Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–11

DOI: 10.1177/20552076231180523
journals.sagepub.com/home/dhj

Sage

Evaluating the utility of daily speech assessments for monitoring depression symptoms

Melisa Gumus[1,2], Danielle D DeSouza[1], Mengdan Xu[1], Celia Fidalgo[1], William Simpson[1,3] and Jessica Robin[1]

Abstract

Objective: Depression is a common mental health disorder and a major public health concern, significantly interfering with the lives of those affected. The complex clinical presentation of depression complicates symptom assessments. Day-to-day fluctuations of depression symptoms within an individual bring an additional barrier, since infrequent testing may not reveal symptom fluctuation. Digital measures such as speech can facilitate daily objective symptom evaluation. Here, we evaluated the effectiveness of daily speech assessment in characterizing speech fluctuations in the context of depression symptoms, which can be completed remotely, at a low cost and with relatively low administrative resources.

Methods: Community volunteers ($N = 16$) completed a daily speech assessment, using the Winterlight Speech App, and Patient Health Questionnaire-9 (PHQ-9) for 30 consecutive business days. We calculated 230 acoustic and 290 linguistic features from individual's speech and investigated their relationship to depression symptoms at the intra-individual level through repeated measures analyses.

Results: We observed that depression symptoms were linked to linguistic features, such as less frequent use of dominant and positive words. Greater depression symptomatology was also significantly correlated with acoustic features: reduced variability in speech intensity and increased jitter.

Conclusions: Our findings support the feasibility of using acoustic and linguistic features as a measure of depression symptoms and propose daily speech assessment as a tool for better characterization of symptom fluctuations.

Keywords

Speech, depression, PHQ, linguistics, digital health

Submission date: 6 September 2022; Acceptance date: 19 May 2023

Introduction

Depression is a leading cause of disability worldwide and a major global health concern.[1] Approximately 280 million people in the world suffer from this common but serious mental health condition.[1,2] A person with depression is at increased risk of experiencing irritability, sadness, loss of appetite, cognitive or social impairment and suicidal thoughts.[1,2] Clinical diagnosis of depression relies on self or caregiver reports of symptoms according to Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) or International Classification of Diseases 10th Revisions (ICD-10).[3,4] The subjective and potentially biased assessment relies on the retrospective recall of episodes, complicating symptom evaluation even further. Digital measures can benefit clinical practice as they can

[1]Winterlight Labs, Toronto, Ontario, Canada
[2]Department of Psychology, University of Toronto, Toronto, Ontario, Canada
[3]McMaster University, Hamilton, Ontario, Canada

Corresponding author:
Jessica Robin, Winterlight Labs, Toronto, Ontario, Canada.
Email: jessica@winterlightlabs.com

address these problems through objective assessment of symptoms.[5,6] Digital tools are more accessible and can be lower burden as they can be administered remotely and reduce time and resource investment.[7] Moreover, digital tools enable frequent assessments of symptoms at multiple time points, creating more detailed data at the individual level.

There have been efforts toward developing objective assessment tools for depression symptoms using innovative techniques from neuroimaging to speech.[8,9] Neuroimaging can reveal brain changes in various medical conditions, and thus can be effective in understanding heterogeneous symptom profiles of patients.[9,10] Similarly, speech production recruits many brain regions for generating linguistic representations from thoughts, and thus provides a window into cognitive and emotional processing through simple assessments.[11] Common acoustic features of speech (e.g. jitter and shimmer) represent the sound of the voice through mathematical deconstructions of the sound wave.[12,13] On the other hand, linguistic features represent sentence generation, vocabulary and syntactic structures and sentiment, reflecting the content of speech and language.[13,14] Through cutting edge analysis methods such as signal processing, computational linguistics, and machine learning, changes in speech and language patterns can provide insights into psychiatric disorders.[15–18]

Speech changes that accompany depression symptoms have been recognized for many years.[17–19] Patients with depression show reduced fundamental frequency range, perceived as pitch, which reflects the monotonous speech often observed clinically.[20–22] On the other hand, another set of acoustic features including the variability in fundamental frequency, jitter and shimmer tend to increase with the severity of depression, which are thought to be related to motor speech control and laryngeal musculature.[21,22] In addition to the alteration in the way speech is produced, individuals with depression also exhibit changes in the linguistic content. They tend to use more first-person pronouns and negative emotional words.[23–26] This may be reflective of a more self-regulatory cycle in which an individual's thoughts are focused on themselves.[27] As their thoughts about themselves, world and life become more negative, they begin using positive words much less frequently. Interestingly, this has been observed even among those who are not experiencing depression symptoms.[28]

Speech and language measurements have the additional benefit of being easily collected remotely, increasing accessibility and lowering administrative burden. With the coronavirus (COVID-19) pandemic, the world entered a digital era in which more of our day-to-day activities could be performed remotely. This new shift in our lives also precipitated a shift to decentralized clinical trials and research, utilizing remote clinical assessments and requiring validation of such digital tools.[29] Another benefit of remote assessments is the ability to collect frequent data. Speech-based assessments are not only easily adaptable to decentralized, remote methods, but they also allow for daily administration with minimal instructions or involvement of clinicians.

