# Facial and vocal markers of schizophrenia measured using remote smartphone assessments

Isaac R. Galatzer-Levy[1,2], Anzar Abbas[1], Vidya Koesmahargyo[1], Vijay Yadav[1], M. Mercedes Perez-Rodriguez[3], Paul Rosenfield[3], Omkar Patil[4], Marissa F. Dockendorf[4], Matthew Moyer[4], Lisa A. Shipley[4] and Bryan J. Hansen[4]

[1] AiCure, New York, NY
[2] Psychiatry, New York University School of Medicine, New York, NY
[3] Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY
[4] Merck & Co., Inc., Kenilworth, NJ, USA

**Corresponding Author**
Vidya Koesmahargyo
AiCure, LLC
19 W 24th St
New York, NY 11375, United States
Tel: +1 (646)-301-5037
Email: vidya.koesmahargyo@aicure.com

## Abstract

**Background:** Machine learning-based facial and vocal measurements have demonstrated relationships with schizophrenia diagnosis and severity. Here, we determine their accuracy of when acquired through automated assessments conducted remotely through smartphones. Demonstrating utility and validity of remote and automated assessments conducted outside of controlled experimental settings can facilitate scaling such measurement tools to aid in risk assessment and tracking of treatment response in difficult to engage populations. **Methods:** Measurements of facial and vocal characteristics including facial expressivity, vocal acoustics, and speech prevalence were assessed in 20 schizophrenia patients over the course of 2 weeks in response to two classes of prompts previously utilized in experimental laboratory assessments: *evoked* prompts, where subjects are guided to produce specific facial expressions and phonations, and *spontaneous* prompts, where subjects are presented stimuli in the form of emotionally evocative imagery and asked to freely respond. Facial and vocal measurements were assessed in relation to schizophrenia symptom severity using the Positive and Negative Syndrome Scale. **Results:** Vocal markers including speech prevalence, vocal jitter, fundamental frequency, and vocal intensity demonstrated specificity as markers of negative symptom severity while measurement of facial expressivity demonstrated itself as a robust marker of overall schizophrenia severity. **Conclusion:** Established facial and vocal measurements, collected remotely in schizophrenia patients via smartphones in response to automated task prompts, demonstrated accuracy as markers of schizophrenia severity. Clinical implications are discussed.

**Keywords**: digital biomarkers; phenotyping; computer-vision; facial expressivity; negative symptoms; vocal acoustics

# 1 – Introduction

There is rapidly increasing utilization of remote digital measurements in clinical research and practice. The development and validation of digital measurement tools in psychiatry come with both significant opportunities and risks. Significant opportunity arises as psychiatry is already undergoing a paradigm shift towards the utilization of objective markers to assess illness and disease progression (Insel, 2018) and towards the widespread use of telehealth platforms for psychiatric care. This is particularly important when face-to-face medical care is not possible, such as during the COVID-19 pandemic (Figueroa & Aguilera, 2020).

Many behavioral and physiological markers are now accessible through digital technology such as wearables, mobile/web apps, and application programming interfaces (APIs; Insel, 2017). Such advances hold promise in allowing new innovations in neuropsychiatry to truly scale in a manner where they can be utilized to develop and implement treatment for patients suffering from significant psychiatric impairment (Insel, 2019).

Schizophrenia represents a poignant example of both the benefits and challenges of remote digital measurement. Clinical trials for schizophrenia drug development are often site-centric, requiring patients to appear physically at the site for measurement of disease severity. The need to travel to sites can restrict study populations to those that live in geographical proximity to the site, restricting access to participation and limiting patient diversity (Lecomte et al., 2020). Current approaches for measurement of disease rely on clinician administered measures that are costly and time-consuming to administer, leading to infrequent assessment and low adherence to treatment regimens. The instruments themselves are not well-aligned with current neurobiological definitions of illness (Torous et al., 2018). For example, cognitive and motor dysfunction, which are symptoms of the negative subtype of schizophrenia, are not accurately assessed through existing scales.

Remote digital measurements are not prone to the same level of subjectivity as human raters, and can provide both scale and ease of use while better aligning clinical measurement with behavioral and physiological measurements of underlying neurobiological treatment targets (Marsch, 2018; Torous & Keshavan, 2018). Indeed, a large number of physiological and behavioral measures have demonstrated validity as markers of schizophrenia severity or caseness. Such markers, often examined in controlled laboratory settings, hold promise as accessible remote proxies to track clinical functioning in patients with schizophrenia. However, there is a need to determine their reliability when captured in real world settings, where differentiating between significant variability and noise can pose a challenge.

A number of behavioral characteristics of schizophrenia, such as alogia (poverty of speech) and affective flattening (diminished emotional expression/emotional withdrawal; (Tandon et al., 2013) can be quantified directly using standardized tasks and coding schemes (Alberto et al., 2019; de Boer et al., 2020; Kohler, Martin, Milonova, et al., 2008; Mandal et al., 1998; Mattes et al., 1995), which can be automated through use of computer vision (Baltrusaitis et al., 2016) and vocal acoustic (Jadoul et al., 2018) machine learning models. In addition to digital measures

that are directly analogous to core schizophrenia symptomatology, there are a number of other acoustic measures including vocal loudness, pitch variability, fundamental frequency, and jitter that have demonstrated validity as markers of schizophrenia (Alberto et al., 2019; Covington et al., 2012; Martínez-Sánchez et al., 2015; Saxman & Burk, 1968). These markers have demonstrated specificity as measures of the negative symptom cluster in particular (Covington et al., 2012).

In the current investigation, we examine the ability to measure schizophrenia severity through facial and vocal analysis using videos recorded during a remote smartphone-based assessment composed of both evoked and spontaneous prompts. We compare these measures against standard clinical assessments of overall schizophrenia severity (i.e. Positive and Negative Syndrome Scale (PANSS) Total) as well as specific domains of positive (P Total), negative (N Total), and general (G Total) symptoms, measured during clinic visits. We further examine the relationship between digital measures and individual symptoms of schizophrenia.

# 2 – Methods

## 2.1 – Participants

Individuals who passed a phone-screen for a DSM-5 diagnosis of schizophrenia and were on a stable treatment regimen for atypical antipsychotic therapy for two months or more with no intent to change medication during the two-week study were recruited as study participants. A total of 20 individuals with schizophrenia were enrolled (8 males, 12 females) with an age range of 29 to 61 ($\mu$ = 45, $\sigma$ = 11). To be included in the study, participants needed to have the ability to be able to speak, read, hear, and understand the language of the clinical staff and the Informed Consent Form, respond verbally to questions, follow instructions, and be willing and be able to participate in all study activities, including the use of smartphones for data collection as described in Section 2.2. The study was conducted at the Mount Sinai Health System Outpatient Psychiatry Clinics and the protocol was approved by the Biomedical Research Alliance of New York (BRANY).

## 2.2 – Data collection

All study participants were assessed for severity of schizophrenia symptomatology using both in-person clinical assessments and remote smartphone-based assessments over the course of the 14-day observational period.

### 2.2.1 – In-person clinical assessments

The Positive and Negative Syndrome Scale (PANSS) was administered in person for all participants by clinic staff on the first (day 1) and last (day 14) of the study. For all subsequent analyses, the PANSS scores for each study participant were averaged for the two time points. Given the study participants were clinically stable, averaging the two PANSS scores allowed for

reduction in any noise in the measurement. In addition to the PANSS, all participants were assessed for negative symptomatology, with a binary yes/no determination of whether or not they were demonstrating negative symptoms of schizophrenia.

