

Figure 1: Example of the reparandum-interregnum-repair disfluency structure. Note that this example would be classified as a restart disfluency because the repair term is different from the reparandum.

refined to correct typographical errors, mislabeled speakers, and missed redactions. Utterances spoken by interviewers were not considered for the present analysis.

2.2. Language Measures

Fourteen language measures were selected based on prior literature as plausibly reflecting variance in positive and negative symptoms, which can be grouped into (1) lexical measures, which characterize the frequency of spoken words from specific semantic and functional categories; (2) sentence coherence measures, which characterize the typicality and predictability of participants' spoken word sequences; and (3) spoken disfluency measures, which characterize the frequency of involuntary disruptions in participants' speech.

Lexical measures were computed using the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) dictionary, an established and well-validated lexicon, which organizes over 6,000 common English words into semantically and functionally similar "word categories" such as pronouns, swear words, and emotion words. For each transcript, we measured the percentage of the participant's words that could be found in the following ten LIWC word categories, which were identified *a priori* as potentially relevant to positive or negative symptoms. Two word categories were related to affective processes: *Positive Emotion* words such as "happy" and "love" and *Negative Emotion* words such as "annoyed" and "crying." Three word categories were related to cognitive and sensory processing: *Cognitive Processes* words such as "cause" and "know," *Perceptual Processes* words such as "look" or "heard," and *Relativity* words such as "here" and "until." Finally, five word categories were related to drives: *Affiliation* words such as "friend" and "social," *Achievement* words such as "win" and "success," *Power* words such as "superior" and "bully," *Reward* words such as "prize" and "benefit," and *Risk* words such as "danger" and "doubt."

Local (i.e., sentence-level) coherence of the spoken utterances was calculated using a perplexity measure, which quantifies the participant's word sequences' average predictability within a set of utterances. Perplexity is a measure of entropy or randomness, which implicitly reflects both vocabulary choices and syntactic constructions, and was derived from a language model that tried to predict the participant's next word based on their previous words using a *trigram backoff* model (Seymore and Rosenfeld, 1996). The trigram backoff model first tries to predict the next word using the previous two words. However, if the model has never seen the previous two words together, it will try to predict the next word using the last word. Finally, if the model has never seen that last word, it will default to predicting the most common word overall (ignoring context). Having been trained on a sizable corpus of conversational speech and text from telephone conversations (Godfrey et al., 1992), the language model we used can be viewed as an approximation of ordinary spoken language. Someone experiencing psychotic symptoms might show higher perplexity values (i.e., it might be more difficult for the language model to predict their words) due to aberrations in their vocabulary choices and sentence constructions associated with cognitive, affective, and interpersonal deficits (e.g., conceptual disorganization or sparse content).

Speech disfluency was estimated as the frequency of three types of speech repairs: *edits* (i.e., nonverbal fillers such as "uh" and verbal fillers such as "I mean" or "you know"), *repeats* (i.e., repetitions of words or short phrases), and *restarts* (i.e., mid-sentence changes). Work by Shriberg (1994) defined disfluencies within a tripartite reparandum-interregnum-repair structure. A reparandum is an error in speech that the speaker subsequently corrects, and a repair term is that correction. An interregnum term is a filler token or a cue phrase (e.g., "umm" or "I mean") that occurs between the reparandum and repair terms as a way to stall for time while the speaker generates the repair term. Note, however, that one or more of these terms may be implicit (e.g., a long pause may take the place of an interregnum, and an interregnum may occur without an explicit reparandum or repair). An example of this structure is depicted in Figure 1. Using this

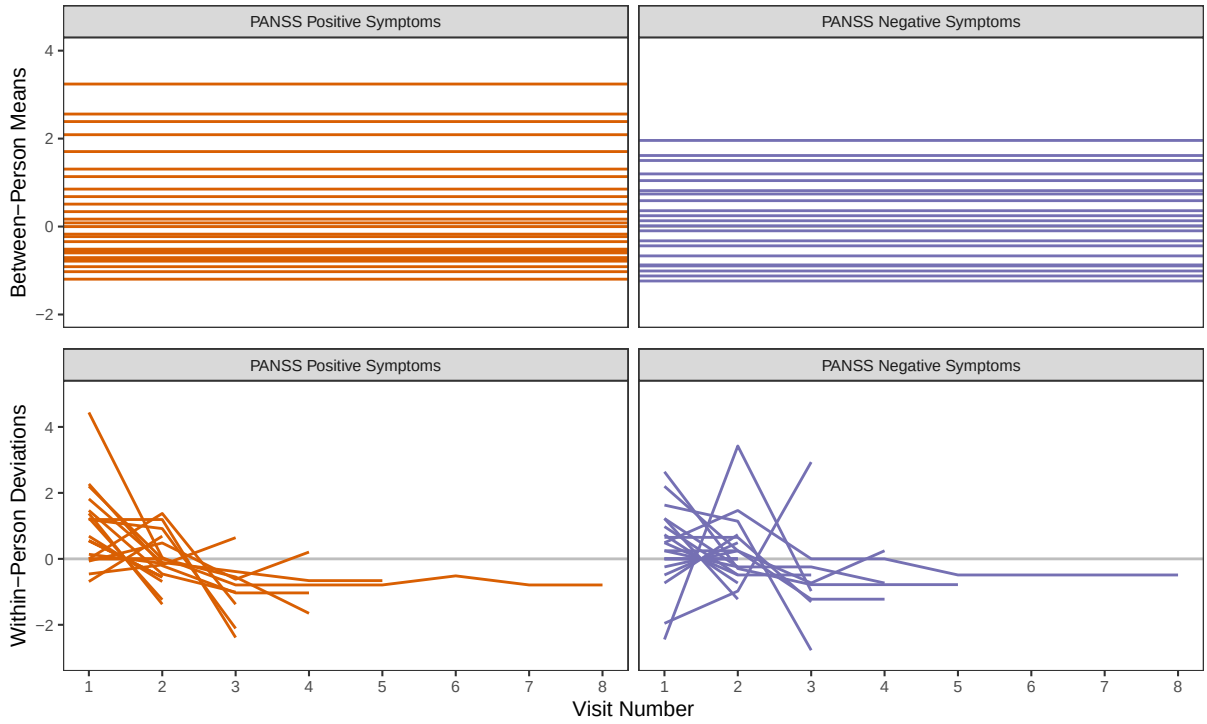


Figure 2: Depiction of the standardized between-person and within-person symptom components (each line is a participant)

structure to illustrate our three examined classes: an *edit* occurs when an interregnum occurs without an explicit reparandum and repair term (e.g., "He... umm... wasn't happy about it"), a *repeat* occurs when the reparandum and repair term are the same (e.g., "I don't... uhh... don't think so"), and a *restart* occurs when the reparandum and repair term are different (e.g., "I loved... I mean... I liked it"). Disfluencies were annotated automatically using an incremental disfluency detection model (Hough and Schlangen, 2017). This language model consists of neural network sequence models that consume incoming words and use word embeddings, part-of-speech tags, and other features to predict disfluency classes for each word in a strictly left-to-right, word-by-word fashion. As in the perplexity calculation, the model was used to tag each utterance within each transcript separately for each disfluency class (edits, repeats, restarts) occurring within each interview. The number of events for each class was then normalized by the total number of words spoken.

2.3. Modeling

Our modeling approach utilized both predictive and inferential methods, which we view as complementary methods to explore different aspects of our research questions. Across both methods, we used a multilevel approach to account for variance between individuals and within individuals over multiple observations, which is both an advantage and challenge for longitudinal, repeated-measures studies. Accounting for a multilevel structure is critical given the potential for dependencies since observations of the same participant are often more similar than observations of different participants, which could bias estimates of model parameters (de Leeuw and Meijer, 2007). Several methods exist for decomposing longitudinal data into separate between-person and within-person components (Hamaker and Muthén, 2019). We calculated each participant's mean score across all observations to estimate each clinical and language measure's between-person components. To estimate each clinical and language measure's within-person components, we subtracted each participant's mean score across all observations from their observed score at each observation (thus restructuring it as that participant's deviation from their mean). Thus, the between-person

components capture whether each participant tended to have lower or higher scores in general, while the within-person components capture whether each observation was particularly low or high for each participant. These components are depicted in Figure 2.

2.3.1. Predictive Modeling

To determine how effectively the clinical measures could be estimated from the selected language measures, we developed separate predictive models for the PANSS Positive and PANSS Negative scores. We first discuss the preprocessing steps we took to prepare the data for predictive modeling and then the techniques we used to train and cross-validate the predictive models.

2.3.2. Predictive Model Preprocessing

Following the decomposition of variables into between- and within-person components, we developed separate models to predict the between-person clinical components from the between-person language components and predict the within-person clinical components from the within-person language components. These models had different sample sizes: the between-person models had one data point per participant ($n = 38$), and the within-person models had one data point per session, but only for those participants who had more than one session ($n = 60$). We will refer here to the language measures as "features" and the clinical measures as "labels."

For all models, we preprocessed the data by dropping the features with near-zero variance, filtering out highly correlated features (i.e., looking for two features with an absolute correlation greater than 0.9 and then removing the feature with the greater mean absolute correlation with the other features), and z-transforming (i.e., centering and scaling) the features. Only a single feature (i.e., between-person restarts) was removed due to its high correlation with another feature (i.e., between-person repeats). The labels were not z-transformed to keep the predictions and performance measures on the same metric as the PANSS scores.

2.3.3. Predictive Model Training and Cross-Validation

Three different predictive modeling algorithms were selected for estimation due to their ability to achieve competitive performance in relatively small samples (Appendix A). First, we used the Elastic Net algorithm (Zou and Hastie, 2005), a penalized form of linear regression that blends the L_1 and L_2 penalties of the lasso and ridge regression methods. This approach contains automatic feature selection (via the penalties) but does not automatically include interaction effects or nonlinear relationships. Second, we used the Support Vector Regression (SVR) algorithm (Drucker et al., 1997), a form of large-margin learning that can capture both interaction effects and nonlinear relationships through a radial basis function kernel but does not include automatic feature selection. Finally, we used the Random Forest algorithm (Breiman, 2001), which combines multiple decision trees into an ensemble to combat overfitting while still performing automatic feature selection and capturing interaction effects and nonlinear relationships. Note that our goal in training multiple algorithms was to maximize our chances of finding one that would perform well in this specific dataset (and not to find which is "best" in an overall sense).

To estimate a predictive model's performance accurately, we must assess its ability to make predictions on data that it has never seen. To accomplish this, we followed the recommendations of Cearns et al. (2019) and used a nested cross-validation procedure to train our models, tune their hyperparameters, and test their performance (Appendix B). Outer cross-validation was used to test the model on unseen data, and inner cross-validation was used to tune the model hyperparameters. Both the inner and the outer procedures took the form of stratified 10-fold cross-validation, and each was repeated three times with different partitionings. Each predictive model was thus evaluated across 30 unseen test sets (i.e., 10 test folds per outer cross-validation times three repetitions). All sessions from any given participant were always assigned to the same fold for the within-person models to prevent data leakage (i.e., training on one session from a participant and then testing on another session from the same).