Sampling speech more than once is necessary for accurate monitoring of specific speech characteristics related to depression[30,31] as its symptoms exhibit irregular patterns and stochastic fluctuations.[32] In this study, our goals were to (a) evaluate the feasibility of collecting remote speech data, (b) determine the value in daily speech assessment, using the Winterlight Speech Assessment App, and (c) investigate which aspects of speech relate to depression symptoms, in order to validate the utility of daily speech assessment for detecting and monitoring symptoms of depression.

Methods

Procedure

This longitudinal, remote study included participants with or without psychiatric diagnoses. Participants completed daily speech assessments and mental health surveys remotely for 30 business days. This longitudinal data enabled us to examine how speech features and symptoms changed over different intervals of time and the reliability of frequent test administration. The study protocol was reviewed and approved by the Advarra Research Ethics Board and registered at clinicaltrials.gov (NCT04851912).

All participants completed three assessment types across 30 days: a practice assessment, followed by daily and bi-weekly assessments (Figure 1). The first session on Day 1 was a practice session. It allowed participants to get familiar with the speech tasks and data collection software, the Winterlight Speech App. Bi-weekly assessments were administered on Days 2 (first study session), 16 (midway through the study) and 30 (last study session). The remaining 26 days in between the three bi-weekly assessment days were part of the daily assessments.

Bi-weekly assessments included six speech assessments: picture description, paragraph reading/recall, semantic fluency, phonemic fluency and journaling. There were five different mental health assessments: Patient Health Questionnaire-9 (PHQ-9),[33] Generalized Anxiety Disorder (GAD-7),[34] Sheehan Disability Scale (SDS),[35] a rumination scale, and a mood questionnaire. These assessments were selected and administered to reflect ecological momentary assessments (EMAs), which allow for repeated assessment of an individual in their naturalistic environment.[36,37] On the other hand, daily assessments only included an open-ended journaling speech task as well as modified versions of the PHQ-9 and mood assessments. We only analyzed journaling tasks and PHQ-9 scores from the bi-weekly assessments to be consistent with the daily data.

Participants

Participants were recruited from the community through online advertisement. Written consent was obtained from

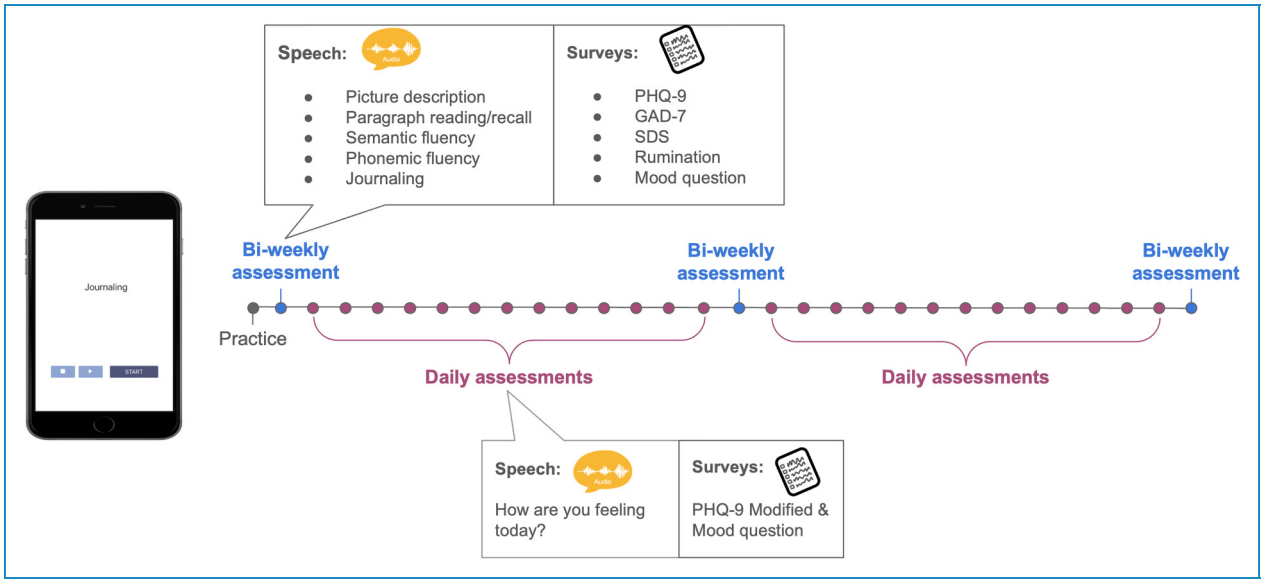

Figure 1. Study design integrating Winterlight Speech App. Participants completed 30 sessions in total: a practice, 3 sessions (start, midway and last day) as part of the bi-weekly assessments, 26 sessions in the daily assessments.

each participant prior to their participation in the study. To be eligible for the study, participants needed to be between the ages of 18 and 95, fluent English speakers (i.e. either English as their first language or they can speak with conversational proficiency) and own an iOS device (iPhone or iPad). The exclusion criteria included the following: residing outside of Canada or the United States, having experienced a chronic alcohol or drug dependence within the last 5 years, and being identified with a clinically significant vision or hearing impairment. After obtaining the written consent, participants were invited for a conference call with a member of the study team, M.G. During these sessions, M.G. collected demographic information and assisted participants with Winterlight Speech App installation on their devices. If participants preferred to skip the tutorial session and were comfortable with app installation, then their demographic information was collected via email. They all received the tutorial slides via email. Participants were compensated for $5 per session, earning up to $155 in total for participating in all 30 sessions and the tutorial.