### 2.2.2 – Remote smartphone-based assessments

On the first day of the study, all study participants were trained by clinic staff on how to use the smartphone application (www.aicure.com) for remote data collection. The smartphone application was used to participate in remote assessments that would capture video and audio of participant behavior using the smartphone front-facing camera as they responded to on-screen prompts (Figure 1). Participants were allowed to use their own smartphones or use smartphones provisioned to them by the clinic staff for the duration of the study. The assessments were taken at scheduled time points over the course of the 14 days and were designed to capture two main kinds of behaviors as described below.

*Free speech and spontaneous expressivity:* Participants were shown images from the Open Affective Standard Image Set (Kurdi et al., 2017) and asked to describe the images and talk about how they made them feel (Figure 1b). The participants' speech and facial expressivity in response to the prompts were captured (Alberto et al., 2019; Cohen et al., 2016; Kohler, Martin, Milonova, et al., 2008; Kohler, Martin, Stolar, et al., 2008; Mandal et al., 1998; Mattes et al., 1995; Schwartz et al., 2006). This assessment was conducted on days 2, 7, and 14 of the study. Measurements acquired from each timepoint of the assessments were averaged for reduction of noise before comparison with PANSS.

*Evoked facial and vocal expressions:* Participants were asked separately to make the most expressive face they could and hold it for 3 seconds (Figure 1c) and then say the names of the days of the week out loud (Figure 1d). These prompts were selected based on prior experimental tasks used to examine emotional activity and speech in schizophrenia (Alpert et al., 1997; Kohler, Martin, Stolar, et al., 2008). The captured video and audio was used to measure facial expressivity and acoustic characteristics of voice during the evoked expressions. These assessments were scheduled on days 1, 7, and 14 of the study. Measurements across the time points were averaged for reduction of noise before comparison with PANSS.

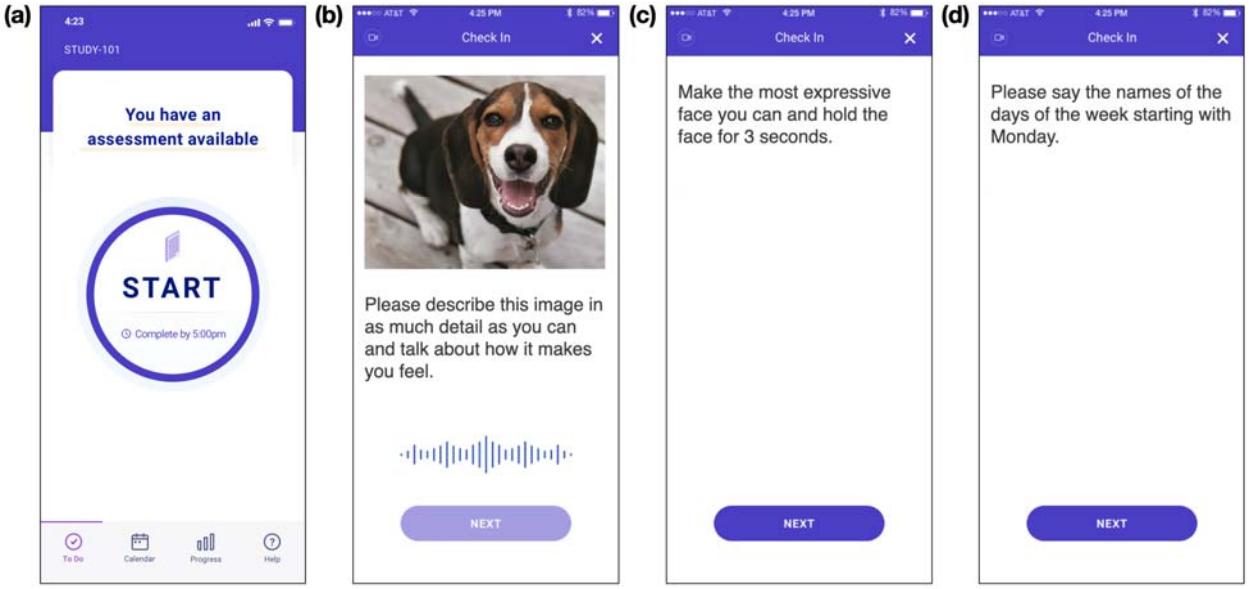


**Figure 1:** Example screenshots from the smartphone assessment all study participants took for remote and automated collection of video and audio data. During each of the prompts, the app speaks the text displayed on the screen and awaits a verbal and visual response from the participant, all while recording video and audio from the front-facing camera and microphone. **(a)** Screen displayed before the participant begins the assessment. **(b)** Prompt for collection of free behavior in response to images, showing one example image. **(c)** Prompt for collection of evoked facial expression behavior. **(d)** Prompt for collection of evoked vocal expression behavior.

## 2.3 – Measurement of digital markers

Video and audio of participant behavior collected during the remote smartphone assessments was securely uploaded for processing. A combination of computer vision and digital signal processing tools were used for quantification of facial and vocal behavior and subsequent derivation of visual and auditory markers of schizophrenia as described below. The code with which these measures were acquired have been packaged as an open source software library and made available for use to all researchers on Github: https://github.com/aicure/open_dbm.

### 2.3.1 – Measurement of facial expressivity

The software library OpenFace (https://github.com/TadasBaltrusaitis/OpenFace) was used to measure framewise facial expressivity through quantification of action units (AUs; Supplementary Table 1) using a computer vision-based implementation of the Facial Action Coding System (FACS). All framewise AU measurements were normalized through division by a timepoint-specific baseline value acquired at the beginning of each assessment when the participant is not presented with any stimulus. The normalization allows for correction of any inter- and intra-individual variability; this methodology has previously been demonstrated to be necessary for measurement of facial behavior using computer vision tools and for subsequent analyses of facial expressivity (Alvino et al., 2007; Wang et al., 2007; Wang et al., 2008). *Facial*

*expressivity* was calculated by taking the mean framewise intensity of all AUs over the course of the video. The method for quantifying *facial expressivity* was the same for both spontaneous and evoked expressivity. For each frame of video, OpenFace provides a confidence score denoting the likelihood that it is accurately detecting a face; only frames with a confidence score of 80% or higher were used for all downstream analysis. While OpenFace provides large amounts of information on specific AUs and emotions, in the current investigation, we focused only on *facial expressivity* because of significant evidence that patients with schizophrenia display less overall affect (e.g. blunted affect; (Berenbaum & Oltmanns, 1992; Henry et al., 2007).

### 2.3.2 – Measurement of vocal acoustics

The software library Parselmouth (https://github.com/YannickJadoul/Parselmouth), which is a python implementation of the Praat software library (www.praat.org), was used for measurement of all vocal acoustic characteristics. All audio analyzed was first passed through the LogMMSE noise-reduction algorithm for speech enhancement (Cannizzaro et al., 2005; Jadoul et al., 2018). Measurements of the verbal acoustic features listed in Table 1 were extracted from the audio collected. Each of the features were calculated separately during free speech and evoked vocal expression.

Despite the exploratory nature of this study and given the small data sample, we attempted to be parsimonious in selection of markers to reduce the likelihood of false discovery. Analysis of vocal markers included those that have previously demonstrated effects in studies of individuals with schizophrenia (Alberto et al., 2019; Martínez-Sánchez et al., 2015). These properties, recorded both during free speech and evoked vocal expressions, include loudness of the individual's voice in Decibels (*vocal intensity*), average fundamental frequency in Hertz (*fundamental frequency mean*), standard deviation of fundamental frequency (*fundamental frequency stdev*), jitter (*vocal jitter*), harmonics-to-noise ratio (*harmonics to noise ratio*) and the percentage of time with detected speech in an audio file (*speech prevalence)* (Cannizzaro et al., 2005; Covington et al., 2012; Kliper et al., 2019; Sarioglu Kayi et al., 2017; Saxman & Burk, 1968).