Predictive performance was tuned and measured using the Root Mean Squared Error (RMSE) metric, which captures the *error* of the model in absolute terms, i.e., the degree to which the model predictions differed from the labels (using the same metric as the label). RMSE squares the errors before averaging

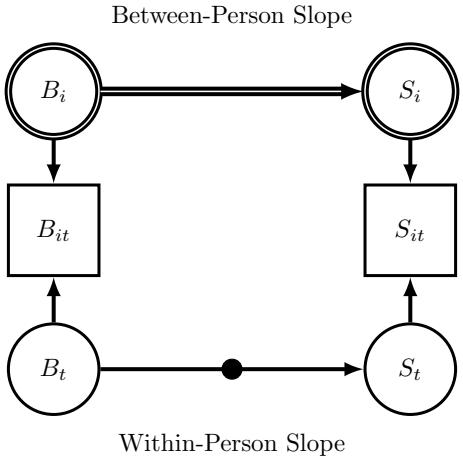


Figure 3: Diagram showing the first set of models. The behavior measures (B_{it}) and symptom scores ($S{+}it$) are decomposed into into between-person (i) components and within-person (t) components, and each symptom score component is regressed on the corresponding behavior component (double-lines represent between-person components and regression pathways). Furthermore, the within-person slopes are estimated as varying across participants (as indicated by the dot on the regression pathway).

and rooting them, meaning that larger errors are penalized more than smaller errors; thus, it is a more conservative estimate of performance than alternatives like the mean absolute error. Given the relatively small sample size, we did not expect our predictive models to achieve their maximum possible performance, and we expected our performance estimates to contain quite a bit of uncertainty. However, we assessed whether our models performed significantly better than a baseline to illustrate how, in future extensions of this work, we hope to improve model performance and hold ourselves to stringent performance criteria to map to demonstrable clinical utility. The specific approach we used to estimate these models was based on Bayesian multilevel modeling (Gelman et al., 2014; McElreath, 2016) and is detailed in Appendix C.

2.3.4. Inferential Modeling

To gain insight into the nature, reliability, and specificity of the statistical relationships between specific clinical and language measures, we also built two sets of linear regression models. We first describe the motivation for and formulation of our models, then the procedures we used to estimate their parameters, and finally the approach we used to interpret their results.

2.3.5. Inferential Model Formulation

The first set of inferential models regressed either the PANSS Positive or PANSS Negative scores on each language measure separately (Figure 3). The slope parameters in these models capture the nature (i.e., size and sign) of the relationship between the participants' symptoms and each language measure, and the uncertainty in the estimation of these parameters captures their level of reliability. The second set of inferential models regressed each language measure simultaneously on both the PANSS Positive and PANSS Negative scores (Figure 4). Including both clinical measures as predictors in the same models allowed us to control for their shared variance statistically. The slope parameters in these models capture the degree to which each is *uniquely* related to the language measures. In all inferential models, we also controlled for participant sex as male participants tended to have higher PANSS Positive scores than female participants.

As described in subsection 2.3, we needed to account for the fact that we had multiple observations of the same participants. Instead of building separate models for the between-person and within-person components as we did for the predictive models, our approach to inferential modeling allowed us to include both components in a single model. We used multilevel regression models (de Leeuw and Meijer, 2007) with a two-level structure, nesting observations within participants. Each language measure was decomposed into between-person and within-person components and these components were added to the models as

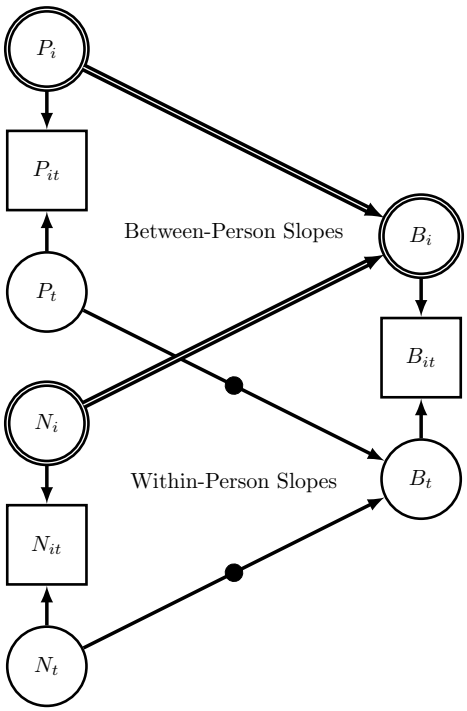


Figure 4: Diagram showing the second set of models. The positive symptom scores (P_{it}), negative symptom scores (N_{it}), and behavior measures (B_{it}) are decomposed into into between-person (i) components and within-person (t) components, and each behavior measure component is regressed on the corresponding symptom components (double-lines represent between-person components and regression pathways). Furthermore, the within-person slopes are estimated as varying across participants (as indicated by the dots on the regression pathways).

separate predictor variables. Varying effects (i.e., random effects) were estimated to allow participants to have different average levels of variables (i.e., intercepts) and different relationships between variables (i.e., slopes). In interpreting the results, we focused on the population-level effects (i.e., fixed effects), which estimate the central tendencies of the distribution of varying effects (e.g., the typical intercept or slope in the population of participants). To promote interpretability and comparability across models, we calculated standardized slopes in each model using the approach described by Schuurman et al. (2016): standardizing the within-person slopes using the within-person standard deviations and standardizing the between-person slopes using the between-person standard deviations.

Finally, to account for the possibility that our outcome variables might be skewed in their distributions, we characterized them using the skew-normal distribution (O'Hagan and Leonard, 1976), which can be parameterized to add a skewness parameter α to the mean and standard deviation parameters of the Gaussian distribution to allow non-zero skewness (and simplifies to the Gaussian distribution when $\alpha = 0$). The exact parameterization used is provided in Appendix D.

2.3.6. Inferential Model Estimation

We implemented our models within a Bayesian multilevel modeling framework (Gelman et al., 2014; McElreath, 2016). In brief, Bayesian methods combine existing knowledge about the probability of different parameter values (in the form of prior distributions) with observed data to generate updated knowledge about the parameter values (in the form of posterior distributions). Statistical inferences can then be made using this updated knowledge (e.g., estimating the central tendency and spread of the posterior distributions). We estimated our models using the `brms` package (Bürkner, 2017, 2018) as a high-level interface to the `Stan` platform for statistical computing (Gelman et al., 2015). Model estimation was performed through Markov chain Monte Carlo (Neal, 1993) via the No-U-Turn Sampler algorithm (Hoffman and Gelman,

2014), which converges quickly in high-dimensional models and eliminates the need for any hand-tuning of hyperparameters.

In setting the prior distributions for our model parameters, we strove to exclude unreasonable values without ruling out reasonable values (Gelman et al., 2014). Skew-normal multilevel regression models have six main types of parameters: slopes, intercepts, standard deviations of the varying effects, correlations between the varying effects, the standard deviation of the residuals, and the skewness parameter. For the slope parameters, because these are the main parameters of interest, we used more conservative normal priors centered on zero to apply light regularization and deter overfitting; the widths of these priors were set to the sample standard deviation of the outcome variable. We used less conservative Student's t priors for the intercept parameters to reflect that we did not have substantive hypotheses for these parameters and wanted them to be well-estimated; these priors were centered around the outcome variable's sample mean. We used nonnegative Student's t priors centered on zero to rule out unreasonable negative values for the varying effects' standard deviations. For the correlations between varying effects, we used the approach of Lewandowski et al. (2009) to assign an equal prior probability to all valid correlation matrices. For the standard deviation of the residuals, we used Student's t priors centered on zero. Finally, we used normal priors centered on zero for the skewness parameter to assign a high prior probability to typical skewness values. Implementation details, including exact priors are provided in Appendix E and Appendix F.

2.3.7. Inferential Model Interpretation

To interpret our inferential models' results, we represented each slope's magnitude as the central tendency of its posterior distribution and the precision of each effect as the spread of its posterior distribution. Specifically, we used the posterior median to measure central tendency and the 89% highest density interval (HDI) to measure spread. The posterior median minimizes the expected absolute error, and the 89% HDI is the narrowest continuous interval that contains 89% of the posterior density. The 89% HDI has become common in Bayesian data analysis because it is more stable than the 95% HDI (Kruschke, 2014) and because it highlights the arbitrariness of such threshold conventions in the first place (McElreath, 2016). Second, for each effect, we calculated the probability of direction (*pd*), which is an "effect existence index" that varies from 50% to 100% and can be interpreted as the probability that a parameter is strictly positive or negative (Makowski et al., 2019). We interpreted effects with *pd* values above 95% as "significant." However, we appreciate the arbitrariness of this cutoff and encourage readers to consider the HDIs directly.

3. Results

A total of 49 eligible participants diagnosed with either schizophrenia, bipolar disorder, major depressive disorder, or a related condition participated in the study between April 2015 and October 2018. Because this paradigm was in development during the initial data collection phase, the first eight participants were unusable for subsequent analysis, as we made refinements to the data acquisition procedures. We did not collect or store any of these participants' personally identifying or demographic data. Three participants had no clinically scaled sessions using PANSS, leaving 38 participants with a clinically scaled session available for analysis.

3.1. Background Characteristics

Demographics were collected from these 38 individuals and are reported in Table 2. Out of the 99 clinical encounters recorded from these individuals, 82 were accompanied by a research interview for PANSS scoring. Four of these recordings were excluded based on low quality or missing audio data caused by equipment malfunction, poor microphone placement, or miscellaneous issues with transcription. This exclusion resulted in a final sample of 78 sessions from 38 participants that were considered for analysis.

The semi-structured MD interviews lasted from 3.1 to 18.9 minutes ($M = 9.4$, $SD = 3.8$) in duration. The total number of recorded sessions per participant ranged from 1 to 8 ($M = 2.1$, $SD = 1.5$), and interviews occurred between 1 and 30 days after study enrollment($M = 4.1$, $SD = 4.7$), all before hospital discharge.

Table 2: Distribution of demographic variables in the sample of participants

	Count	Percent		Count	Percent
Sex			Education		
Female	15	39%	12 or fewer years	3	8%
Male	23	61%	13 years	1	3%
Race			14 years	1	3%
White	30	79%	16 years	10	26%
Black	2	5%	17 or more years	16	42%
Asian	1	3%	Age		
Multiple	1	3%	18-24 years old	10	26%
(Not Reported)	4	11%	25-29 years old	12	32%
Ethnicity			30-39 years old	6	16%
Latino or Hispanic	3	8%	40-49 years old	7	18%
Not Latino or Hispanic	26	68%	50-59 years old	2	5%
(Not Reported)	9	24%	(Not Reported)	1	3%

3.2. Clinical Results

Both PANSS scales range from a minimum possible score of 7 to a maximum possible score of 49. Across the 78 included sessions, symptom severity spanned the continuum from low to high severity, as expected based on the acute inpatient setting. PANSS Positive scores ranged from 7 to 36; as illustrated in Figure 5a, sessions were distributed across all four of the quartiles reported by Kay et al. (1987): low (46 sessions; 59%), medium (12 sessions, 15%), high (7 sessions, 9%), and severe (13 sessions, 17%). PANSS Negative scores ranged from 7 to 27; as illustrated in Figure 5b, sessions were distributed mostly among the low (59 sessions, 76%) and medium (16 sessions, 21%) quartiles, with few falling in the high (2 sessions, 3%) or severe (1 session, 1%) quartiles.

3.3. Predictive Modeling Results

Our predictive models' performance quantifies the extent to which the selected language measures could be combined to predict the PANSS Positive and PANSS Negative scores using our data. For both scales, we evaluated both the ability to predict the between-person component and the within-person component.

The results for predicting the PANSS Positive scores are provided in Table 3, and the model predictions in each repetition of the cross-validation procedure are depicted in Figure 6. In predicting the between-person labels from the between-person features, the Random Forest algorithm had the best performance with an estimated RMSE score of 5.60, which is better than the estimated baseline of 6.40 (which would be achieved by guessing the mean PANSS Positive score for all sessions). Our performance comparison model found an 86.7% chance (i.e., posterior probability) that this algorithm was better than the baseline. In predicting the within-person labels from the within-person features, the Elastic Net algorithm had the best performance with an estimated RMSE of 3.81, which is better than the estimated baseline of 4.15. Our performance comparison model found a 78.5% chance that this algorithm was better than the baseline.