Winterlight Speech Assessment (iOS App)

All speech assessments were conducted using the Winterlight Lab Speech App. Participants logged into their accounts to complete the speech assessments using their individual username and password. Upon login, the speech assessments associated with that day were available to them. All 30 days included a journaling task, with the prompt, "How are you feeling today?." Participants were instructed to provide as much detail as they preferred, and there was no time limit for the task. Once participants pressed start, the device's microphone recorded their speech until they indicated they were finished with the task. Audio recordings were uploaded to secure servers for analysis upon connection to wifi or cellular signal.

Speech features

The speech samples recorded by the participants were transcribed by trained transcriptionists who ensured the audio contained participant speech and flagged samples with audio issues (e.g. no audible speech, very poor-quality audio), for removal. Acoustic and linguistic features were then computed using the Winterlight Labs pipeline (www.winterlightlabs.com), which uses Python-based acoustic and natural language processing libraries and custom code to compute 520 speech features based on each speech recording and its accompanying transcript. These variables reflect the acoustic (e.g. properties of the sound wave, speech rate, number of pauses), lexical (e.g. rates and types of words used, and their characteristics such as frequency and imageability, which reflect how commonly words are used and how easy they are to picture, respectively), semantic (relating to the meaning of the words, e.g. semantic relatedness of subsequent utterances, semantic relatedness of utterances to the items in the picture) and syntactic (relating to the grammar of the sentences, e.g. syntactic complexity, use of different syntactic constructions) properties of the sample. Open source packages include SpaCy for parts-of-speech tagging and morphological features,[38] the Stanford NLP parser for syntactic features,[39] Praat and Parselmouth for acoustic features,[40,41] and GloVe and FastText models for semantic features.[42,43]

Table 1. Speech feature overview, definitions, numbers, and examples.

Speech feature category	Definition	Number of features	Examples
Acoustic	Variables describing the acoustic properties of the sound wave	209	Fundamental frequency; Mel-frequency Cepstral Coefficients (MFCCs); zero crossing rate
Timing	Variables relating to the rate of speech and total speech output	21	Speech rate (words/minute); articulation rate (syllables/second); number of pauses; pause duration; total duration of speech
Parts of speech	Variables enumerating the rate of usage of different parts of speech	72	Use of nouns, pronouns, verbs, adjectives, etc.
Lexical	Variables describing the characteristics of words used	32	Frequency, familiarity, imageability of words; measures of vocabulary diversity such as type-token ratio
Syntactic	Variables enumerating the rate of usage of different syntactic structures and measures of syntactic complexity	144	Number of clauses per sentence; use of noun phrases, verb phrases, subordinate phrases, etc.
Discourse	Variables using cosine distance and graph theoretical measures to calculate the organization and repetition of utterances	18	Average cosine distance between utterances; graph density, number of nodes and diameter
Coherence	Variables using word vector models to calculate the semantic similarity between utterances	15	Average, minimum and maximum cosine distances between subsequent utterances in word vector space
Sentiment	Variables describing the sentiment of the words used	9	Valence, arousal and dominance scores for all words and word types

The pipeline also uses custom code to compute additional features based on the transcript and audio file, using lexical norms from previous publications[44–47] or previously published models and features.[48] An overview of the features categories, definitions, feature numbers, and examples is provided in Table 1.

Clinical assessments

We only analyzed the PHQ-9 in this paper for consistency between daily and bi-weekly data. The PHQ-9 was administered through a link to an online form. The questionnaires used in this study are presented in Supplementary Material. All other mental health assessments were combined with the PHQ-9 into one form for the convenience of the participants. The PHQ-9 is a well-established self-rated measure for depression symptoms.[33] It contains 9 questions which correspond to the core DSM criteria for depression, with each question ranging from 0 to 3 points for a total of 27 points. It has been validated against clinician rated measures and cut-off scores are well established. In bi-weekly assessments, we used the PHQ-9 with a traditional scoring system. For daily assessments, we used a modified version of the PHQ-9. The questions remained the same but were modified to refer to the current day (rather than past 2 weeks) and answers ranged from 1 to 7 points for a total of 63 points, with higher scores indicating more depression systems. The reason for using modified PHQ-9 was to adapt it to reflect daily assessments and increase the answer range due to concerns that the original scale may not having enough dynamic range to detect small daily variations. For interpretability of our results, we downscaled the total scores obtained through modified PHQ-9, so they reflect the original scale with a total of 27 points.

Statistical analysis

All analyses were completed on R statistical software, version 4.1.2.[49] Repeated measures correlation (rmcorr)[50] was used to evaluate within-individual associations between the speech features and depression/mood symptoms across multiple time points. We eliminated acoustic and linguistic features that had empty values for at least 20% of participants. The empty values were due to most of these features being specific to other tasks such as

picture description and not relevant to the task of interest (i.e. journaling). This data cleaning process yielded 262 acoustic and linguistic features in our analyses. We fit separate rmcorr models to each of these features to investigate their unique relationship with modified PHQ-9 scores at the individual level. This linear association was represented with a correlation coefficient (r_{rm}) and allowed us to investigate common intra-individual associations without violating independence assumptions or requiring simple averaging across sessions. Repeated measure correlation plots demonstrate the linear fit for each participant, providing a visual representation of a particular speech-depression relationship across participants. Statistical significance was set to an alpha level equal to 0.0002 taking multiple comparisons into consideration through the Bonferroni correction method (0.05/262 features).