**Table 1:** List of vocal acoustic variables extracted from audio files collected during participation in remote smartphone assessments.

| Variable | Description |
|---|---|
| *Vocal intensity* | Volume of participant's speech, measured in dB |
| *Fundamental frequency mean* | Average fundamental frequency of participant speech in Hz |
| *Fundamental frequency stdev* | Standard deviation in fundamental frequency in Hz |

| | |
|---|---|
| *Vocal jitter* | Degree of irregularity in the frequency of the participant's speech, measured in Hz |
| *Speech prevalence* | Percentage of the audio file where participant speech was detected as opposed to silence |
| *Harmonics to noise ratio* | Quantification of additive noise in the participant's speech |

## 2.4 – Data analysis

Both facial expressivity and vocal characteristics were assessed during free behavior following spontaneous prompts. Facial expressivity was also assessed during evoked facial expressions and vocal characteristics were assessed during evoked vocal expression following evoked prompts. Evaluation of vocal characteristics during the evoked expression task allowed for measurement of specific characteristics that have been previously shown to be effective measures of schizophrenia during phonation (e.g. *fundamental frequency mean and stdev, jitter, harmonics-to-noise ratio*) while also measuring speech characteristics such as amount of time spoken (i.e. *speech prevalence*) (Cannizzaro et al., 2005; Covington et al., 2012; Kliper et al., 2019; Sarioglu Kayi et al., 2017; Saxman & Burk, 1968). A large number of variables can be calculated from video and audio data sources; however, the analyses presented herein were limited to features that have evidence and a theoretical basis for relationship to schizophrenia severity and symptoms in the scientific literature.

**Table 2:** All variables described in Section 2.3 were calculated separately for distinct behaviors captured during the remote smartphone assessments. Each of the behaviors that were elicited and captured during the smartphone assessment and the digital markers calculated from those behaviors are listed here.

| Behavior | On-screen prompt | Digital markers measured |
|---|---|---|
| Free behavior | *Please describe what you see in this image and talk about how it makes you feel* (Figure 1b) | *Facial expressivity*<br>*Fundamental frequency mean*<br>*Fundamental frequency stdev*<br>*Vocal Jitter*<br>*Harmonics-to-noise ratio*<br>*Speech prevalence* |
| Evoked facial expression | *Please make the most expressive face you can and hold it for 3 second*s (Figure 1c) | *Facial expressivity* |

| | | |
|---|---|---|
| Evoked vocal expression | *Please say the names of the days of the week starting with Monday* (Figure 1d) | *Fundamental frequency mean*<br>*Fundamental frequency stdev*<br>*Vocal Jitter*<br>*Harmonics-to-noise ratio*<br>*Speech prevalence* |

#### 2.4.1 – Comparison to PANSS subscale scores

Digital measures were compared to schizophrenia severity overall using the PANSS total severity score (*PANSS Total*) along with the three subscales reflecting negative symptom severity (*N Total*), positive symptom severity (*P Total*), and general severity (*G Total*) using Pearson's correlation. When comparing negative symptoms, we utilized the PANSS Marder Symptom Factor which includes two symptoms that are traditionally included in the general severity score: *Motor Retardation* and *Social Avoidance and Isolation*.

#### 2.4.2 – Comparison to individual PANSS items

Digital measurements that demonstrated significance in relation to specific subscales were then further explored in relation to the specific symptoms that derive those subscales, correcting for multiple comparisons using a Benjamini-Hochberg adjusted p-value (Li & Barber, 2019). This was an exploratory analysis conducted to further disaggregate the heterogeneity within the symptom scales to understand more specifically which clinical features were reflected in the digital measurement.

# 3 – Results

## 3.1 – Comparison to PANSS scores

### 3.1.1 – Vocal markers during evoked vocal expression

Results demonstrate that multiple digital measures are significantly correlated with overall negative symptom severity (N Total) after correcting for multiple comparisons. This includes *fundamental frequency mean* ($r$ = -0.64; *adjusted* $p$ = .02), *vocal jitter* ($r$ = 0.56; *adjusted* $p$ = .02), and *harmonics to noise ratio* ($r$ = -0.61; *adjusted* $p$ = .02). Two other features demonstrated marginal significance after correction for false discovery, including *speech prevalence* ($r$ = -0.47; *adjusted* $p$ = .06) and *fundamental frequency stdev* ($r$ = -0.44; *adjusted* $p$ = .07; see Table 3 for full results). Importantly, the directionality of results was consistent with prior research. For example, increased negative symptom severity was reflected in decreased speech prevalence, decreased tonal qualities of speech, and increased noise to speech sounds, consistent with the literature (Alberto et al., 2019; Covington et al., 2012; Martínez-Sánchez et al., 2015; Saxman & Burk, 1968).

**Table 3:** Correlation between vocal markers during evoked vocal expression and PANSS score showed a relationship between vocal characteristics and schizophrenia severity.

| Variable | | N Total | P Total | G Total | Total | Vocal intensity | Fundamental frequency stdev | Fundamental frequency mean | Vocal jitter | Speech prevalence |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. N Total | Pearson's r | — | | | | | | | | |
| | p-value | — | | | | | | | | |
| 2. P Total | Pearson's r | 0.452* | — | | | | | | | |
| | p-value | 0.045 | — | | | | | | | |
| 3. G Total | Pearson's r | 0.572** | 0.806*** | — | | | | | | |
| | p-value | 0.008 | < .001 | — | | | | | | |
| 4. Total | Pearson's r | 0.757*** | 0.870*** | 0.947*** | — | | | | | |
| | p-value | < .001 | < .001 | < .001 | — | | | | | |
| 5. Vocal intensity | Pearson's r | -0.091 | -0.250 | -0.088 | -0.152 | — | | | | |
| | p-value | 0.710 | 0.903 | 0.720 | 0.642 | — | | | | |
| 6. Fundamental frequency stdev | Pearson's r | -0.436 | -0.068 | 0.098 | -0.090 | -0.081 | — | | | |
| | p-value | 0.074 | 0.782 | 0.827 | 0.714 | 0.743 | — | | | |
| 7. Fundamental frequency mean | Pearson's r | -0.644* | -0.253 | -0.218 | -0.373 | 0.475 | 0.577* | — | | |
| | p-value | 0.018 | 0.296 | 0.371 | 0.696 | 0.1 | 0.020 | — | | |
| 8. Vocal jitter | Pearson's r | 0.563* | 0.229 | 0.122 | 0.293 | -0.176 | -0.695*** | -0.823*** | — | |
| | p-value | 0.024 | 0.519 | 0.928 | 0.336 | 0.786 | < .001 | < .001 | — | |
| 9. Speech prevalence | Pearson's r | -0.470 | -0.247 | -0.292 | -0.362 | 0.611* | 0.043 | 0.781*** | -0.373 | — |
| | p-value | 0.063 | 0.614 | 0.225 | 0.381 | 0.025 | 0.863 | < .001 | 0.116 | — |
| 10. Harmonics to noise ratio | Pearson's r | -0.610* | -0.195 | -0.126 | -0.297 | 0.154 | 0.773*** | 0.868*** | -0.965*** | 0.422 |
| | p-value | 0.018 | 0.507 | 0.606 | 0.434 | 0.66 | < .001 | < .001 | < .001 | 0.072 |

* p < 0.05; ** p < 0.01; *** p < 0.001

Upon examination of the relationship between vocal measures and individual negative symptoms, results demonstrated that specific symptoms including *Emotional Withdrawal (r = -0.55 ; p = .015)*, *Poor Rapport (r = -0.59 ; p = .008)*, *Lack of Spontaneity (r = -0.59; p = .008* , and *Motor Retardation (r = -0.53; p = .02)* demonstrated significant correlations with multiple auditory measures in a direction consistent with their relationship to overall negative symptom severity (Supplementary Table 2).