The results for predicting PANSS Negative scores are provided in Table 4, and the model predictions are depicted in Figure 7. In predicting the between-person labels from the between-person features, the Random Forest algorithm had the best performance with an estimated RMSE of 4.05, which is better than the estimated baseline of 4.32. Our performance comparison model found a 78.9% chance that this algorithm was better than the baseline. In predicting the within-person labels from the within-person features, the Elastic Net algorithm had the best performance with an estimated RMSE of 2.29, which is better than the estimated baseline of 4.32. Our performance comparison model found a 56.8% chance that this algorithm was better than the baseline.

Table 3: Results of the Bayesian model comparing the PANSS Positive predictive models' performance to the baseline

	RMSE in Predicting PANSS Positive			
	Median	89% HDI	Baseline	Prob.
Between-Person				
Elastic Net	5.97	[4.72, 7.17]	6.40	71.7%
Support Vector	6.49	[4.96, 7.84]	6.40	45.6%
Random Forest	5.60	[4.37, 6.81]	6.40	86.7%
Within-Person				
Elastic Net	3.81	[3.06, 4.55]	4.15	78.5%
Support Vector	4.23	[3.31, 5.19]	4.15	44.6%
Random Forest	4.05	[3.23, 4.80]	4.15	58.8%

HDI = Highest density interval, Baseline = Score to beat,
Prob. = Posterior probability that median < baseline.

Table 4: Results of the Bayesian model comparing the PANSS Negative predictive models' performance to the baseline

	RMSE in Predicting PANSS Negative			
	Median	89% HDI	Baseline	Prob.
Between-Person				
Elastic Net	4.20	[3.43, 4.91]	4.32	61.1%
Support Vector	4.24	[3.53, 4.97]	4.32	57.3%
Random Forest	4.05	[3.44, 4.59]	4.32	78.9%
Within-Person				
Elastic Net	2.29	[1.80, 2.75]	2.34	56.8%
Support Vector	2.30	[1.80, 2.77]	2.34	55.6%
Random Forest	2.38	[1.85, 2.87]	2.34	44.3%

HDI = Highest density interval, Baseline = Score to beat,
Prob. = Posterior probability that median < baseline.

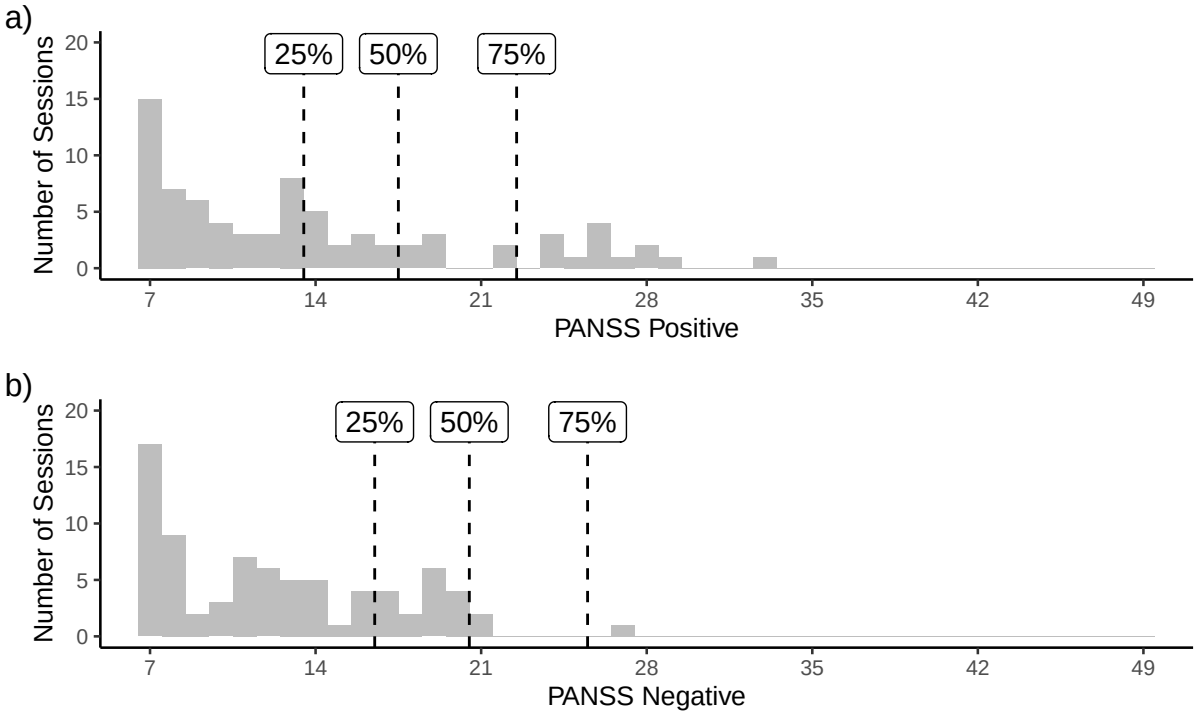


Figure 5: Histograms depicting the distributions of the (a) PANSS Positive scores and (b) PANSS Negative scores in the current sample with dashed reference lines depicting the quartiles reported by Kay et al. (1987)

3.4. Inferential Modeling Results

3.4.1. Language Measures Associated with Positive and Negative Symptoms

For each of the two clinical outcomes measures (PANSS Positive and PANSS Negative), we first tested for associations with the 14 language measures independently (i.e., without controlling for the other clinical measure).

The results for PANSS Positive are provided in Table 5 and depicted in Figure 8. Between-person slopes were significant for four lexical features: *increased* use of Perceptual Process words with higher PANSS Positive ($B = 0.45$, $pd = 100\%$), *increased* use of Affiliation words with higher PANSS Positive ($B = 0.27$, $pd = 97.0\%$), *decreased* use of Achievement words ($B = -0.24$, $pd = 96.5\%$), and *decreased* use of Reward words with higher PANSS Positive ($B = -0.29$, $pd = 98.6\%$). Between-person slopes for the remaining six lexical features did not show any reliable relationship with PANSS Positive. The between-person slopes were also non-significant for the coherence and all three disfluency features. Within-person slopes were significant for one lexical feature: *increased* use of Negative Emotion words were associated with higher PANSS Positive ($B = 0.41$, $pd = 98.9\%$). Within-person slopes were significant for the coherence feature: *increased* perplexity was associated with higher PANSS Positive ($B = 0.26$, $pd = 95.1\%$). Within-person slopes were non-significant for the other nine lexical features (although the slope for Risk words was suggestive) and non-significant for all disfluency features.

The results for PANSS Negative are provided in Table 6 and depicted in Figure 9. Between-person slopes were significant for two lexical features and one disfluency feature: *increased* Negative Emotion words were associated with higher PANSS Negative scores ($B = 0.32$, $pd = 98.4\%$), *decreased* Relative words were associated with higher PANSS Negative scores ($B = -0.25$, $pd = 96.0\%$), and *increased* edits were associated with higher PANSS Negative ($B = 0.27$, $pd = 97.1\%$). Within-person slopes were significant for two lexical features: *decreased* Cognitive Processes words were associated with higher PANSS Negative scores ($B = -0.37$, $pd = 97.9\%$), and *decreased* Power words were associated with higher PANSS Negative

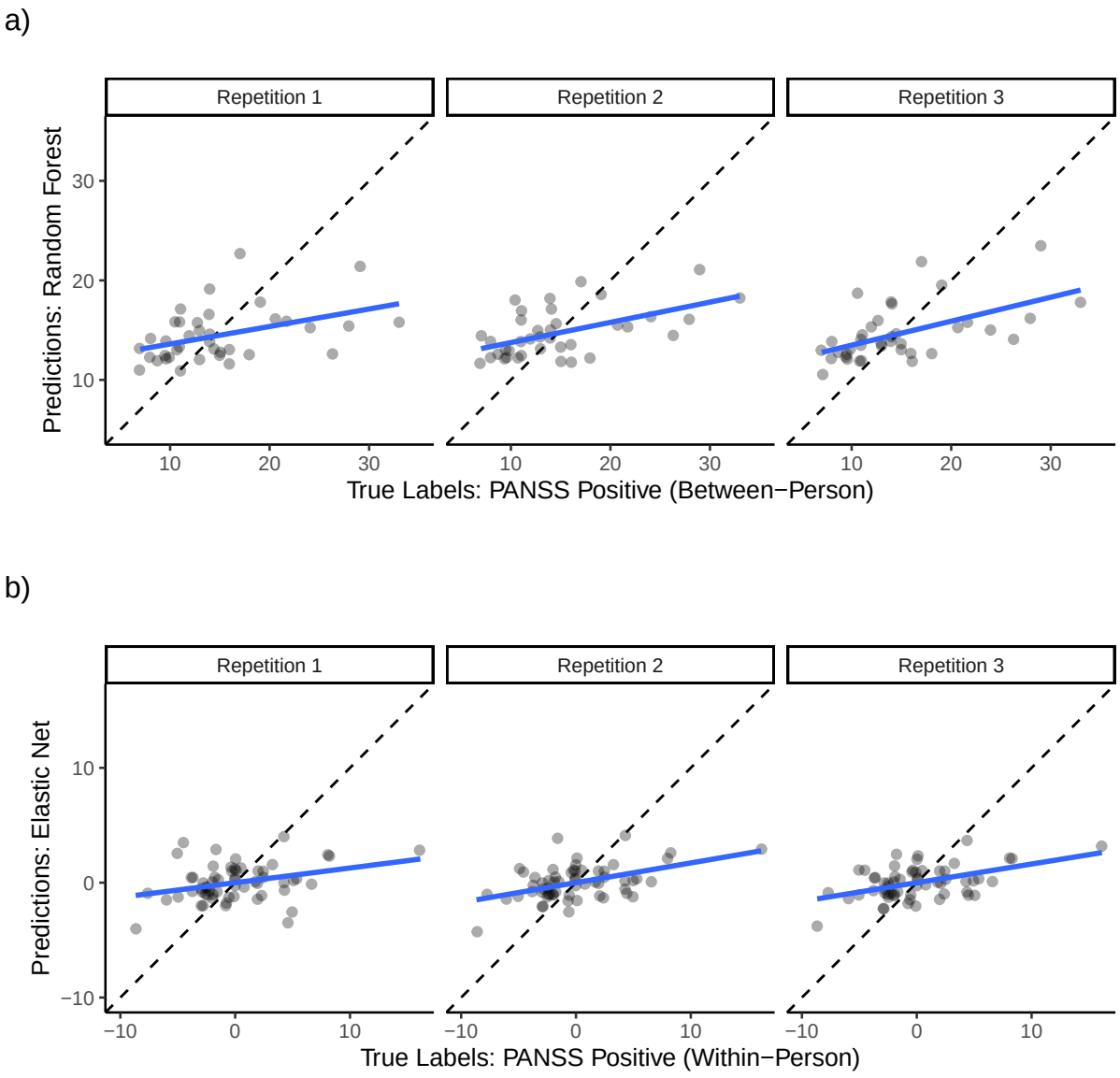


Figure 6: Predictions from each repetition of the outer cross-validation in predicting (a) the PANSS Positive Between-Person component using Random Forest and (b) the PANSS Positive Within-Person component using Elastic Net.