Results

Demographics

The study included 16 participants (11 females, 5 males) with an age range of 21–54 ($M = 30.75 \pm 9.28$). Four participants reported to be on emotional or behavioral medications and five reported using medications for their physical health (Table 2). The daily assessments had a completion rate of 93% for both speech assessments and surveys with only 1 participant discontinuing participation in the second half of the study due to personal reasons. At baseline (on Day 2 as Day 1 was practice), participants reported a mean depression score of 6.47 ± 6.24 on PHQ-9 assessment.

Table 2. Participant demographics.

Demographic variable	*N*/Mean ± SD
Sample size	16
Age	30.75 ± 9.28 (21–54)
Sex	11 female (69%)/5 male (31%)
Education	16.00 ± 1.97
Native language	10 English 2 Chinese 2 Tamil 1 Hindi 1 Gujarati
Ethnicity	6 Brown or South Asian 5 White or Caucasian 3 Asian or Pacific Islander 2 Black or African American
Medications for emotion/behavior	4 (Vortioxetine, Venlafaxine, Citalopram, Escitalopram)
Medications for physical health	5
Daily speech assessment completion rate	93%
Daily survey completion rate	93%

Daily speech assessments relate to depression

We examined correlations between speech characteristics and depression symptoms collected daily for 26 sessions. At the intra-individual level, we found a significant relationship between linguistic features such as sentiment scores (i.e. a measure of the average valence or dominance score of each word used) and depression scores reported on the PHQ-9. According to repeated measures correlations, sentiment dominance, r_{rm} (370) = −0.31, 95% CI [−0.40, −0.22], $p < 0.001$, and sentiment valence, r_{rm} (370) = −0.34, 95% CI [−0.43, −0.25], $p < 0.001$, were related to depression scores (Figure 2(a) and 2(b)), with higher scores relating to more negative and less dominant words used. The sentiment features were calculated by taking the average valence scores, representing the positivity or negativity of each word, and dominance scores, representing whether a word denotes being in control or feeling controlled, for all words in the transcript that had normative values available.[47] These associations remained significant following a Bonferroni correction for multiple comparisons at the alpha level of 0.0002.

We also observed significant correlations between acoustic features and depression scores in the repeated measures analyses. At the intra-individual level, higher depression scores were associated with lower vocal intensity range (i.e. speech volume), r_{rm} (370) = −0.22, 95% CI [−0.32, −0.12], $p < 0.001$, but greater jitter, r_{rm} (370) = 0.24, 95% CI [0.14, 0.34], $p < 0.001$ (Figure 2(c) and (d)). Intensity range is the difference between the maximum and minimum of the intensity curve of the recording, representing the range of perceived loudness. Jitter is a calculation of the average absolute difference between consecutive periods in an acoustic signal, used as a measure of vocal quality. Significance remained after a Bonferroni correction for multiple comparisons at the alpha level of 0.0002

Unclear link between speech and depression in bi-weekly data

Depression scores exhibited daily variability within individuals over 20+ sessions (Figure 3(a)). We investigated whether the speech-depression relationships observed in daily data, discussed above, could be detected from our bi-weekly assessments alone, whose sessions are spaced 2 weeks apart. We chose this cadence because the commonly used PHQ-9 asks participants to reflect on their feelings

from the past 2 weeks. Thus, a 2-week data collection is typical. We note that by virtue of being bi-weekly, this analysis contained fewer observations and therefore had reduced power compared to the daily assessment analysis.

When we restricted our analyses to bi-weekly time points (i.e. Days 2, 16, and 30), we found that acoustic and linguistic features were not significantly associated with depression scores as reported using the standard PHQ-9. This indicates the importance of large sample size, especially in the studies with relatively low number of study sessions. For example, although nonsignificant, the link between sentiment dominance and depression scores had a negative trend on Day 2, $r^2 = 0.08$, $F(1, 13) = 2.17$, $\beta = -0.02$, $p = 0.16$, 95% CI [−0.04, 0.008], but had a positive trend on Day 30, $r^2 = 0.03$, $F(1, 12) = 1.38$, $\beta = 0.01$, $p = 0.26$, 95% CI [−0.01, 0.03] (Figure 3(b) and (c)), showing that patterns were inconsistent with bi-weekly sampling and small sample size. Similarly, we investigated the relationship between sentiment dominance and depression in bi-weekly data, using repeated measures analysis. Three visits were not sufficient to capture the significant negative association between the sentiment dominance and depression scores reported within the daily assessments, r_{rm} (28) = 0.01, 95% CI [−0.36, 0.39], $p = 0.94$ (Figure 3(d)).

Bi-weekly data also did not show any significant correlations between acoustic features and depression scores. For example, jitter in speech and depression had a negative, but not significant relationship on both Day 2, $r^2 = -0.04$, $F(1, 13) = 0.47$, $\beta = -0.0002$, $p = 0.50$, 95% CI [−0.001, 0.0005], and Day 30, $r^2 = -0.07$, $F(1, 12) = 0.12$, $\beta = -0.0001$, $p = 0.74$, 95% CI [−0.001, 0.0007] (Figure 3(e) and (f)). Repeated measure analysis over three time points was also not sufficient to capture the significant positive relationship reported between jitter and depression in the daily data, r_{rm} (28) = −0.05, 95% CI [−0.41, 0.33], $p = 0.80$ (Figure 3(g)).