### 3.1.2 – Evoked facial expression

Results demonstrated that *facial expressivity* demonstrated significant relationships with overall schizophrenia severity PANSS Total (*r* = -0.71; *adjusted p* = .002) and severity on all PANSS subscales (N Total, *r* = -0.50; *adjusted p* = .035; P Total, *r* = -0.63; *adjusted p* = .006; G Total, *r* = -0.70; *adjusted p* = .009). See Table 4.

**Table 4:** Correlation between *facial expressivity* during evoked facial expression and PANSS score showed a relationship between facial affect and schizophrenia severity.

| Variable | | Facial expressivity | N Total | P Total | G Total |
|---|---|---|---|---|---|
| 1. Facial expressivity | Pearson's r | — | | | |
| | p-value | — | | | |
| 2. N Total | Pearson's r | -0.500* | — | | |
| | p-value | 0.035 | — | | |
| 3. P Total | Pearson's r | -0.628** | 0.452* | — | |
| | p-value | 0.010 | 0.045 | — | |
| 4. G Total | Pearson's r | -0.695** | 0.572** | 0.806*** | — |
| | p-value | 0.009 | 0.008 | < .001 | — |
| 5. Total | Pearson's r | -0.714** | 0.757*** | 0.870*** | 0.947*** |
| | p-value | 0.002 | < .001 | < .001 | < .001 |

* p < 0.05; ** p < 0.01; *** p < 0.001

Pearson's correlations with individual symptoms demonstrates significant or marginally significant relationships with multiple individual negative symptoms including *Blunted Affect* ($r$ = -0.49; $p$ = .04), *Social Withdrawal* ($r$ = -0.63; $p$ = .005), *Stereotyped Thinking* ($r$ = -0.41; $p$ = .09), and *Motor Retardation* ($r$ = -0.45; $p$ = .06). Further, facial expressivity was significantly negatively correlated with positive symptoms including *Delusions* ($r$ = -0.81; $p$ < .001), *Hallucinations* ($r$ = -0.50; $p$ = .04), and *Paranoid Ideation* ($r$ = -0.64; $p$ = .005). Finally, *facial expressivity* demonstrated significant negative correlations with symptoms in the general symptom factor including *Guilt* ($r$ = -0.50; $p$ = .035), *Tension* ($r$ = -0.40; $p$ = .10), *Unusual Thought Content* ($r$ = -0.65; $p$ = .02), *Disturbances in Volition* ($r$ = -0.78; $p$ < .001), and *Social Avoidance* ($r$ = -0.79; $p$ < .001). See Supplementary Table 3 for complete results.

### 3.1.3 – Free behavior in response to images

Spontaneous measurement of voice and facial expressions, as elicited by emotionally valenced images, demonstrated relationships between multiple vocal markers and the negative symptom cluster. Highly consistent with results of vocal measurements in response to evoked prompts, the following measures demonstrated significance in relation to negative symptom severity (N Tota)l: *fundamental frequency mean* ($r$ = -0.61; *adjusted* $p$ = .04), *harmonics to noise ratio* ($r$ = -0.58; *adjusted* $p$ = .03), *speech prevalence* ($r$ = -0.57; *adjusted* $p$ = .025). *Vocal jitter* demonstrated a marginally significant adjusted p-value ($r$ = 0.43; *adjusted* $p$ = .09), and *fundamental frequency stdev* did not approach significance (See Table 5). In contrast to measurement after the evoked task, *vocal intensity* measured during free behavior did demonstrate significance ($r$ = 0.50; *adjusted* $p$ = .05).

**Table 5:** Correlation between facial and vocal markers during free behavior and PANSS score showed a relationship between facial affect and vocal characteristics with schizophrenia severity.

| Variable | | N Total | P Total | G Total | Total | Facial expressivity | Vocal intensity | Fundamental frequency mean | Fundamental frequency Stdev | Harmonics to noise ratio | Vocal jitter |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. N Total | Pearson's r | — | | | | | | | | | |
| | p-value | — | | | | | | | | | |
| 2. P Total | Pearson's r | 0.452* | — | | | | | | | | |
| | p-value | 0.045 | — | | | | | | | | |
| 3. G Total | Pearson's r | 0.572** | 0.806*** | — | | | | | | | |
| | p-value | 0.008 | < .001 | — | | | | | | | |
| 4. Total | Pearson's r | 0.757*** | 0.870*** | 0.947*** | — | | | | | | |
| | p-value | < .001 | < .001 | < .001 | — | | | | | | |
| 5. Facial expressivity | Pearson's r | 0.142 | -0.113 | 0.090 | 0.056 | — | | | | | |
| | p-value | 0.563 | 0.644 | 0.834 | 0.819 | — | | | | | |
| 6. Vocal intensity | Pearson's r | -0.502* | -0.332 | -0.225 | -0.386 | 0.364 | — | | | | |
| | p-value | 0.050 | 0.165 | 0.826 | 0.240 | 0.126 | — | | | | |
| 7. Fundamental frequency mean | Pearson's r | -0.606* | -0.288 | -0.268 | -0.428 | 0.184 | 0.935*** | — | | | |
| | p-value | 0.042 | 0.812 | 0.268 | 0.476 | 0.451 | < .001 | — | | | |
| 8. Fundamental frequency stdev | Pearson's r | -0.304 | -0.189 | -0.127 | -0.225 | 0.179 | 0.581** | 0.529* | — | | |
| | p-value | 0.240 | 0.613 | 0.605 | 0.495 | 0.464 | 0.009 | 0.020 | — | | |
| 9. Harmonics to noise ratio | Pearson's r | -0.584* | -0.224 | -0.097 | -0.312 | 0.174 | 0.654** | 0.774*** | 0.476* | — | |
| | p-value | 0.031 | 0.624 | 0.970 | 0.339 | 0.475 | 0.002 | < .001 | 0.039 | — | |
| 10. Vocal jitter | Pearson's r | 0.426 | 0.147 | 0.015 | 0.194 | -0.097 | -0.541* | -0.691** | -0.278 | -0.937*** | — |
| | p-value | 0.096 | 0.640 | 0.951 | 0.498 | 0.692 | 0.017 | 0.001 | 0.249 | < .001 | — |
| 11. Speech prevalence | Pearson's r | -0.567* | -0.260 | -0.261 | -0.403 | 0.161 | 0.869*** | 0.923*** | 0.260 | 0.575** | -0.510* |
| | p-value | 0.025 | 0.660 | 0.980 | 0.304 | 0.510 | < .001 | < .001 | 0.283 | 0.010 | 0.026 |

* p < 0.05; ** p < 0.01; *** p < 0.001

Individual negative symptoms broadly demonstrated relationships with multiple vocal markers (See Supplementary Table 4 for complete results). *Blunted affect* only demonstrated a marginally significant relationship with *speech prevalence* ($r$ = 0.40; $p$ = .09) while *vocal intensity* only demonstrated a significant relationship with the symptom of *Emotional Withdrawal* ($r$ = -0.53; $p$ = .02).

## 4 – Discussion

In the current investigation, we sought to test the hypothesis that facial and vocal markers of schizophrenia can be captured remotely in patients using brief automated smartphone-based assessments and that such measures would be well-correlated to standard clinical measures of schizophrenia symptom severity. Such measures show promise of objective and automated methods of assessing illness severity in the context of treatment development and decision making. Prompts and vocal/facial measures that have previously demonstrated accuracy in controlled research settings were simplified and deployed as a brief assessment via a smartphone application in an observational study with schizophrenia patients. Results support the ability to measure meaningful clinical markers of schizophrenia severity via a brief smartphone based assessment that captures data remotely and processes it through back-end deep machine learning algorithms to create vocal and facial markers.