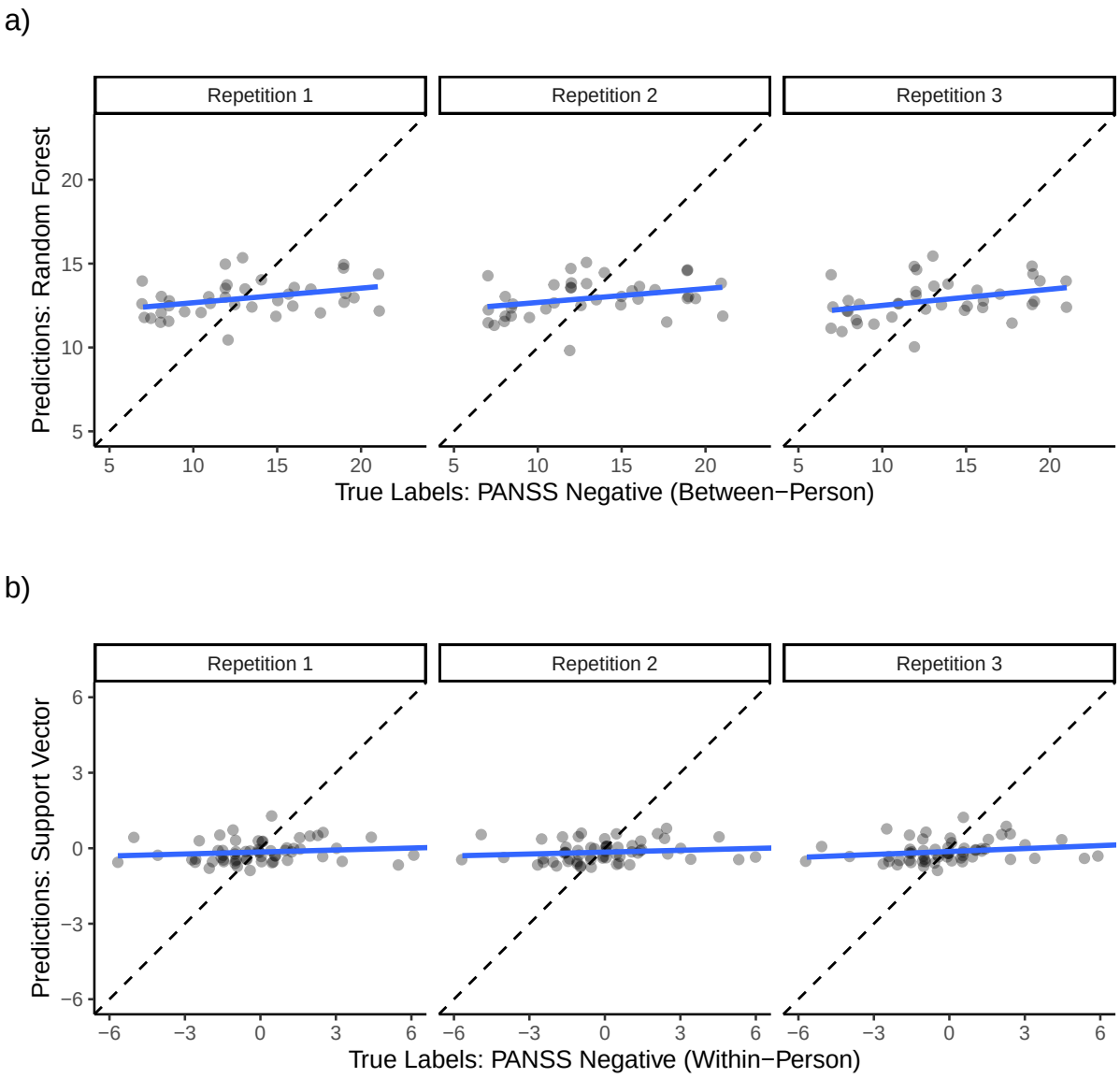


Figure 7: Predictions from each repetition of the outer cross-validation in predicting (a) the PANSS Negative Between-Person component using Random Forest and (b) the PANSS Negative Within-Person component using Support Vector.

Table 5: Population-level effects from the first set of inferential models predicting PANSS Positive scores from the decomposed language measures. Note that each row represents a separate model and all models also control for participant sex.

	Between-Person Slope			Within-Person Slope		
	Median	89% HDI	Sig.	Median	89% HDI	Sig.
Lexical Features						
Positive Emotion Words	0.11	[−0.13, 0.33]		−0.13	[−0.45, 0.15]	
Negative Emotion Words	0.10	[−0.15, 0.32]		0.41	[0.15, 0.68]	⋆
Cognitive Processes Words	−0.16	[−0.37, 0.06]		−0.18	[−0.49, 0.12]	
Perceptual Processes Words	0.45	[0.27, 0.60]	⋆⋆	0.00	[−0.23, 0.24]	
Affiliation Words	0.27	[0.06, 0.48]	⋆	0.04	[−0.20, 0.29]	
Achievement Words	−0.24	[−0.46, −0.01]	⋆	−0.09	[−0.41, 0.20]	
Power Words	0.02	[−0.19, 0.22]		−0.14	[−0.47, 0.14]	
Reward Words	−0.29	[−0.52, −0.08]	⋆	0.00	[−0.25, 0.24]	
Risk Words	−0.12	[−0.33, 0.07]		0.20	[−0.05, 0.44]	
Relativity Words	0.08	[−0.11, 0.30]		0.05	[−0.19, 0.27]	
Coherence Features						
Perplexity	0.05	[−0.17, 0.28]		0.26	[0.00, 0.52]	⋆
Disfluency Features						
Edits	−0.09	[−0.31, 0.11]		−0.16	[−0.43, 0.12]	
Restarts	0.05	[−0.18, 0.26]		0.03	[−0.21, 0.28]	
Repeats	0.05	[−0.18, 0.26]		0.01	[−0.25, 0.23]	

HDI = Highest density interval, Sig. = Significance, ⋆ $pd > 95\%$, ⋆⋆ $pd > 99\%$.

Table 6: Population-level effects from the first set of inferential models predicting PANSS Negative scores from the decomposed language measures. Note that each row represents a separate model and each model also controls for participant sex.

	Between-Person Slope			Within-Person Slope		
	Median	89% HDI	Sig.	Median	89% HDI	Sig.
Lexical Features						
Positive Emotion Words	0.12	[−0.07, 0.33]		0.04	[−0.26, 0.34]	
Negative Emotion Words	0.32	[0.08, 0.54]	⋆	0.15	[−0.18, 0.48]	
Cognitive Processes Words	0.09	[−0.13, 0.30]		−0.37	[−0.70, −0.06]	⋆
Perceptual Processes Words	0.06	[−0.15, 0.27]		0.09	[−0.18, 0.36]	
Affiliation Words	0.03	[−0.18, 0.23]		−0.04	[−0.32, 0.23]	
Achievement Words	0.07	[−0.17, 0.30]		0.07	[−0.16, 0.31]	
Power Words	−0.02	[−0.24, 0.19]		−0.48	[−0.82, −0.16]	⋆⋆
Reward Words	−0.09	[−0.32, 0.14]		0.07	[−0.26, 0.43]	
Risk Words	0.16	[−0.06, 0.41]		0.02	[−0.22, 0.27]	
Relativity Words	−0.25	[−0.47, −0.02]	⋆	−0.20	[−0.49, 0.09]	
Coherence Features						
Perplexity	−0.12	[−0.37, 0.13]		0.09	[−0.27, 0.43]	
Disfluency Features						
Edits	0.27	[0.04, 0.49]	⋆	0.24	[−0.14, 0.62]	
Restarts	0.12	[−0.12, 0.36]		0.04	[−0.23, 0.30]	
Repeats	0.15	[−0.07, 0.40]		0.00	[−0.25, 0.26]	

HDI = Highest density interval, Sig. = Significance, ⋆ $pd > 95\%$, ⋆⋆ $pd > 99\%$.

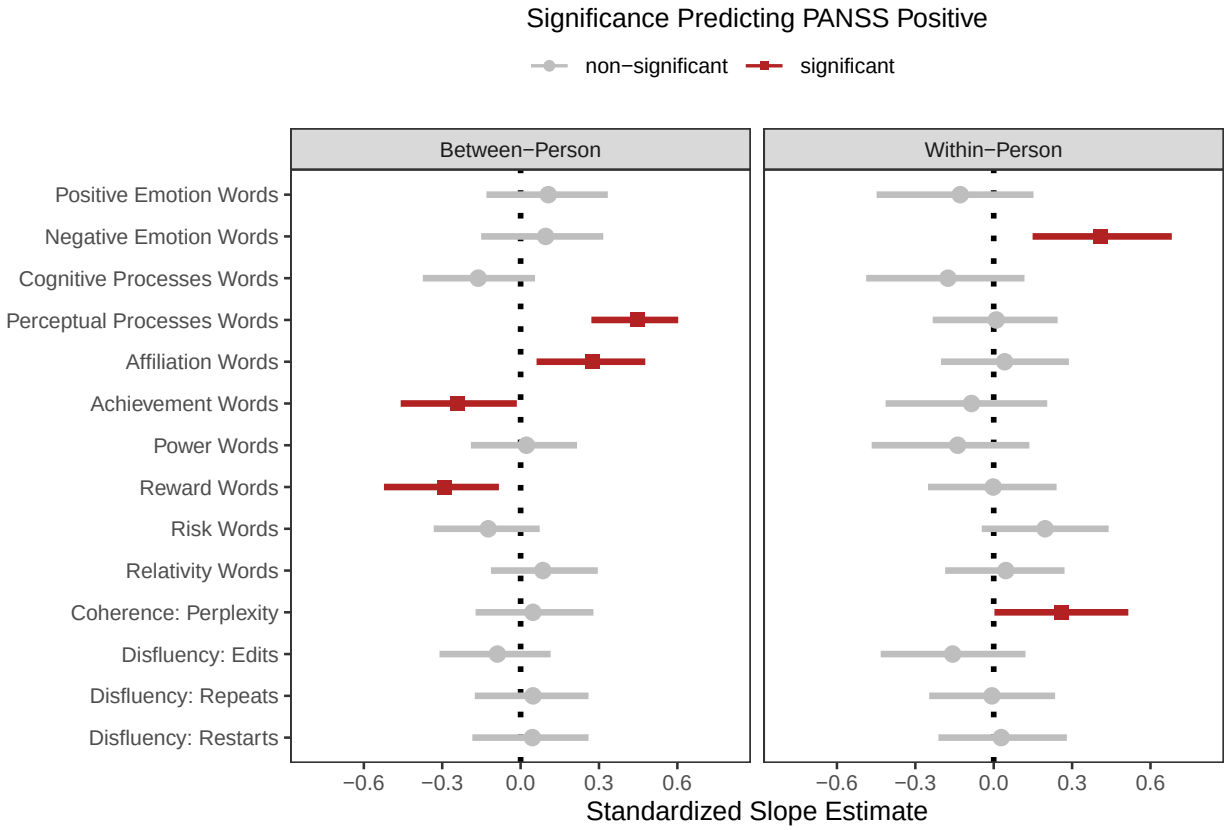


Figure 8: Forest plot depicting the standardized slope estimates in the PANSS Positive inferential models (points are posterior medians and intervals are posterior 89% highest density intervals)