Discussion

This study provided evidence for the feasibility of daily speech assessments in monitoring depression scores and demonstrated the success of remote, app-based speech data collection. The Winterlight Speech Assessment app was easily accessible and required relatively less time and resources than in-person assessments. It also enabled collection of daily speech recordings, providing extensive data for each individual over a short study duration. Our daily assessments captured daily variations in individual depression scores. Fluctuations in depression symptoms have been previously reported in the literature.[32] Our results indicated that daily speech assessments could help characterize individual variability in depression scores as they capture increased variability in the depressive states of an individual. For example, those with high depression scores used more negative, less authoritative words (i.e. low sentiment and dominance scores), and their speech intensity range was reduced with increased vocal jitter. However, these speech-depression relationships were not detected in the bi-weekly assessments (i.e. over 1–3 sessions). Although the daily and bi-weekly assessments differed in the amount of data, we hypothesize that larger sample sizes would be required to detect the speech-depression associations in bi-weekly data. On the other hand, daily data even in small studies can more accurately characterize the relationship between depression, and linguistic and acoustic speech features at the individual level, capturing daily fluctuations and improving the sensitivity of assessments.

Individuals with higher depression scores tend to use less dominant and more negatively valenced words. Linguistic speech features, specifically sentiment dominance and valence, presented a significant association with depression scores measured with the PHQ-9. As the participants reported feeling more depressed, their sentiment dominance and valence scores decreased, meaning the words they chose to describe their daily feelings were not as positive and denoted feeling less in control. These findings are in line with the literature where depressed individuals were shown to use less positive or emotional words.[23,24] More importantly, our findings revealed this relationship between the use of less dominant and more negative words and depression symptoms, in a non-clinical, normative population. A recent study also showed that subtle changes in speech can related to depression symptoms in non-clinical population.[51] These relationships may be stronger in clinical samples with higher depression severity. The linguistic speech features provide us a potential objective measure to monitor the symptoms of not only depressed individuals but also those at risk for or more vulnerable to experiencing depression symptoms. These findings also overlap with Beck's cognitive theory of depression: the negative thoughts are central to depression, which, in fact, precedes any other physiological or mood symptoms.[52,53]

In addition to the content of the speech, the way speech is produced (i.e. acoustic speech features) was also altered with depression symptoms. Our results revealed that higher depression scores were correlated with reduced intensity range and increased jitter (i.e. variability in speech frequency) in speech, which are consistent with the literature.[17,18] Acoustic features are directly affected by laryngeal muscles, and the changes in acoustic features could imply changes to vocal prosody and muscle tension.[54] Depressed patients have previously been reported to exhibit reduced speaking intensity.[55] This could be one of the factors contributing to monotonous speech prosody observed in patients with depression.[31] Similarly, previous literature reports greater jitter in the speech of patients with depression which could be the signature of muscle tension, resulting in more rough or hoarse speech.[56] Observing similar findings in a non-clinical population suggests that speech has the potential as a