Results demonstrate that vocal characteristics such as fundamental frequency, loudness, non-verbal vocal tones and prevalence of speech serve as specific markers of negative symptom severity. The majority of these markers demonstrate a robust signal of negative symptom severity regardless of whether prompts were evoked or spontaneous.

The observation that vocal markers provide specificity as a metric of negative symptom severity has significant practical implications for clinical research and decision making. Recent advances in the mechanistic understanding of negative symptomatology have led to a number of promising pharmacological and cognitive treatments for negative symptoms of schizophrenia (Erhart et al., 2006; Fusar-Poli et al., 2015; Millan et al., 2014; Singh et al., 2010). Such initiatives are important given the lack of FDA-approved treatments for negative symptoms (Kirkpatrick et al., 2006). However, reliable and change-sensitive measures of negative symptomatology to assess the efficacy of these treatments are sparse (King, 1998; Möller, 2007; Prikryl et al., 2007; Walther et al., 2009).

Facial expressivity only demonstrated a relationship with schizophrenia severity when captured using evoked prompts. This may indicate that either greater structure is needed to assess this marker remotely or that the prompts that were utilized were not a strong enough elicitation. Indeed, prior work has demonstrated that video rather than still images are stronger evocations to assess emotional variability in schizophrenia (Bersani et al., 2013). Despite this, we do observe that facial expressivity in response to evoked prompts provides a robust signal for overall symptom severity. Analysis of single symptoms demonstrates face validity for this marker as it reduced facial expressivity relates to greater blunted affect, social withdrawal, social avoidance, and motor retardation. We also observe that facial expressivity was decreased as subjects endorsed greater positive symptoms of hallucinations, delusions, and paranoid ideation. This finding is consistent with evidence that social and cognitive impairment, which falls into the negative symptom cluster, mechanistically relates to neurodegeneration that also impacts visual and auditory hallucinations (Bersani et al., 2013; Jenkins et al., 2018; Zhuo et al., 2020).

The current study presents with a number of important limitations. While the primary hypotheses were supported, not all effects were consistent across prompts. Given the small sample size, it is impossible to conclude definitively which markers can be utilized to robustly assess schizophrenia symptom severity or impairment. Indeed, a number of relatively large correlation coefficients demonstrated only marginal significance, likely due to sample size constraints. Further, despite the markers being hypothesized a priori, the current work is exploratory in nature given the small sample size, limited number of assessments, and the short duration of the study. A larger assessment will be needed to replicate the current findings. Despite the above limitations, the current work provides evidence that facial and vocal digital measures can be remotely captured in schizophrenia patients and that such measures demonstrate statistically significant relationships with established measures of schizophrenia symptom severity, demonstrating promise that these tools could be used to remotely measure and track schizophrenia symptoms and severity in an objective manner.

Finally, while the app-based video/audio capture utilizes a proprietary platform, this investigation utilized open-source Python-based software, available to all researchers (https://github.com/AiCure/open_dbm/). As an additional measure, the code that implemented the open-source software for this investigation and subsequent analyses of results have been provided by the authors in the Methods section. This allows for the expansion of the experiment to a wider patient population as mentioned above and the independent validation of the methods and their implementation in this investigation by other researchers in academic and clinical research, following an open-science framework for the development of digital tools for objective, accurate, and scalable measurement of disease symptomatology in both mental and physical health.

## 5 – Conclusions

In this investigation, we demonstrate that facial and vocal markers, measured using computer vision and vocal analytics from video captured remotely via smartphones demonstrates validity as a marker of schizophrenia and is a promising metric for negative symptom severity. Use of such technology in clinical care and clinical research settings could allow for more frequent, remotely assessed, objective measurement of disease symptomatology and treatment response in a scalable and accessible manner, which can support development of novel treatments and risk assessment among individuals with schizophrenia.

## Acknowledgments

The authors appreciate the involvement of the clinical, research, and operations staff at both Mount Sinai and AiCure for the development, deployment, and implementation of the technology presented here and the participants who volunteered to be involved in the research.

## Declaration of Interest

Authors IGL, AA, VY and VK were employed and own shares at AiCure, LLC at the time of the study. Authors OP, MD, MM, LS, and BH are employees of Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA and may own stock/stock options in Merck & Co., Inc., Kenilworth, NJ, USA.

## References


Alberto, P., Arndis, S., Vibeke, B., & Riccardo, F. (2019). *Voice Patterns in Schizophrenia: A systematic Review and Bayesian Meta-Analysis* [Preprint]. Bioinformatics. https://doi.org/10.1101/583815

Alpert, M., Kotsaftis, A., & Pouget, E. R. (1997). Speech Fluency and Schizophrenic Negative Signs. *Schizophrenia Bulletin*, *23*(2), 171–177. https://doi.org/10.1093/schbul/23.2.171

Alvino, C., Kohler, C., Barrett, F., Gur, R. E., Gur, R. C., & Verma, R. (2007). Computerized measurement of facial expression of emotions in schizophrenia. *Journal of Neuroscience Methods*, *163*(2), 350–361. https://doi.org/10.1016/j.jneumeth.2007.03.002

Baltrusaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. https://doi.org/10.1109/WACV.2016.7477553

Berenbaum, H., & Oltmanns, T. F. (1992). Emotional experience and expression in schizophrenia and depression. *Journal of Abnormal Psychology*, *101*(1), 37–44. https://doi.org/10.1037//0021-843x.101.1.37

Bersani, G., Polli, E., Valeriani, G., Zullo, D., Melcore, C., Capra, E., Quartini, A., Marino, P., Minichino, A., Bernabei, L., Robiony, M., Bersani, F. S., & Liberati, D. (2013). Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders. *Neuropsychiatric Disease and Treatment*, *9*.

Cannizzaro, M. S., Cohen, H., Rappard, F., & Snyder, P. J. (2005). Bradyphrenia and Bradykinesia Both Contribute to Altered Speech in Schizophrenia: A Quantitative Acoustic Study. *Cognitive and Behavioral Neurology*, *18*(4), 206–210. https://doi.org/10.1097/01.wnn.0000185278.21352.e5

Cohen, A. S., Mitchell, K. R., Docherty, N. M., & Horan, W. P. (2016). Vocal expression in schizophrenia: Less than meets the ear. *Journal of Abnormal Psychology*, *125*(2), 299–309. https://doi.org/10.1037/abn0000136

Covington, M. A., Lunden, S. L. A., Cristofaro, S. L., Wan, C. R., Bailey, C. T., Broussard, B., Fogarty, R., Johnson, S., Zhang, S., & Compton, M. T. (2012). Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia Research*, *142*(1–3), 93–95. https://doi.org/10.1016/j.schres.2012.10.005

de Boer, J. N., van Hoogdalem, M., Mandl, R. C. W., Brummelman, J., Voppel, A. E., Begemann, M. J. H., van Dellen, E., Wijnen, F. N. K., & Sommer, I. E. C. (2020). Language in schizophrenia: Relation with diagnosis, symptomatology and white matter tracts. *Npj Schizophrenia*, *6*(1), 1–10. https://doi.org/10.1038/s41537-020-0099-3

Erhart, S. M., Marder, S. R., & Carpenter, W. T. (2006). Treatment of schizophrenia negative symptoms: Future prospects. *Schizophrenia Bulletin*, *32*(2), 234–237. https://doi.org/10.1093/schbul/sbj055

Figueroa, C. A., & Aguilera, A. (2020). The Need for a Mental Health Technology Revolution in the COVID-19 Pandemic. *Frontiers in Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.00523

Fusar-Poli, P., Papanastasiou, E., Stahl, D., Rocchetti, M., Carpenter, W., Shergill, S., &

McGuire, P. (2015). Treatments of Negative Symptoms in Schizophrenia: Meta-Analysis of 168 Randomized Placebo-Controlled Trials. *Schizophrenia Bulletin*, *41*(4), 892–899. https://doi.org/10.1093/schbul/sbu170