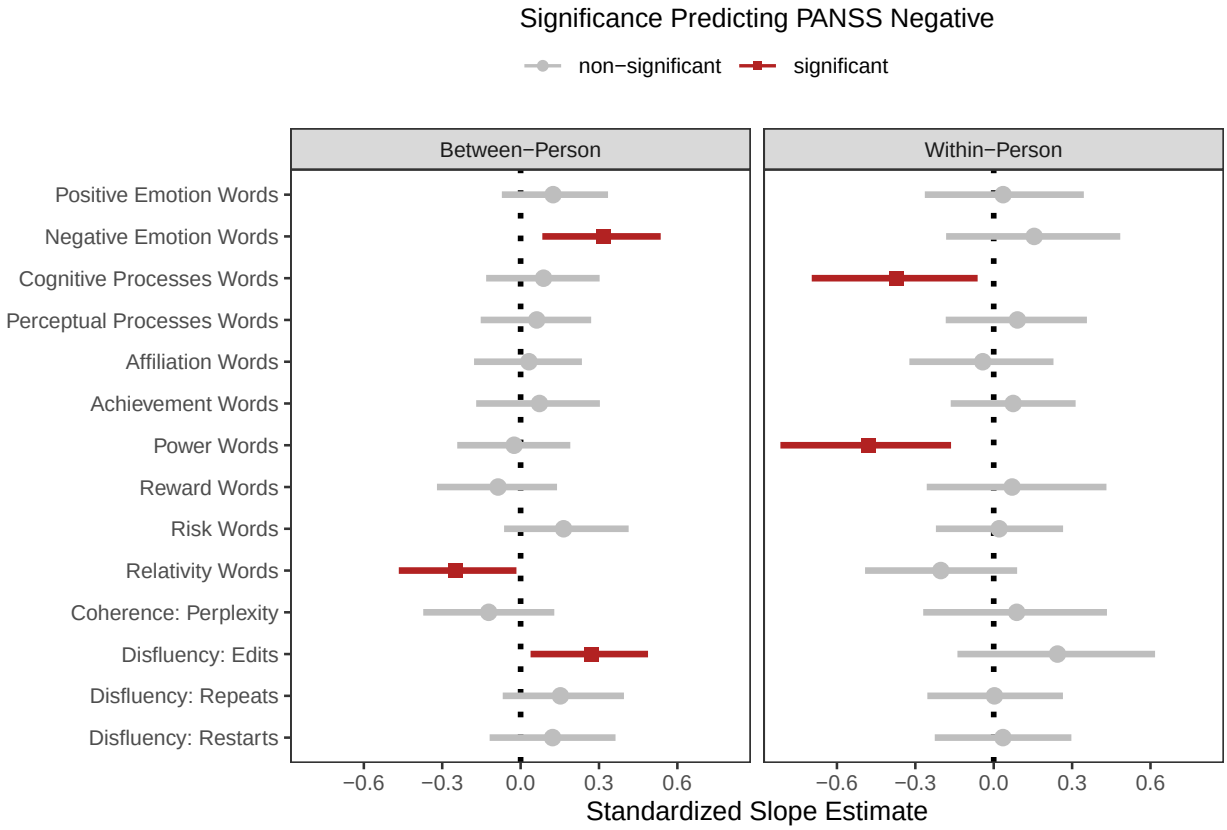


Figure 9: Forest plot depicting the standardized slope estimates in the PANSS Negative inferential models (points are posterior medians and intervals are posterior 89% highest density intervals)

Table 7: Population-level effects from the second set of inferential models predicting each language measure from the decomposed PANSS Positive and PANSS Negative scores. Note that each row is a separate model and that all models control for participant sex.

	Between-Person Slope		Within-Person Slope	
	PANSS+	PANSS−	PANSS+	PANSS−
Lexical Features				
Positive Emotion Words	−0.11	0.16	−0.16	0.21
Negative Emotion Words	−0.08	0.42 ⋆	0.49 ⋆	0.04
Cognitive Processes Words	−0.39 ⋆	0.27	0.02	−0.18
Perceptual Processes Words	0.57 ⋆⋆	−0.15	0.11	0.02
Affiliation Words	0.09	−0.05	0.09	0.01
Achievement Words	−0.49 ⋆⋆	0.22	−0.10	0.21
Power Words	−0.07	0.03	0.17	−0.45 ⋆
Reward Words	−0.46 ⋆⋆	0.03	−0.01	0.10
Risk Words	−0.32 ⋆	0.24	0.29	−0.13
Relativity Words	0.38 ⋆	−0.47 ⋆⋆	0.37	−0.52 ⋆
Coherence Features				
Perplexity	0.19	−0.22	0.19	0.04
Disfluency Features				
Edits	−0.37 ⋆	0.49 ⋆⋆	−0.24	0.19
Repeats	−0.13	0.13	0.02	0.03
Restarts	−0.14	0.19	−0.10	0.04

PANSS+ = PANSS Positive, PANSS− = PANSS Negative, ⋆ $pd > 95\%$, ⋆⋆ $pd > 99\%$.

scores ($B = -0.48$, $pd = 99.4\%$). Within-person slopes related to PANSS negative were non-significant for the other eight lexical features, the coherence feature, and the disfluency features.

3.4.2. Specificity of Associations to Positive and Negative Symptoms

Our second set of inferential models sought to identify associations with language measures specific to positive and negative symptoms, i.e., statistical relationships that persisted after controlling for the other clinical measure (and participant sex). Although the slopes in these models are standardized and can be interpreted similarly to those in the first set of inferential models, these are *unique* or *partial* effects, controlling for the clinical scales' shared variance. That is, language measures identified by these models (as significantly associated with one clinical score or the other) are linked only to *unique* variance in that score and *not* with variance shared with the other clinical score. The estimates for the partial effects are presented in Table 7. As in the first set of inferential models, we first present these selective results by clinical measure (positive and negative symptoms) and then by component (between-person and within-person slopes).

Between-person PANSS Positive symptoms were significantly and *uniquely* associated with six lexical features and one disfluency feature: *increased* use of Perceptual Processes words ($B = 0.57$, $pd = 100\%$) and Relativity words ($B = 0.38$, $pd = 97.9\%$) were each associated with higher PANSS Positive across individuals; *decreased* use of Cognitive Processes words ($B = -0.39$, $pd = 97.6\%$), Achievement words ($B = -0.49$, $pd = 99.7\%$), Reward words ($B = -0.46$, $pd = 99.8\%$), and Risk words ($B = -0.32$, $pd = 95.4\%$) were also associated with higher PANSS Positive across individuals; and *decreased* edits were associated with higher PANSS Positive across individuals ($B = -0.37$, $pd = 97.8\%$). The other four lexical features, the other two disfluency features, and the coherence feature had non-significant unique between-person effects. Within-person PANSS Positive symptom slopes were significantly and uniquely associated with a single lexical feature: *increased* use of Negative Emotion words was linked to higher PANSS Positive ($B = 0.49$, $pd = 98.1\%$). Two lexical features were suggestive of a unique within-person association with PANSS

Positive: *increased* use of Risk ($B = 0.29$, $pd = 91.4\%$) and Relativity ($B = 0.37$, $pd = 94.4\%$) words were linked to higher PANSS Positive within an individual. The other seven lexical features, the coherence feature, and the disfluency features had non-significant unique within-person effects.

Between-person PANSS Negative slopes were significantly and *uniquely* associated with two lexical features and one disfluency feature: *increased* use of Negative Emotion words ($B = 0.42$, $pd = 99.0\%$), *decreased* use of Relativity words ($B = -0.47$, $pd = 99.4\%$), and *increased* edits ($B = 0.49$, $pd = 99.5\%$) were each associated with higher PANSS Negative across participants. Between-person PANSS Negative slopes were suggestive for two lexical features: *increased* use of Cognitive Processes ($B = 0.27$, $pd = 91.9\%$) and Risk ($B = 0.24$, $pd = 90.2\%$) words were associated with higher PANSS Negative across individuals. The other six lexical features, the coherence feature, and the other disfluency features had non-significant unique between-person effects. Within-person PANSS Negative Symptoms were significantly and uniquely associated with two lexical features: *decreased* use of Power words ($B = -0.45$, $pd = 96.8\%$) and *decreased* use of Relativity words ($B = -0.52$, $pd = 97.6\%$) were associated with higher PANSS Negative within participants. The other eight lexical features, the coherence feature, and disfluency features had non-significant unique within-person effects.

4. Discussion

Changes in expressive and receptive language are common to many prominent mental state abnormalities and vary as a function of clinical severity in both psychotic and affective disorders. Here, we used a semi-structured clinical encounter to capture language measures in individuals hospitalized for mental disturbances. We focused on linking specific language measures — which can be readily quantified using available computational approaches developed by the natural language processing (NLP) community — with core clinical outcome assessment measures of "positive" and "negative" symptoms of psychosis. We first discuss our overall experiences with feasibility and tolerability of performing audio-visual recordings in acutely ill individuals, including what we have found to be essential and non-essential components of a successful research program leveraging inpatient psychiatric settings to study individuals whose mental states vary both relative to one another and over time. We first summarize and discuss statistical inference testing results, where each language measure was studied for relationships to positive and negative symptoms. We then summarize the predictive models' results, where we used machine learning to combine informative features and studied models' performances at estimating symptom severity levels. Finally, we acknowledge some limitations of the present work and highlight critical outstanding questions that will be addressed in future studies.

4.1. Applying a Computational Approach to Clinical Instruments

Although the behavioral health community is increasingly embracing evidence-based clinical practices, neither psychiatry nor psychology has adopted a single objective marker of illness into standard clinical practice. The judgment of expert clinicians and raters to detect the diagnostic and prognostic features of psychiatric illness is not only expensive but almost impossible to monitor for efficacy since heterogeneous training and credentialing practices can lead to substantial variance in clinical practice that limits generalizability. Performance of behavioral task paradigms—as employed in experimental psychology, cognitive neuroscience, or computational neuroscience—provides one strategy for probing latent model parameters related to mental state disturbances (Browning et al., 2020; Huys et al., 2016; Maia et al., 2017). However, there has been a recent push in cognitive neuroscience toward using task-free naturalistic paradigms that, compared to modular task-based paradigms, provide a far richer, contextual, and more realistic probe of mental functioning. While task-based paradigms may generate powerful evidence in controlled settings or less acutely ill individuals, deriving reliable illness severity measures based on task performance alone is challenging in real-world settings and at the higher end of the severity spectrum. For example, severely ill individuals may be unable or unwilling to perform tasks designed to probe the very functions that are potentially impaired in these individuals (e.g., aphasic individuals may be unable to perform some language tasks), complicating the interpretation of these assessment strategies.

Our research program is focused on developing novel methods for assessing clinical outcomes more effectively than previously possible with the aid of novel measurement sources, computational modeling, and machine teaching, which leverages clinical expertise to improve model performance while addressing feasibility in real-world settings, including both traditional care settings and home environments. The application of such novel approaches to the study of psychiatric populations could revolutionize the mental healthcare system, providing an urgently needed backstop of objective behavioral data to aid clinical decision-making and estimation of clinical outcomes. Importantly, this work aims not to replace clinicians or expert clinical raters but rather to augment their abilities, expedite training, make more practitioners more effective in more settings, and begin to provide much-needed evidence to scaffold and help shape clinical practices.

4.2. Use of Multilevel Models in Repeated Measures Studies of Individuals

Here, we developed separate models to predict the between-person clinical components (from the between-person language components) and the within-person clinical components (from the within-person language components) to examine a longitudinal sample comprised of different individuals providing a different number of repeated samples and showing different illness trajectories over their hospital course (and study participation). Multilevel models are especially critical when developing objective markers with putative clinical utility since the relationship between two variables may differ across levels (Hamaker and Muthén, 2019).

The challenge of using multilevel models to examine the association of features both between- and within-individuals is how best to interpret features showing association in one but not both contexts. While less intuitive, we feel it is critical to help clinicians build the understanding that the two sets of relationships are often statistically uncoupled despite any default assumptions. Therefore, we were careful in this study to report the results for both types of effects, providing some educated guesses around the pattern of results, even though highly accurate and precise estimation of these effects will likely require larger samples than we had here.

4.3. Language Features Associated with Positive Symptoms

We found that participants who experienced more positive symptoms on average tended to express more words related to Perceptual Processes and Affiliation and fewer words related to Achievement and Reward. When participants experienced more positive symptoms than usual (relative to their baseline), they tended to express more Negative Emotion words than usual, and their language tended to be less difficult to predict than usual (Figure 8, Table 5).

When controlling for between-person negative symptom means, experiencing more positive symptoms on average predicted more frequent expression of Perceptual Processes and Relativity words, and less frequent expression of Cognitive Processes, Achievement, Reward, and Risk words and fewer edit disfluencies. When controlling for within-person negative symptom deviations, experiencing more positive symptoms than usual (relative to one's baseline) predicted more frequent expression of Negative Emotion words (and possibly more Risk and Relativity words; Table 7).

More frequent use of words related to Perceptual Processes associated with more severe positive symptoms could reflect an excess or distortion of normal perception, possibly driven by delusions and hallucinations. In contrast, the less frequent use of Achievement and Reward words associated with more severe positive symptoms could indicate a lower level of normal drives characteristic of healthy mental functioning. This suggestion is consistent with prior work using LIWC to compare written language between a group of individuals with a psychotic disorder and a group of non-psychotic individuals, showing relatively lower written expression of words related to work, achievement, and leisure in the psychotic group (Mitchell et al., 2015). See Figure 10 for examples of words that were frequently used in patients who experienced more positive symptoms.