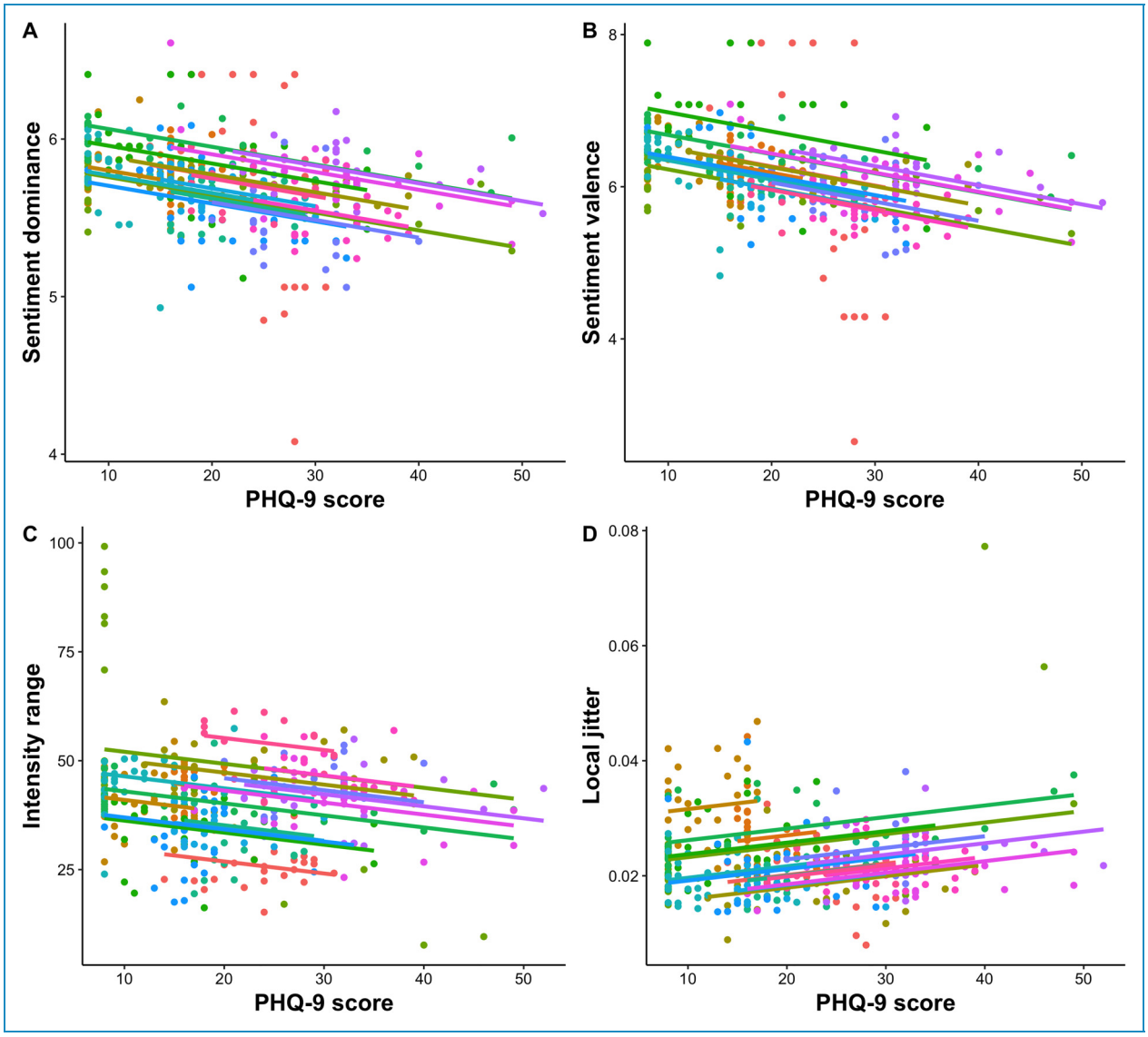

Figure 2. Daily assessments, through repeated measure analyses (rmcorr), capture the intra-individual associations between depression scores on modified PHQ-9, and acoustic and linguistic speech features. Each dot represents the feature value and corresponding PHQ-9 score from each assessment. Each participant is represented with a different color with corresponding lines showing the rmcorr fit for each participant. Linguistic features including (A) sentiment dominance and (B) valence were negatively correlated with depression scores. Among the acoustic features, (C) intensity range was negatively associated with depression while (D) jitter showed a positive relationship.

digital measure for detecting and monitoring even the mild symptoms of depression.

Longitudinal studies enable detection of speech changes over time, accounting for symptom fluctuations in psychiatric disorders. For example, changes in speech rate over time were shown to be correlated with depression scores in patients with major depressive disorder and bipolar disorder.[57] Similarly, speech articulatory coordination was reported as a way of detecting depression symptoms in patients who were receiving treatment.[58] Speech was also investigated in response to antidepressant treatment and as a biomarker for depression severity. Although patients with depression initially showed longer speech pauses, their speech was shortened following treatment.[31] These studies are crucial in assessing speech and symptom fluctuations. However, following up with participants over time, at multiple time points, requires extensive time, effort and resources. In this validation study, we propose that daily speech assessments as an efficient and effective tool for data collection and tracking individual differences over time.

One of the limitations of the study was that the data was collected during the COVID-19 pandemic. The stress related to work and life adjustments, reduced social interactions, and the resulting anxiety and worry might have impacted self-reported depression symptoms as well as