Henry, J. D., Green, M. J., de Lucia, A., Restuccia, C., McDonald, S., & O'Donnell, M. (2007). Emotion dysregulation in schizophrenia: Reduced amplification of emotional expression is associated with emotional blunting. *Schizophrenia Research*, *95*(1–3), 197–204. https://doi.org/10.1016/j.schres.2007.06.002

Insel, T. R. (2017). Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*, *318*(13), 1215. https://doi.org/10.1001/jama.2017.11295

Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry*, *17*(3), 276–277. https://doi.org/10.1002/wps.20550

Insel, T. R. (2019). Bending the Curve for Mental Health: Technology for a Public Health Approach. *American Journal of Public Health*, *109*(Suppl 3), S168–S170. https://doi.org/10.2105/AJPH.2019.305077

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Jenkins, L. M., Bodapati, A. S., Sharma, R. P., & Rosen, C. (2018). Working memory predicts presence of auditory verbal hallucinations in schizophrenia and bipolar disorder with psychosis. *Journal of Clinical and Experimental Neuropsychology*, *40*(1), 84–94. https://doi.org/10.1080/13803395.2017.1321106

King, D. J. (1998). Drug treatment of the negative symptoms of schizophrenia. *European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology*, *8*(1), 33–42. https://doi.org/10.1016/s0924-977x(97)00041-2

Kirkpatrick, B., Fenton, W. S., Carpenter, W. T., & Marder, S. R. (2006). The NIMH-MATRICS Consensus Statement on Negative Symptoms. *Schizophrenia Bulletin*, *32*(2), 214–219. https://doi.org/10.1093/schbul/sbj053

Kliper, R., Vaizman, Y., Weinshall, D., & Portuguese, S. (2019). *Evidence for depression and schizophrenia in speech prosody*. 85–88. https://doi.org/10.36505/ExLing-2010/03/0022/000142

Kohler, C. G., Martin, E. A., Milonova, M., Wang, P., Verma, R., Brensinger, C. M., Bilker, W., Gur, R. E., & Gur, R. C. (2008). Dynamic evoked facial expressions of emotions in schizophrenia. *Schizophrenia Research*, *105*(1–3), 30–39. https://doi.org/10.1016/j.schres.2008.05.030

Kohler, C. G., Martin, E. A., Stolar, N., Barrett, F. S., Verma, R., Brensinger, C., Bilker, W., Gur, R. E., & Gur, R. C. (2008). Static posed and evoked facial expressions of emotions in schizophrenia. *Schizophrenia Research*, *105*(1), 49–60. https://doi.org/10.1016/j.schres.2008.05.010

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Lecomte, T., Potvin, S., Corbière, M., Guay, S., Samson, C., Cloutier, B., Francoeur, A., Pennou, A., & Khazaal, Y. (2020). Mobile Apps for Mental Health Issues: Meta-Review of Meta-Analyses. *JMIR MHealth and UHealth*, *8*(5), e17458.

https://doi.org/10.2196/17458
Mandal, M. K., Pandey, R., & Prasad, A. B. (1998). Facial expressions of emotions and schizophrenia: A review. *Schizophrenia Bulletin*, *24*(3), 399–412. https://doi.org/10.1093/oxfordjournals.schbul.a033335
Marsch, L. A. (2018). Opportunities and needs in digital phenotyping. *Neuropsychopharmacology*, *43*(8), 1637–1638. https://doi.org/10.1038/s41386-018-0051-7
Martínez-Sánchez, F., Muela-Martínez, J. A., Cortés-Soto, P., García Meilán, J. J., Vera Ferrándiz, J. A., Egea Caparrós, A., & Pujante Valverde, I. M. (2015). Can the Acoustic Analysis of Expressive Prosody Discriminate Schizophrenia? *The Spanish Journal of Psychology*, *18*, E86. https://doi.org/10.1017/sjp.2015.85
Mattes, R. M., Schneider, F., Heimann, H., & Birbaumer, N. (1995). Reduced emotional response of schizophrenic patients in remission during social interaction. *Schizophrenia Research*, *17*(3), 249–255. https://doi.org/10.1016/0920-9964(95)00014-3
Millan, M. J., Fone, K., Steckler, T., & Horan, W. P. (2014). Negative symptoms of schizophrenia: Clinical characteristics, pathophysiological substrates, experimental models and prospects for improved treatment. *European Neuropsychopharmacology: The Journal of the European College of Neuropsychopharmacology*, *24*(5), 645–692. https://doi.org/10.1016/j.euroneuro.2014.03.008
Möller, H.-J. (2007). Clinical evaluation of negative symptoms in schizophrenia. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *22*(6), 380–386. https://doi.org/10.1016/j.eurpsy.2007.03.010
Prikryl, R., Kasparek, T., Skotakova, S., Ustohal, L., Kucerova, H., & Ceskova, E. (2007). Treatment of negative symptoms of schizophrenia using repetitive transcranial magnetic stimulation in a double-blind, randomized controlled study. *Schizophrenia Research*, *95*(1–3), 151–157. https://doi.org/10.1016/j.schres.2007.06.019
Sarioglu Kayi, E., Diab, M., Pauselli, L., Compton, M., & Coppersmith, G. (2017). Predictive Linguistic Features of Schizophrenia. *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, 241–250. https://doi.org/10.18653/v1/S17-1028
Saxman, J. H., & Burk, K. W. (1968). Speaking Fundamental Frequency and Rate Characteristics of Adult Female Schizophrenics. *Journal of Speech and Hearing Research*, *11*(1), 194–203. https://doi.org/10.1044/jshr.1101.194
Schwartz, B. L., Mastropaolo, J., Rosse, R. B., Mathis, G., & Deutsch, S. I. (2006). Imitation of facial expressions in schizophrenia. *Psychiatry Research*, *145*(2–3), 87–94. https://doi.org/10.1016/j.psychres.2005.12.007
Singh, S. P., Singh, V., Kar, N., & Chan, K. (2010). Efficacy of antidepressants in treating the negative symptoms of chronic schizophrenia: Meta-analysis. *The British Journal of Psychiatry: The Journal of Mental Science*, *197*(3), 174–179. https://doi.org/10.1192/bjp.bp.109.067710
Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., Malaspina, D., Owen, M. J., Schultz, S., Tsuang, M., Van Os, J., & Carpenter, W. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research*, *150*(1), 3–10. https://doi.org/10.1016/j.schres.2013.05.028

Torous, J., & Keshavan, M. (2018). A new window into psychosis: The rise digital phenotyping, smartphone assessment, and mobile monitoring. *Schizophrenia Research*, *197*, 67–68. https://doi.org/10.1016/j.schres.2018.01.005

Torous, J., Staples, P., Barnett, I., Sandoval, L. R., Keshavan, M., & Onnela, J.-P. (2018). Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ Digital Medicine*, *1*, 15. https://doi.org/10.1038/s41746-018-0022-8

Walther, S., Koschorke, P., Horn, H., & Strik, W. (2009). Objectively measured motor activity in schizophrenia challenges the validity of expert ratings. *Psychiatry Research*, *169*(3), 187–190. https://doi.org/10.1016/j.psychres.2008.06.020

Wang, P., Kohler, C., Barrett, F., Gur, R., Gur, R., & Verma, R. (2007). Quantifying Facial Expression Abnormality in Schizophrenia by Combining 2D and 3D Features. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/CVPR.2007.383061

Wang, Peng, Barrett, F., Martin, E., Milanova, M., Gur, R. E., Gur, R. C., Kohler, C., & Verma, R. (2008). Automated Video Based Facial Expression Analysis of Neuropsychiatric Disorders. *Journal of Neuroscience Methods*, *168*(1), 224–238. https://doi.org/10.1016/j.jneumeth.2007.09.030