4.4. Language Features Associated with Negative Symptoms

We found that participants who experienced more negative symptoms on average tended to express more Negative Emotion words and fewer Relativity words and tended to have more edit disfluencies. When

(a) Negative Emotion words

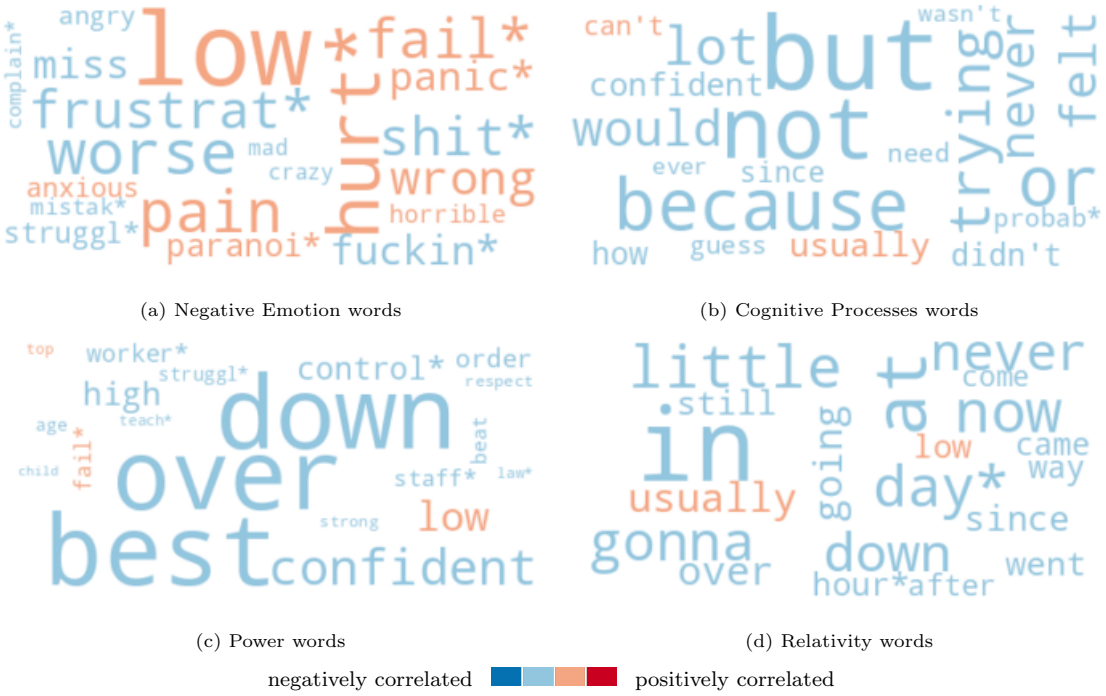
(b) Perceptual Processes words

(c) Affiliation words

(d) Reward words

negatively correlated positively correlated

Figure 10: **Words Associated with Positive Symptoms.** Each panel illustrates a word cloud, in which words associated with higher positive symptom scores are shown in red and words associated with lower positive symptom scores are shown in blue, with word size indicative of the overall frequency of occurrence. Lighter shades represent Pearson correlations between ±0.15 and ±0.30 and darker shades represent correlations between ±0.30 and ±1.00. Each panel illustrates words drawn from one of four lexical categories found to be associated with higher positive symptom scores: (a) Negative Emotion words, (b) Perceptual Processes words, (c) Affiliation words, and (d) Reward Words. Asterisks (*) indicate stems; e.g., 'attack*' represents the words 'attack', 'attacks', 'attacking', 'attacker', and so on.



(a) Negative Emotion words

(b) Cognitive Processes words

(c) Power words

(d) Relativity words

negatively correlated positively correlated

Figure 11: **Words Associated with Negative Symptoms.** Each panel illustrates a word cloud, in which words associated with higher negative symptom scores are shown in red and words associated with lower negative symptom scores are shown in blue, with word size indicative of the overall frequency of occurrence. Lighter shades represent Pearson correlations between ±0.15 and ±0.30 and darker shades represent correlations between ±0.30 and ±1.00. Each panel illustrates words drawn from one of four lexical categories found to be associated with higher negative symptom scores: (a) Negative Emotion words, (b) Cognitive Processes words, (c) Power words, and (d) Relativity Words. Asterisks (*) indicate stems; e.g., 'hurt*' represents the words 'hurt', 'hurts', 'hurting', 'hurtful', and so on.

participants experienced more negative symptoms than usual (relative to their baseline), they tended to express fewer words related to Power and Relativity than usual (Figure 9, Table 6).

When controlling for positive symptoms, experiencing more negative symptoms on average continued to show more frequent expression of Negative Emotion words, fewer Relativity words, and more edit disfluencies, similar to the negative symptom models that did not control for positive symptoms. The selective models also suggested possible relationships between higher negative symptoms with more Cognitive Processes and Risk words and possibly higher language coherence. Experiencing more negative symptoms than usual (controlling for positive symptoms) was associated with the expression of fewer Power and Relativity words than usual (Table 7).

The less frequent use of words expressing relativity to time and space in association with more severe negative symptoms (and less severe positive symptoms) is consistent with the possibility that this feature primarily captures the "poverty of speech" aspect of the negative symptom domain. This result supports prior work suggesting that computational measures of poverty of language content can predict psychotic symptoms (Rezaii et al., 2019). This result also suggests that, in our sample, the frequency of occurrence of Relativity words did not index the extent of reality monitoring because less frequent use of Relativity words would be expected in association with an increase in severity of positive symptoms (and not a decrease as observed here). See Figure 11 for examples of words that were frequently used in patients who experienced more negative symptoms.

The increased occurrence of speech edits in association with more severe negative symptoms (and less significantly, less severe positive symptoms) is consistent with the possibility that this language feature captures the speech disfluency often associated with the negative symptom domain. This finding complements prior work suggesting that some quantitative measures of speech disfluency (specifically, pauses in speech)

could distinguish individuals with schizophrenia from healthy controls (Cohen et al., 2014).

4.5. Language Features Associated with Both Negative and Positive Symptoms

The more frequent use of Negative Emotion words associated with more severe positive and negative symptoms suggests that this language feature may not be associated with specific aspects of psychotic illness. This finding is consistent with a body of work demonstrating negative biases in both the comprehension and expression of language across primary psychotic and primary affective disorders, including schizophrenia, major depressive disorder, and bipolar disorder (Dar et al., 2021; Fineberg et al., 2016; Hitczenko et al., 2020; Holt et al., 2006; Jalenques et al., 2013). Similarly, the less frequent expression of words related to Cognitive Processes associated with more severe positive *and* negative symptoms may reflect a general cognitive deficit, consistent with the idea that cognitive deficits are characteristic of mental disorders across multiple diagnoses (McTeague et al., 2017).

Thus, several of the language features predicted the PANSS Positive and Negative scores, consistent with our working hypothesis that automatic measurements of these aspects of language contain signals associated with illness severity in specific domains. Furthermore, several of the language features uniquely predicted the symptoms in one of the domains (i.e., positive or negative). While, from the present results, we cannot establish the relationship between language features and specific illness symptoms, our findings help generate testable and specific hypotheses, as detailed below.

4.6. Predicting Psychiatric Assessments Using Natural Language Processing

When predicting participants' positive symptoms, our best-performing model had an average error of around 5.6 points (on the 7-to-49 PANSS scale) when predicting participants' between-person means and an average error of around 3.8 points when predicting participants' within-person deviations. When compared to the standard deviations (SDs) of these variables, this means that errors on the between-person level tended to be around 0.88 SDs and errors on the within-person level tended to be around 0.92 SDs. When predicting participants' negative symptoms, our best-performing model had an average error of around 4.1 points (or 0.94 SDs) when predicting participants' between-person means and an average error of around 2.3 points (or 0.98 SDs) when predicting participants' within-person deviations. None of these models were significantly better than the baseline.

We view these model performances as currently below what would be expected to provide clinical actionability under most common clinical conditions. However, this is not too surprising given that machine learning models typically require much more data than we currently have to achieve their maximal performance. In this study, we have provided a rigorous framework for developing and evaluating predictive performance. As detailed below, future studies will collect more data and seek to improve predictive performance as we determine optimal clinical use cases where such tools might help aid assessments, including care triage and hospital discharge.

4.7. Feasibility and Tolerability of Audiovisual Assessment of Acutely Ill Individuals

In addition to AV methods refinements (as noted in subsubsection 2.1.2), our efforts allowed us to assess the feasibility and tolerability of our data collection protocol. Although we could not systematically collect information from non-consenting individuals about their reasons for declining to participate, our impression was that most hospitalized participants were interested in the study so as to have an opportunity to speak with a clinician. Enrolled participants did not seem to mind the presence of a camera or speaking into a microphone and, using the described AV methods, the quality of the recordings was sufficient for quantitative assessment of facial (e.g., facial expressions, eye gaze) and voice (e.g., speech rate, vowel space) features (Vail et al., 2018; Vijay et al., 2016; Wörtwein et al., 2017).

4.8. Urgent Need for Objective Markers of Psychotic and Affective Pathology

In psychiatric assessments, clinical instruments and extended, detailed interviews (e.g., First et al., 2015)) provide many numeric values to enter into computational analyses. Indeed, most of us in this field have attempted to tie biological endpoints to the numerical assignments of our "masters-level clinicians" and other research assistants, and even physicians, who spend hours with research participants asking questions and filling out forms. This process can be cumbersome and unreliable at the item-level, even with months of supervised training. In any case, a significant workforce cost and burden to patients is expended in generating these error-prone data sets that serve as the behavioral scaffold for most clinical decision-making, insurance claims, and translational research studies.

In clinical research, the principal reason we rely on this strategy for collecting behavioral phenotypic data from research participants is the great challenge of quantifying their mental suffering, paranoia, and thoughts of harming themselves. Patients' presenting complaints are often philosophical/existential and exceedingly personal, so we have a strong implicit bias that these idiosyncrasies would themselves be too challenging to quantify into some clinically meaningful system. There is intense pressure to adhere to quantitative systems, particularly from payer entities that are structured to optimize around numeric targets, thus having strong financial incentives to develop them. However, as clinicians, we tend to push back against the quantitative frameworks (e.g., DSM or ICD code, billable fields in a mental status exam, minutes or visits spent with the patient) as demeaning of the true nature of our clinical work and capturing almost no nuance in their presentation. These quantitative systems we rely on now are, for the most part, a cumbersome reality for most clinicians, not a learning healthcare system, in which knowledge generation processes are embedded in daily practice to produce continual improvement in care.

Eventually, we envision comprehensive behavioral monitoring solutions whereby suites of hardware (e.g., smartphone, computer, wearables, and other sensors), coupled with software (i.e., sophisticated computational analyses) will provide continuous assessment of a wide array of neuropsychiatric functions using entirely or predominantly naturalistic assessment via usage patterns and sensor data from devices. In the near term, strategies that incrementally advance on currently accepted practice are more pragmatic as proof-of-concept studies, in which the primary goal is to demonstrate the feasibility and utility of the approach. We, therefore, chose to focus on clinical instruments that are already used to assess patients. Applying a computational approach to one of these existing systems afforded minimal systemic burden, at least in principle, and could therefore be scalable if deemed to provide any useful or "clinically actionable" information.