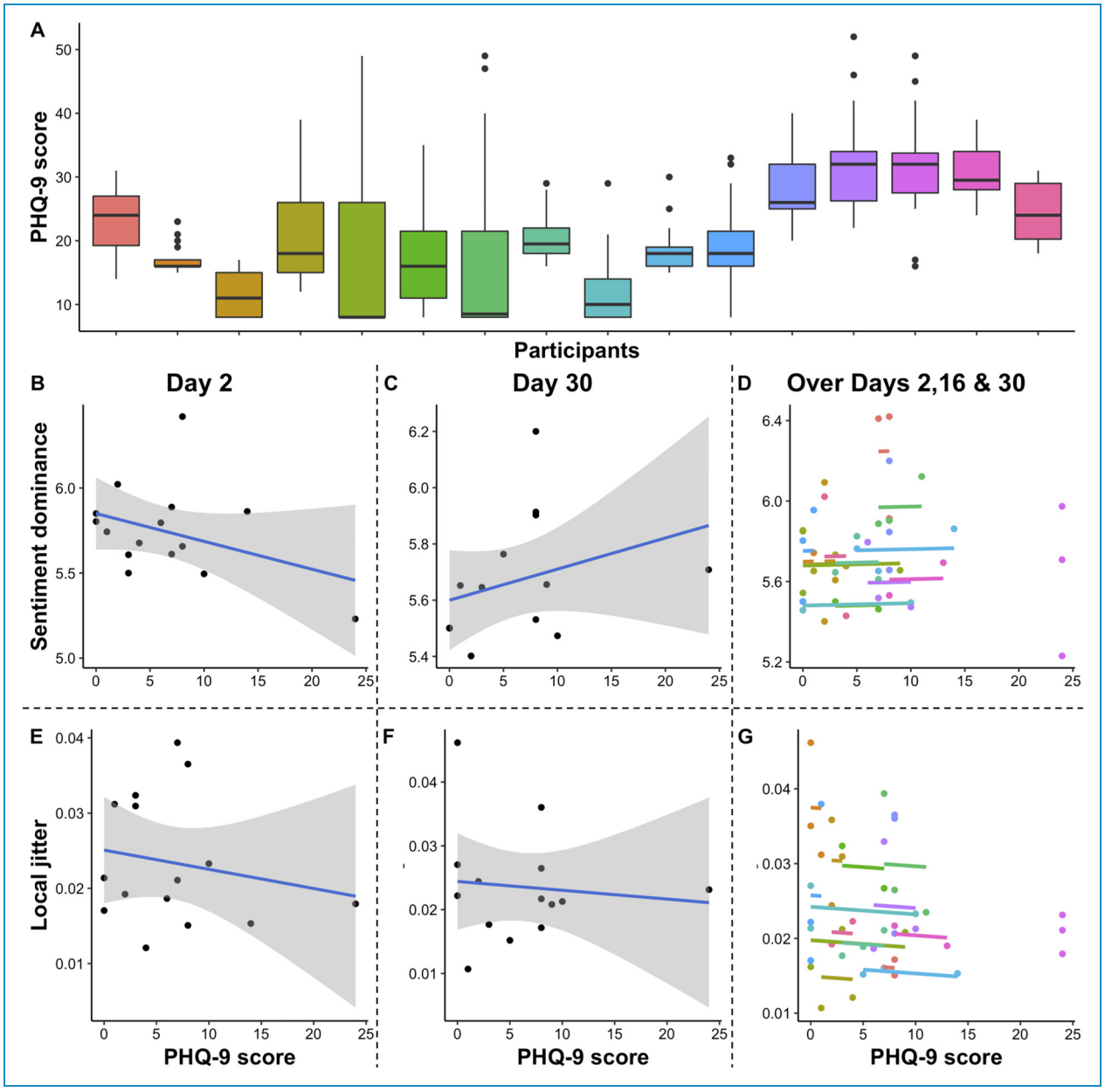

Figure 3. Daily fluctuations of individual depression scores are not captured with bi-weekly assessments. (A) Participants' PHQ-9 scores in daily assessments (20+ days) are plotted separately, depicting daily symptom fluctuations within individuals. Each color represents a different participant in the study, with boxplots depicting median PHQ-9 scores and their variation. Regarding linguistic features, the linear relationship between sentiment dominance and depression is negative on Day 2 (B) but positive on Day 30 (C), both of which are not significant. Assessments over three sessions (D) are not capturing the negative association between sentiment scores and depression observed in daily assessments in repeated measures analyses through rmcorr. For the acoustic features, the linear relationship between jitter and depression is negative on both Days 2 (E) and 30 (F), both of which are also not significant. Repeated measure analysis over three sessions (G) through rmcorr is also not capturing the positive relationship between jitter and depression that we have seen in the daily analyses. (B), (C), (E) and (F) show simple regressions: each dot represents an independent observation on specific time points. The purple line is the simple regression fit. In (D) and (G), rmcorr: each participant is represented with different colors with corresponding lines showing the rmcorr fit for each participant.

the content of participants' speech. Regardless of the source of the depression-like symptoms, we were able to observe a relationship between speech features and depression scores at the individual level. However, future studies should investigate the differences between clinical and non-clinical populations using daily speech assessments and evaluate the robustness of speech-depression relationship over time. In addition, the scales of questions in daily and

bi-weekly assessments were not ideally matched. We implemented the original PHQ-9 in the bi-weekly assessments such that each question ranged between 0 and 3, referring to the frequency of a symptom occurring within the last 2 weeks. In daily assessments, the PHQ-9 was modified so that each question remained the same, but the question referred to the current day and ranged between 1 and 7. The larger scale was implemented to better capture the daily fluctuations, which was later rescaled for comparison with the bi-weekly assessments. Future research should implement assessments with similar scales to control for any potential influences of scale differences on individuals' choice of scores.

Although our sample size was relatively small, we assessed daily speech and depression scores of each participant, sampling them at multiple time points. This study was conducted as an exploratory study to test the effectiveness of daily speech assessments. Repeated measures analyses with 20+ time points allowed us to focus on intra-individual variability, and thus, enabled us to use a smaller sample. The power of repeated measures analyses through rmcorr, increases exponentially by either the sample size or the number of observations. Although repeated measure analyses on daily assessments had a relatively high power, it is necessary to investigate the findings from bi-weekly assessments in more than three time points. These findings should be replicated in larger samples as well as in more sophisticated analyses such as deep learning. Our focus on this paper was to validate that daily speech assessments provide a feasible, objective digital tool, and to identify candidate speech features that track fluctuations in symptoms. Future research should build on this work, and others, to implement deep learning or classification models to potentially develop speech biomarkers based on daily assessments and investigate how speech can potentially inform symptom changes in remission and relapse.

Conclusion

The present study tested the potential of using daily speech assessment as a digital tool for monitoring depression symptoms. Here, we demonstrated that daily speech assessments can capture signals relating to the daily depression symptom fluctuations, providing a replication of the speech-depression associations reported in the literature. Our findings highlight that this speech-depression relationship is not only specific to high severity clinical populations and suggests that speech has the potential to be used as a digital measure for detecting those at risk of or more vulnerable to depression. Speech, as a promising digital measure, not only provides an opportunity to objectively assess depression symptoms, but also allows for remote data collection. While many other techniques are costly and require in-person administration, speech assessments are accessible and allow for frequent administration, which can complement more in-depth clinical assessments and help make tools for monitoring symptoms more accessible. Daily speech assessment holds the potential to be used as a digital measure for monitoring symptoms in depression and many other psychiatric diseases.