Zhuo, C., Chen, M., Xu, Y., Jiang, D., Chen, C., Ma, X., Li, R., Sun, Y., Li, Q., Zhou, C., & Lin, X. (2020). Reciprocal deterioration of visual and auditory hallucinations in schizophrenia presents V-shaped cognition impairment and widespread reduction in brain gray matter-A pilot study. *Journal of Clinical Neuroscience*, *79*, 154–159. https://doi.org/10.1016/j.jocn.2020.07.054

## Supplementary Materials

**Supplementary Table 1:** List of facial action units (AUs) whose frame-wise intensity was quantified using computer vision; AU intensities were normalized and then combined to measure *facial expressivity*.

| Action Unit | Description |
|---|---|
| AU1 | Inner brow raiser |
| AU2 | Outer brow raiser |
| AU4 | Brow lowerer |
| AU5 | Upper lid raiser |
| AU6 | Cheek raiser |
| AU7 | Lid tightener |
| AU9 | Nose wrinkler |
| AU12 | Lip corner puller |
| AU15 | Lip corner depressor |
| AU16 | Lower lip depressor |
| AU20 | Lip stretcher |
| AU23 | Lip tightener |
| AU26 | Jaw drop |

**Supplementary Table 2:** Correlation between vocal markers during evoked vocal expression and PANSS individual items.

| Variable | | Blunted Affect | Emotional Withdrawal | Poor Rapport | Social Withdrawal | Abstract Thinking | Lack of Spontaneity | Stereotyped Thinking | Motor Retardation | Social Avoidance | Vocal intensity | Fundamental frequency skew | Fundamental frequency mean | Vocal jitter | Speech prevalence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Blunted Affect | Pearson's r | — | | | | | | | | | | | | | |
| | p-value | — | | | | | | | | | | | | | |
| 2. Emotional Withdrawal | Pearson's r | 0.454* | — | | | | | | | | | | | | |
| | p-value | 0.044 | — | | | | | | | | | | | | |
| 3. Poor Rapport | Pearson's r | 0.598** | 0.647** | — | | | | | | | | | | | |
| | p-value | 0.005 | 0.002 | — | | | | | | | | | | | |
| 4. Social Withdrawal | Pearson's r | 0.453* | 0.765*** | 0.508* | — | | | | | | | | | | |
| | p-value | 0.045 | < .001 | 0.022 | — | | | | | | | | | | |
| 5. Abstract Thinking | Pearson's r | 0.103 | 0.451* | 0.347 | 0.457* | — | | | | | | | | | |
| | p-value | 0.665 | 0.046 | 0.134 | 0.043 | — | | | | | | | | | |
| 6. Lack of Spontaneity | Pearson's r | 0.605** | 0.635** | 0.721*** | 0.595** | 0.466* | — | | | | | | | | |
| | p-value | 0.005 | 0.003 | < .001 | 0.006 | 0.038 | — | | | | | | | | |
| 7. Stereotyped Thinking | Pearson's r | 0.524* | 0.660** | 0.714*** | 0.818*** | 0.570** | 0.717*** | — | | | | | | | |
| | p-value | 0.018 | 0.002 | < .001 | < .001 | 0.009 | < .001 | — | | | | | | | |
| 8. Motor Retardation | Pearson's r | 0.714*** | 0.663** | 0.690*** | 0.702*** | 0.346 | 0.720*** | 0.601** | — | | | | | | |
| | p-value | < .001 | 0.001 | < .001 | < .001 | 0.136 | < .001 | 0.005 | — | | | | | | |
| 9. Social Avoidance | Pearson's r | 0.385 | 0.580** | 0.233 | 0.804*** | 0.133 | 0.182 | 0.528* | 0.451* | — | | | | | |
| | p-value | 0.093 | 0.007 | 0.322 | < .001 | 0.575 | 0.443 | 0.017 | 0.046 | — | | | | | |
| 10. Vocal intensity | Pearson's r | -0.237 | -0.040 | 0.004 | -0.204 | 0.269 | -0.083 | -0.094 | -0.200 | -0.152 | — | | | | |
| | p-value | 0.330 | 0.872 | 0.986 | 0.401 | 0.266 | 0.735 | 0.702 | 0.413 | 0.533 | — | | | | |
| 11. Fundamental frequency skew | Pearson's r | -0.144 | -0.549* | -0.591** | -0.092 | -0.227 | -0.450 | -0.197 | -0.408 | 0.035 | -0.081 | — | | | |
| | p-value | 0.558 | 0.015 | 0.008 | 0.707 | 0.349 | 0.053 | 0.419 | 0.083 | 0.888 | 0.743 | — | | | |
| 12. Fundamental frequency mean | Pearson's r | -0.518* | -0.500* | -0.538* | -0.381 | -0.268 | -0.590** | -0.429 | -0.527* | -0.308 | 0.470* | 0.577** | — | | |
| | p-value | 0.023 | 0.029 | 0.017 | 0.107 | 0.268 | 0.008 | 0.067 | 0.020 | 0.200 | 0.040 | 0.010 | — | | |
| 13. Vocal jitter | Pearson's r | 0.383 | 0.557* | 0.573* | 0.333 | 0.227 | 0.485* | 0.258 | 0.513* | 0.284 | -0.176 | -0.695*** | -0.822*** | — | |
| | p-value | 0.106 | 0.013 | 0.010 | 0.164 | 0.351 | 0.035 | 0.286 | 0.025 | 0.239 | 0.472 | < .001 | < .001 | — | |
| 14. Speech prevalence | Pearson's r | -0.520* | -0.235 | -0.232 | -0.356 | -0.076 | -0.475* | -0.418 | -0.308 | -0.313 | 0.611** | 0.043 | 0.781*** | -0.373 | — |
| | p-value | 0.023 | 0.332 | 0.339 | 0.134 | 0.752 | 0.040 | 0.075 | 0.200 | 0.191 | 0.005 | 0.863 | < .001 | 0.116 | — |
| 15. Harmonics to noise ratio | Pearson's r | -0.376 | -0.603** | -0.624** | -0.327 | -0.334 | -0.504* | -0.300 | -0.512* | -0.265 | 0.154 | 0.773*** | 0.868*** | -0.965*** | 0.422 |
| | p-value | 0.113 | 0.006 | 0.004 | 0.172 | 0.162 | 0.028 | 0.212 | 0.025 | 0.272 | 0.528 | < .001 | < .001 | < .001 | 0.072 |

* p < .05, ** p < .01, *** p < .001

**Supplementary Table 3:** Correlation between facial markers during evoked facial expression and PANSS individual items.

| Item | | Facial expressivity |
|---|---|---|
| 1. Facial expressivity | Pearson's r | — |
| | p-value | — |
| 2. Delusions | Pearson's r | -0.813*** |
| | p-value | < .001 |
| 3. Disorganization | Pearson's r | -0.375 |
| | p-value | 0.125 |
| 4. Hallucinations | Pearson's r | -0.498* |
| | p-value | 0.036 |
| 5. Excitement | Pearson's r | 0.206 |
| | p-value | 0.412 |
| 6. Grandiosity | Pearson's r | -0.154 |
| | p-value | 0.542 |
| 7. Paranoia | Pearson's r | -0.637** |
| | p-value | 0.005 |
| 8. Hostility | Pearson's r | -0.386 |
| | p-value | 0.114 |
| 9. Blunted Affect | Pearson's r | -0.488* |
| | p-value | 0.040 |
| 10. Emotional Withdrawal | Pearson's r | -0.523* |
| | p-value | 0.026 |
| 11. Poor Rapport | Pearson's r | -0.231 |
| | p-value | 0.356 |
| 12. Social Withdrawal | Pearson's r | -0.629** |
| | p-value | 0.005 |
| 13. Abstract Thinking | Pearson's r | 0.175 |
| | p-value | 0.486 |
| 14. Lack of Spontaneity | Pearson's r | -0.244 |
| | p-value | 0.330 |