4.9. Limitations and Future Directions

As already stated, a larger sample would be needed to achieve more competitive predictive performance, and yield more conclusive insights into the relationships between symptoms of a psychotic disorder and patterns of expressive language usage. Additionally, our sample represented the demographics of inpatients at McLean Hospital but it is currently unknown how well these results would generalize to samples drawn from different populations. For example, our sample was relatively high in education level, which may have influenced various aspects of their expressive language. In order to ensure that medical technologies such as the ones proposed here are developed and applied equitably across groups, it is critical that future work in this space (1) transparently reports samples' demographic characteristics, (2) acknowledges that the generalizability of results across contexts and groups cannot be assumed without evidence, (3) collects and analyzes data from increasingly diverse samples, and (4) systematically checks for and rigorously quantifies predictive biases and performance gaps across different populations.

We represented participants' psychiatric state using the total scores in two of three PANSS categories, the PANSS Positive and PANSS Negative. These measures are appropriate in that they are among the primary scales used to represent the severity of positive and negative symptoms of psychotic disorders, both in the clinic and in clinical research. However, several factor analyses of the PANSS suggest that the complex symptomatology of psychotic disorders is more accurately represented by a five-category classification (positive, negative, cognitive, excitement/hostility, anxiety/depression) than a three-category classification (positive, negative, cognitive) (Lindenmayer et al., 1994; van der Gaag et al., 2006; Citrome et al., 2011).

Thus, future work examining the relationships of language features with more precise and detailed symptoms categories (as defined in factor analyses using large samples (e.g., Citrome et al., 2011), or with specific symptoms (e.g., delusions and hallucinations), may improve our understanding of these relationships and yield more reliable language biomarkers.

We characterized participants' expressive language using ten lexical measures, a coherence measure, and several disfluency measures. However, the NLP community has developed other relevant measures that could be brought to bear in this type of work, such as use of metaphorical language (e.g., Gutiérrez et al., 2017) and complexity of syntax (Bedi et al., 2015; Corcoran et al., 2018; Elvevåg et al., 2007). In addition, receptive conversational skills (i.e., language comprehension) could be assessed by examining the coherence of participants' responses in the context of the MD's prompts and questions. We also note that the utility of our findings and approach using human transcription will only increase as technologies for automatic speech transcription become more sophisticated, accurate, and widely available.

Finally, future work should integrate spoken language measures with non-verbal measures of face, voice, and body expressions. Nonverbal behavior plays a crucial role in human communication: spoken language conveys in tandem linguistic information (i.e., the meaning of words and sentences) and paralinguistic, nonverbal information (i.e., the form of speech delivery). The nonverbal information, consisting of voice acoustic characteristics (e.g., tone, amplitude, prosody), nonverbal sounds (e.g., laugh, grunt, cry), facial expressions (e.g., eye gaze, frown, smile), and body posture (e.g., leaning, hand gestures), provides the context for understanding the linguistic information in spoken language and both are key for decoding the speaker's emotional and mental state. Both the verbal and nonverbal aspects of spoken language communication are an essential part of clinical assessments in mental health settings, and have been found to relate to symptom severity in preliminary studies in a subset of the present cohort where each type of feature was examined separately (Vail et al., 2017; Vijay et al., 2016; Wörtwein et al., 2017). Integrating the verbal and nonverbal measures of spoken language in larger samples will lead to more accurate modeling and better understanding of their relationships with mental disease symptoms.

5. Conclusions

We examined the feasibility and value of measuring lexical, coherence, and disfluency features of expressive language in individuals hospitalized with a psychotic disorder during brief semi-structured clinical interviews conducted over the course of their hospitalization. We found that our recording procedures were well tolerated by both clinical staff and hospitalized inpatients and produced data that were of sufficient quality for human transcription and computational analysis using natural language processing methods. Inferential modeling revealed language features that tracked with positive symptoms alone (e.g., Perceptual words), negative symptoms alone (e.g. disfluencies), as well as some features that tracked with both (e.g., Negative Emotion words). We conclude that automatic detection of language abnormalities associated with illness severity in specific symptoms domains of psychosis is possible and informative, even in acute clinical settings. Future studies with larger, more diverse samples, over extended periods of time, and in relation to other non-verbal signals such as vocal acoustics, face expression, and body movements, are necessary to establish the full potential of naturalistic recordings to generate rich, detailed information about mental health status. Nonetheless, these studies demonstrate how computationally-defined behavioral phenotypes, as can be derived from natural language, can serve to scaffold neurobiological observations seeking to establish causal links between specific circuit abnormalities and subjective experiences, by providing discrete and reproducible behavioral signals that can be measured with fidelity and temporal granularity. Given that the ultimate goal of neurobiological understanding is to alleviate subjective suffering of those experiencing neuropsychiatric conditions, our work demonstrates how computationally rigorous measurement and analysis of human behavior in individuals experiencing the conditions could provide a bridge between subjective and neurobiological observations, leading to a more progressive, rather than idiosyncratic, accumulation of scientific knowledge regarding the neurobiological mechanisms of disease risk, emergence, progression, and recovery.

Acknowledgements

The authors wish to thank Nathaniel Shogren and Claire Foley for their assistance in quality assurance of the audio-visual recordings.

Role of the Funding Source

This material is based upon work partially supported by the National Science Foundation (#1722822), the National Institutes of Health (U01-MH116925, R01-MH096951), and the Taplin Family Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material do not necessarily reflect the views of these funding sources, and no official endorsement should be inferred.

Conflicts of Interest

JTB has received consulting fees and equity from Mindstrong, Inc., unrelated to the present work. JTB has also received consulting fees from Google, Verily, Apple, Blackthorn Therapeutics, Pear Therapeutics, AbleTo, and Niraxx Therapeutics, all unrelated to the present work.

References

Andreasen, N.C., 1979. Thought, language, and communication disorders: II. Diagnostic significance. Arch. Gen. Psychiatry 36, 1325.

Andreasen, N.C., Grove, W.M., 1986. Thought, language, and communication in schizophrenia: Diagnosis and prognosis. Schizophr. Bull. 12, 348–359.

Andreasen, N.C., Olsen, S., 1982. Negative v positive schizophrenia: Definition and validation. Arch. Gen. Psychiatry 39, 789.

Bearden, C.E., Wu, K.N., Caplan, R., Cannon, T.D., 2011. Thought disorder and communication deviance as predictors of outcome in youth at clinical high risk for psychosis. J. Am. Acad. Child Adolesc. Psychiatry 50, 669–680.

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 1, 15030.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Browning, M., Carter, C.S., Chatham, C., Den Ouden, H., Gillan, C.M., Baker, J.T., Chekroud, A.M., Cools, R., Dayan, P., Gold, J., et al., 2020. Realizing the clinical potential of computational psychiatry: report from the banbury center meeting, february 2019. Biol. Psychiatry 88, e5–e10.

Bürkner, P., 2017. Brms: An R package for Bayesian multilevel models using Stan. J. Stat. Softw. 80, 1–28.

Bürkner, P., 2018. Advanced Bayesian multilevel modeling with the R package Brms. R J. 10, 395–411.

Cearns, M., Hahn, T., Baune, B.T., 2019. Recommendations and future directions for supervised machine learning in psychiatry. Transl. Psychiatry 9, 1–12.

Citrome, L., Meng, X., Hochfeld, M., 2011. Efficacy of iloperidone in schizophrenia: a PANSS five-factor analysis. Schizophr. Res. 131, 75–81.

Cohen, A.S., Mitchell, K.R., Elvevåg, B., 2014. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. Schizophr. Res. 159, 533–538.

Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., Bearden, C.E., Cecchi, G.A., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17, 67–75.

Corcoran, C.M., Mittal, V.A., Bearden, C.E., Gur, R.E., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G.A., Wolff, P., 2020. Language as a biomarker for psychosis: A natural language processing approach. Schizophr. Res. 226, 158–166.

Dar, S., Liebenthal, E., Pan, H., Smith, T., Savitz, A., Landa, Y., Silbersweig, D., Stern, E., 2021. Abnormal semantic processing of threat words associated with excitement and hostility symptoms in schizophrenia. Schizophr. Res. 228, 394–402.

de Boer, J.N., Voppel, A.E., Begemann, M.J.H., Schnack, H.G., Wijnen, F., Sommer, I.E.C., 2018. Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. Neurosci. Biobehav. Rev. 93, 85–92.

de Leeuw, J., Meijer, E. (Eds.), 2007. Handbook of Multilevel Analysis. Springer.

Docherty, N.M., Cohen, A.S., Nienow, T.M., Dinzeo, T.J., Dangelmaier, R.E., 2003. Stability of formal thought disorder and referential communication disturbances in schizophrenia. J. Abnorm. Psychol. 112, 469–475.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V., 1997. Support vector regression machines, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA. pp. 155–161.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. Schizophr. Res. 93, 304–316.

Fineberg, S.K., Leavitt, J., Deutsch-Link, S., Dealy, S., Landry, C.D., Pirruccio, K., Shea, S., Trent, S., Cecchi, G., Corlett, P.R., 2016. Self-reference in psychosis and depression: A language marker of illness. Psychol. Med. 46, 2605–2615.

First, M.B., Williams, J.W., Karg, R.S., Spitzer, R.L., 2015. Structured clinical interview for DSM-5 — Research Version (SCID-5-RV). American Psychiatric Association, Arlington, VA.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data Analysis. CRC Press, Boca Raton, FL. third edition.
Gelman, A., Lee, D., Guo, J., 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. J. Educ. Behav. Stat. 40, 530–543.
Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 517–520.
Gooding, D.C., Ott, S.L., Roberts, S.A., Erlenmeyer-Kimling, L., 2013. Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: Findings from the New York High-Risk Project. Psychol. Med. 43, 1003–1012.
Gutiérrez, E.D., Cecchi, G., Corcoran, C., Corlett, P., 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 2923–2930.
Hamaker, E.L., Muthén, B., 2019. The fixed versus random effects debate and how it relates to centering in multilevel modeling. Psychol. Methods .
Harrow, M., Marengo, J.T., 1986. Schizophrenic thought disorder at followup: Its persistence and prognostic significance. Schizophr. Bull. 12, 373–393.
Hitczenko, K., Mittal, V.A., Goldrick, M., 2020. Understanding language abnormalities and associated clinical markers in psychosis: The promise of computational methods. Schizophr. Bull. .
Hoffman, M.D., Gelman, A., 2014. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15, 1593–1623.
Holmes, A.J., Patrick, L.M., 2018. The myth of optimality in clinical neuroscience. Trends Cogn. Sci. 22, 241–257.
Holt, D.J., Titone, D., Long, L.S., Goff, D.C., Cather, C., Rauch, S.L., Judge, A., Kuperberg, G.R., 2006. The misattribution of salience in delusional patients with schizophrenia. Schizophr. Res. 83, 247–256.
Hough, J., Schlangen, D., 2017. Joint, incremental disfluency detection and utterance segmentation from speech, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Valencia, Spain. pp. 326–336.
Huys, Q.J., Maia, T.V., Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat. Neurosci. 19, 404.
Jalenques, I., Enjolras, J., Izaute, M., 2013. Valence émotionnelle des mots dans la schizophrénie. Encephale 39, 189–197.
Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.
Kruschke, J., 2014. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. Academic Press, Cambridge, MA. second edition.
Lakhan, S.E., Kramer, A., 2009. Schizophrenia genomics and proteomics: Are we any closer to biomarker discovery? Behav. Brain Funct. 5, 2.
Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. J. Multivar. Anal. 100, 1989–2001.
Lindenmayer, J., Bernstein-Hyman, R., Grochowski, S., 1994. A new five factor model of schizophrenia. Psychiatr. Q. 65, 299–322.
Maia, T.V., Huys, Q.J.M., Frank, M.J., 2017. Theory-based computational psychiatry. Biol. Psychiatry 82, 382–384.
Makowski, D., Ben-Shachar, M.S., Chen, S.H.A., Lüdecke, D., 2019. Indices of effect existence and significance in the Bayesian framework. Front. Psychol. 10.
McElreath, R., 2016. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press, New York, NY.
McTeague, L.M., Huemer, J., Carreon, D.M., Jiang, Y., Eickhoff, S.B., Etkin, A., 2017. Identification of common neural circuit disruptions in cognitive control across psychiatric disorders. Am. J. Psychiatry 174, 676–685.
Mitchell, M., Hollingshead, K., Coppersmith, G., 2015. Quantifying the language of schizophrenia in social media, in: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, pp. 11–20.
Montgomery, S.A., Åsberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389.
Nagels, A., Fährmann, P., Stratmann, M., Ghazi, S., Schales, C., Frauenheim, M., Turner, L., Hornig, T., Katzev, M., Müller-Isberner, R., Grosvald, M., Krug, A., Kircher, T., 2016. Distinct neuropsychological correlates in positive and negative formal thought disorder syndromes: The Thought and Language Disorder Scale in endogenous psychoses. Neuropsychobiology 73, 139–147.
Neal, R.M., 1993. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1. University of Toronto.
Neighbors, H.W., Trierweiler, S.J., Ford, B.C., Muroff, J.R., 2003. Racial differences in DSM diagnosis using a semi-structured instrument: The importance of clinical judgment in the diagnosis of African Americans. J. Health Soc. Behav. 44, 237.
O'Hagan, A., Leonard, T., 1976. Bayes estimation subject to uncertainty about parameter constraints. Biometrika 63, 201–203.
Overall, J.E., Gorham, D.R., 1962. The Brief Psychiatric Rating Scale. Psychol. Rep. 10, 799–812.
Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K., 2015. The development and psychometric properties of LIWC2015. University of Texas at Austin, Austin, TX.
Prata, D., Mechelli, A., Kapur, S., 2014. Clinically meaningful biomarkers for psychosis: A systematic and quantitative review. Neurosci. Biobehav. Rev. 45, 134–141.
Rezaii, N., Walker, E., Wolff, P., 2019. A machine learning approach to predicting psychosis using semantic density and latent