| | | |
|---|---|---|
| 15. Stereotyped Thinking | Pearson's r | -0.408 |
| | p-value | 0.093 |
| 16. Somatic | Pearson's r | -0.202 |
| | p-value | 0.421 |
| 17. Anxiety | Pearson's r | -0.381 |
| | p-value | 0.119 |
| 18. Guilt | Pearson's r | -0.499* |
| | p-value | 0.035 |
| 19. Tension | Pearson's r | -0.397 |
| | p-value | 0.102 |
| 20. Mannerisms | Pearson's r | -0.650** |
| | p-value | 0.003 |
| 21. Depression | Pearson's r | -0.263 |
| | p-value | 0.292 |
| 22. Motor Retardation | Pearson's r | -0.445 |
| | p-value | 0.064 |
| 23. Uncooperativeness | Pearson's r | -0.337 |
| | p-value | 0.171 |
| 24. Unusual Thought Content | Pearson's r | -0.537* |
| | p-value | 0.022 |
| 25. Disorientation | Pearson's r | -0.103 |
| | p-value | 0.683 |
| 26. Poor Attention | Pearson's r | -0.226 |
| | p-value | 0.367 |
| 27. Lack of Judgement | Pearson's r | -0.177 |
| | p-value | 0.482 |
| 28. Disturbance of Volition | Pearson's r | -0.775*** |
| | p-value | < .001 |
| 29. Poor Impulse Control | Pearson's r | -0.300 |
| | p-value | 0.226 |
| 30. Preoccupation | Pearson's r | -0.298 |

| | | |
|---|---|---|
| | p-value | 0.230 |
| 31. Social Avoidance | Pearson's r | -0.792*** |
| | p-value | < .001 |

* p < .05, ** p < .01, *** p < .001

**Supplementary Table 4:** Correlation between facial and vocal markers during free behavior and PANSS individual items.

| Variable | | Blunted Affect | Emotional Withdrawal | Poor Rapport | Social Withdrawal | Abstract Thinking | Lack of Spontaneity | Stereotyped Thinking | Motor Retardation | Social Avoidance | Facial expressivity | Vocal intensity | Fundamental frequency | Fundamental frequency standard deviation | Harmonics to noise ratio | Vocal jitter |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Blunted Affect | Pearson's r | — | | | | | | | | | | | | | | |
| | p-value | — | | | | | | | | | | | | | | |
| 2. Emotional Withdrawal | Pearson's r | 0.454* | — | | | | | | | | | | | | | |
| | p-value | 0.044 | — | | | | | | | | | | | | | |
| 3. Poor Rapport | Pearson's r | 0.598** | 0.647** | — | | | | | | | | | | | | |
| | p-value | 0.005 | 0.002 | — | | | | | | | | | | | | |
| 4. Social Withdrawal | Pearson's r | 0.452* | 0.792*** | 0.508* | — | | | | | | | | | | | |
| | p-value | 0.045 | < .001 | 0.022 | — | | | | | | | | | | | |
| 5. Abstract Thinking | Pearson's r | 0.103 | 0.457* | 0.347 | 0.457* | — | | | | | | | | | | |
| | p-value | 0.665 | 0.046 | 0.134 | 0.043 | — | | | | | | | | | | |
| 6. Lack of Spontaneity | Pearson's r | 0.605** | 0.630** | 0.721*** | 0.590** | 0.468* | — | | | | | | | | | |
| | p-value | 0.005 | 0.003 | < .001 | 0.006 | 0.038 | — | | | | | | | | | |
| 7. Stereotyped Thinking | Pearson's r | 0.524* | 0.660** | 0.714*** | 0.618*** | 0.570** | 0.713*** | — | | | | | | | | |
| | p-value | 0.018 | 0.002 | < .001 | < .001 | 0.009 | < .001 | — | | | | | | | | |
| 8. Motor Retardation | Pearson's r | 0.714*** | 0.463** | 0.690*** | 0.752*** | 0.348 | 0.720*** | 0.601** | — | | | | | | | |
| | p-value | < .001 | 0.001 | < .001 | < .001 | 0.134 | < .001 | 0.006 | | | | | | | | |
| 9. Social Avoidance | Pearson's r | 0.380 | 0.590** | 0.233 | 0.804*** | 0.131 | 0.182 | 0.528* | 0.451* | — | | | | | | |
| | p-value | 0.093 | 0.007 | 0.322 | < .001 | 0.570 | 0.443 | 0.017 | 0.046 | — | | | | | | |
| 10. Facial expressivity | Pearson's r | -0.045 | 0.181 | 0.053 | -0.101 | 0.328 | 0.108 | 0.137 | -0.095 | -0.102 | — | | | | | |
| | p-value | 0.854 | 0.452 | 0.826 | 0.681 | 0.171 | 0.661 | 0.575 | 0.697 | 0.678 | — | | | | | |
| 11. Vocal intensity | Pearson's r | -0.302 | -0.530* | -0.337 | -0.512* | -0.366 | -0.418 | -0.279 | -0.540* | -0.425 | 0.364 | — | | | | |
| | p-value | 0.208 | 0.016 | 0.159 | 0.020 | 0.123 | 0.077 | 0.248 | 0.017 | 0.070 | 0.126 | — | | | | |
| 12. Fundamental frequency | Pearson's r | -0.355 | -0.632** | -0.480* | -0.514* | -0.444 | -0.508* | -0.346 | -0.564* | -0.453 | 0.194 | 0.930*** | — | | | |
| | p-value | 0.136 | 0.004 | 0.037 | 0.024 | 0.057 | 0.026 | 0.147 | 0.012 | 0.051 | 0.451 | < .001 | — | | | |
| 13. Fundamental frequency | Pearson's r | -0.192 | -0.441 | -0.371 | -0.398 | -0.030 | -0.282 | -0.170 | -0.365 | -0.404 | 0.179 | 0.561** | 0.528* | — | | |
| | p-value | 0.432 | 0.055 | 0.774 | 0.101 | 0.904 | 0.246 | 0.485 | 0.124 | 0.086 | 0.464 | 0.009 | 0.020 | — | | |
| 14. Harmonics to noise ratio | Pearson's r | -0.229 | -0.668** | -0.583** | -0.490* | -0.412 | -0.460* | -0.380 | -0.481* | -0.303 | 0.174 | 0.654** | 0.774*** | 0.478* | — | |
| | p-value | 0.347 | 0.002 | 0.009 | 0.031 | 0.070 | 0.038 | 0.104 | 0.037 | 0.207 | 0.475 | 0.002 | < .001 | 0.038 | — | |
| 15. Vocal jitter | Pearson's r | 0.057 | 0.564* | 0.457* | 0.353 | 0.332 | 0.328 | 0.209 | 0.308 | 0.209 | -0.097 | -0.541* | -0.691** | -0.279 | -0.937*** | — |
| | p-value | 0.816 | 0.012 | 0.049 | 0.136 | 0.164 | 0.170 | 0.381 | 0.205 | 0.380 | 0.692 | 0.017 | 0.001 | 0.248 | < .001 | — |
| 16. Speech to silence ratio | Pearson's r | -0.396 | -0.510* | -0.468* | -0.442 | -0.453 | -0.547* | -0.526 | -0.552* | -0.359 | 0.161 | 0.868*** | -0.923*** | 0.260 | 0.576** | -0.510* |
| | p-value | 0.093 | 0.028 | 0.043 | 0.058 | 0.051 | 0.017 | 0.174 | 0.014 | 0.133 | 0.510 | < .001 | < .001 | 0.283 | 0.010 | 0.028 |