content analysis. NPJ Schizophr. 5, 1–12.
Saykin, A.J., Shen, L., Foroud, T.M., Potkin, S.G., Swaminathan, S., Kim, S., Risacher, S.L., Nho, K., Huentelman, M.J., Craig, D.W., Thompson, P.M., Stein, J.L., Moore, J.H., Farrer, L.A., Green, R.C., Bertram, L., Jack, C.R., Weiner, M.W., 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimers Dement. 6, 265–273.
Schuurman, N.K., Ferrer, E., de Boer-Sonnenschein, M., Hamaker, E.L., 2016. How to compare cross-lagged associations in a multilevel autoregressive model. Psychol. Methods 21, 206–221.
Seymore, K., Rosenfeld, R., 1996. Scalable backoff language models, in: Proceedings of the 4th International Conference on Spoken Language Processing, IEEE, Philadelphia, PA, USA. pp. 232–235.
Shriberg, E.E., 1994. Preliminaries to a Theory of Speech Disfluencies. Doctoral dissertation. University of California at Berkeley.
Singhal, A., Yu-Yu, T., Hsia, R.Y., 2016. Racial-ethnic disparities in opioid prescriptions at emergency department visits for conditions commonly associated with prescription drug abuse. PLOS ONE 11.
Vail, A.K., Baltrušaitis, T., Pennant, L., Liebson, E., Baker, J., Morency, L.P., 2017. Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE. pp. 490–497.
Vail, A.K., Liebson, E., Baker, J.T., Morency, L.P., 2018. Toward objective, multifaceted characterization of psychotic disorders: Lexical, structural, and disfluency markers of spoken language, in: Proceedings of the International Conference on Multimodal Interaction, pp. 170–178.
van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel, B., van Harten, P.N., Valmaggia, L., de Hert, M., Cuijpers, A., et al., 2006. The five-factor model of the positive and negative syndrome scale II: a ten-fold cross-validation of a revised model. Schizophr. Res. 85, 280–287.
Vijay, S., Baltrušaitis, T., Pennant, L., Ongür, D., Baker, J.T., Morency, L.P., 2016. Computational study of psychosis symptoms and facial expressions, in: Computing and Mental Health Workshop at CHI, pp. 1–4.
Wörtwein, T., Baltrušaitis, T., Laksana, E., Pennant, L., Liebson, E.S., Öngür, D., Baker, J.T., Morency, L.P., 2017. Computational analysis of acoustic descriptors in psychotic patients, in: Interspeech 2017, ISCA. pp. 3256–3260.
Yalincetin, B., Bora, E., Binbay, T., Ulas, H., Akdede, B.B., Alptekin, K., 2017. Formal thought disorder in schizophrenia and bipolar disorder: A systematic review and meta-analysis. Schizophr. Res. 185, 2–8.
Young, R.C., Biggs, J.T., Ziegler, V.E., Meyer, D.A., 1978. A rating scale for mania: Reliability, validity and sensitivity. Br. J. Psychiatry 133, 429–435.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. 67, 301–320.

Appendix A. Predictive Modeling Details

Elastic Net Implementation

`caret` Method "glmnet" using `glmnet::glmnet()`
Mixing Percentage $\alpha \in \{0.100, 0.325, 0.550, 0.775, 1.000\}$, grid search
Regularization $\lambda \in \{0.003, 0.016, 0.074, 0.343, 1.590\}$, grid search

Support Vector Implementation

`caret` method "svmRadial" using `kernlab::ksvm()`
Regularization Parameter $C \in \{0.25, 0.50, 1.00, 2.00, 4.00\}$, grid search
Inverse Kernel Width σ analytically derived using `kernlab::sigest()`

Random Forest Implementation

`caret` method "ranger" using `ranger::ranger()`
Number of Randomly Selected Predictors $m_{try} \in \{2, 5, 8, 11, 14\}$, grid search
Splitting Rule $\in$ {"variance", "extratrees"}, grid search
Minimal Node Size = 5

Appendix B. Cross-Validation Algorithm

1. Partition ALL data into 10 outer folds* using median split stratification†
2. Loop through each outer fold one at a time, doing the following for each:
   (a) Assign the currently selected outer fold to be TESTING data
   (b) Combine the remaining 9 outer folds to be DEVELOPMENT data
   (c) Partition DEVELOPMENT data into 10 inner folds* using median split stratification†
   (d) Loop through each inner fold one at a time, doing the following for each:
      i. Assign the currently selected inner fold to be VALIDATION data
      ii. Combine the remaining 9 inner folds to be TRAINING data
      iii. Train predictive models with various hyperparameter values on TRAINING data
      iv. For each model trained in Step 2(d)iii, make predictions on the VALIDATION data and save its performance in terms of RMSE
   (e) Repeat Steps 2(c) and 2(d) three times to collect 30 total VALIDATION performance scores
   (f) Select the hyperparameters with the best VALIDATION performance using 1.5% tolerance‡
   (g) Re-train the model on all DEVELOPMENT data using the hyperparameters selected in Step 2(f)
   (h) Use the model from Step 2(g) to make predictions on the TESTING data, saving its performance
3. Repeat Steps 1 and 2 three times to collect 30 total TESTING performance scores
4. Store all TESTING performance scores for analysis

* All sessions from a participant were assigned to the same fold in both outer and inner cross-validations. In these cases, stratification was performed using the mean of the label variable across sessions.

† Median split stratification involves splitting the data points into two relatively balanced groups based on whether they are above or below the median value for the label variable.

‡ Tolerance selection involves selecting the least complex model that scored within X% of the best performance. By lightly penalizing complexity, this procedure reduces overfitting compared to selecting the best performance.

Appendix C. Predictive Model Comparison

Save the label and prediction for each data point in each repetition's testing fold. Calculate each data point's squared error. Estimate a model predicting these squared errors with nested varying intercepts to account for the dependency related to repetitions, folds, participants, and sessions. Square root the model's intercept to estimate RMSE.

Comparison Model for Between-Person Predictions

Formula	`sqerr ~ 1 + (1 | repetition/fold) + (1 | participant)`
Priors	Intercept: `rayleigh(sigma = label_between_sd^2)` Participant SDs: `student_t(nu = 3, mu = 0, sigma = sqerr_sd)` All others: default
Hyperparameters	Delta (target acceptance rate) = 0.999999 Maximum tree depth = 14 Number of Markov chains = 8 Warmup iterations per chain = 2500 Inference interations per chain = 2500
Post-Processing	`RMSE = sqrt(Intercept)` Compare RMSE posterior to Baseline (`label_between_sd`)

Comparison Model for Within-Person Predictions

Formula	`sqerr ~ 1 + (1 | repetition/fold) + (1 | participant/session)`
Priors	Intercept: `rayleigh(sigma = label_within_sd^2)` All others: default
Hyperparameters	Delta (target acceptance rate) = 0.9999 Maximum tree depth = 14 Number of Markov chains = 8 Warmup iterations per chain = 2500 Inference interations per chain = 2500
Post-Processing	`RMSE = sqrt(Intercept)` Compare RMSE posterior to Baseline (`label_within_sd`)

Appendix D. Parameterization of the Skew-Normal Distribution

The standard parameterization of the Gaussian distribution uses a mean parameter μ and a standard deviation parameter σ as follows.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

The typical parameterization of the skew-normal distribution uses a location parameter ξ, a positive scale parameter ω, and a skewness parameter α as follows (where `erf` is the error function of the Gaussian distribution).

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{y-\xi}{\omega}\right)^2\right)\left(1 + \text{erf}\left(\alpha\left(\frac{y-\xi}{\omega\sqrt{2}}\right)\right)\right)$$

To reparameterize this distibution in terms of the mean parameter μ, standard deviation parameter σ, and skewness parameter α, the ω and ξ parameters can be calculated from these values as follows.

$$\omega = \frac{\sigma}{\sqrt{1 - \frac{2}{\pi}\frac{\alpha^2}{1+\alpha^2}}}$$

$$\xi = \mu - \omega\frac{\alpha}{\sqrt{1+\alpha^2}}\sqrt{\frac{2}{\pi}}$$

Appendix E. Inferential Model Set 1

Formula	`symptom ~ 1 + sex + behavior_between_z + behavior_within_z + (1 + behavior_within_z | participant)`
Family	`skew_normal(link = "identity", link_sigma = "log", link_alpha = "identity")`
Priors	Alpha (skewness): `normal(mu = 0, sigma = 4)` Betas (slopes): `normal(mu = 0, sigma = symptom_sd)` Intercept: `student_t(nu = 3, mu = symptom_mean, sigma = symptom_sd)` Sigma (Residual SD): `student_t(nu = 3, mu = 0, sigma = symptom_sd)` Varying Effect SDs: `student_t(nu = 3, mu = 0, sigma = symptom_sd)` Varying Effect Correlations: `lkj(eta = 1)`
Hyperparameters	Delta (target acceptance rate) = 0.999 Number of Markov chains = 8 Warmup iterations per chain = 2500 Inference interations per chain = 2500
Post-Processing	Between-person slopes = `Beta / symptom_between_sd` Within-person slopes = `Beta / symptom_within_sd`

Appendix F. Inferential Model Set 2

Formula	`behavior ~ 1 + sex + positive_between_z + positive_within_z + negative_between_z + negative_within_z + (1 + positive_within_z + negative_within_z | participant)`
Family	`skew_normal(link = "identity", link_sigma = "log", link_alpha = "identity")`
Priors	Alpha (skewness): `normal(mu = 0, sigma = 4)` Betas (slopes): `normal(mu = 0, sigma = behavior_sd)` Intercept: `student_t(nu = 3, mu = behavior_mean, sigma = behavior_sd)` Sigma (Residual SD): `student_t(nu = 3, mu = 0, sigma = behavior_sd)` Varying Effect SDs: `student_t(nu = 3, mu = 0, sigma = behavior_sd)` Varying Effect Correlations: `lkj(eta = 1)`
Hyperparameters	Delta (target acceptance rate) = 0.9999 Number of Markov chains = 8 Warmup iterations per chain = 2500 Inference interations per chain = 2500
Post-Processing	Between-person slopes = `Beta / behavior_between_sd` Within-person slopes = `Beta / behavior_within_sd`