Acknowledgements: The authors would like to thank the employees of Winterlight Labs for their help and support throughout the study and the study volunteers who participated in the research.

Contributorship: MG, DDD, CF, WS, and JR designed the study. MG conducted the study, collected/analyzed the data, interpreted the results, and wrote the first draft. MX assisted with analyses. DDD helped with interpretation. JR supervised the project. All authors reviewed and edited the manuscript.

Declaration of conflicting interests: The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: MG, MX, CF, WS and JR are employees of Winterlight Labs.

Ethics approval: The study was approved by the Advarra Research Ethics Board.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is funded by Winterlight Labs.

Guarantor: Not applicable.

ORCID iD: Melisa Gumus https://orcid.org/0000-0003-2478-2572

Supplemental material: Supplemental material for this article is available online.

References

1. Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx) [Internet], http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b (2019).
2. WHO. Mental disorders [Internet], https://www.who.int/news-room/fact-sheets/detail/depression (2021).
3. American Psychiatric Association . *Diagnostic and statistical manual of mental disorders: DSM-5™*. 5th ed. Arlington, VA: American Psychiatric Publishing, Inc., 2013.
4. World Health Organization. *The ICD-10 classification of mental and behavioural disorders : clinical descriptions and diagnostic guidelines [Internet]*. Genève, Switzerland, 1992 , https://apps.who.int/iris/handle/10665/37958.
5. Abbas A, Sauder C, Yadav V, et al. Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study. *Front Digit Health [Internet]* 2021; 3: 610006.

6. Coravos A, Khozin S and Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digit Med* 2019; 2: 1–5.
7. Robin J, Harrison JE, Kaufman LD, et al. Evaluation of speech-based digital biomarkers: review and recommendations. *Digit Biomark* 2020; 4: 99–108.
8. Gururajan A, Clarke G, Dinan TG, et al. Molecular biomarkers of depression. *Neurosci Biobehav Rev* 2016; 64: 101–133.
9. Patel MJ, Khalaf A and Aizenstein HJ. Studying depression using imaging and machine learning methods. *NeuroImage: Clin* 2016; 10: 115–123.
10. Gumus M, Mack ML, Green R, et al. Brain connectivity changes in postconcussion syndrome as the neural substrate of a heterogeneous syndrome. *Brain Connect* 2022; 12(8): 711–724.
11. Dronkers N and Ogar J. Brain areas involved in speech production. *Brain* 2004; 127: 1461–1462.
12. Massaro DW. 3 – Acoustic features in speech perception. In: Massaro DW (ed.) *Understanding language [internet]*. Madison, Wisconsin: Academic Press, 1975 [cited 2022 Jul 23], pp.77–124, https://www.sciencedirect.com/science/article/pii/B9780124783508500088.
13. Voleti R, Liss JM and Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J Sel Top Signal Process* 2020; 14: 282–298.
14. Fraser KC, Meltzer JA and Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimer's Dis* 2016; 49: 407–422.
15. Cohen AS and Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. *Curr Opin Psychiatry* 2014; 27: 203–209.
16. Cohen AS, McGovern JE, Dinzeo TJ, et al. Speech deficits in serious mental illness: a cognitive resource issue? *Schizophr Res* 2014; 160: 173–179.
17. Low DM, Bentley KH and Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngosc Investig Otolaryngol* 2020; 5: 96–116.
18. DeSouza DD, Robin J, Gumus M, et al. Natural language processing as an emerging tool to detect late-life depression. *Front Psychiatry [Internet]* 2021; 12: 719125. https://www.frontiersin.org/article/10.3389/fpsyt.2021.719125. [cited 2022 Apr 30].
19. Moses PJ. *The voice of neurosis*. Oxford, UK: Grune & Stratton, 1954, vi, 131 p. (The voice of neurosis).
20. Cummins N, Scherer S, Krajewski J, et al. A review of depression and suicide risk assessment using speech analysis. *Speech Commun* 2015; 71: 10–49.
21. Kiss G and Vicsi K. Mono- and multi-lingual depression prediction based on speech processing. *Int J Speech Technol* 2017; 20: 919–935.
22. Horwitz R, Quatieri TF, Helfer BS, et al. On the relative importance of vocal source, system, and prosody in human depression. In: 2013 IEEE international conference on body sensor networks. IEEE, 2013, pp.1–6.
23. Rude S, Gortner EM and Pennebaker J. Language use of depressed and depression-vulnerable college students. *Cogn Emot* 2004; 18: 1121–1133.
24. Tølbøll KB. Linguistic features in depression: a meta-analysis. *J Lang Works-Sprogvidenskabeligt Studentertidsskrift* 2019; 4: 39–59.
25. Bernard JD, Baddeley JL, Rodriguez BF, et al. Depression, language, and affect: an examination of the influence of baseline depression and affect induction on language. *J Lang Soc Psychol* 2016; 35: 317–326.
26. Tackman AM, Sbarra DA, Carey AL, et al. Depression, negative emotionality, and self-referential language: a multi-lab, multi-measure, and multi-language-task research synthesis. *J Pers Soc Psychol* 2019; 116: 817–834.
27. Pyszczynski T and Greenberg J. Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression. *Psychol Bull* 1987; 102: 122–138.
28. Molendijk ML, Bamelis L, van Emmerik AAP, et al. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behav Res Ther* 2010; 48: 44–51.
29. Ferrar J, Griffith GJ, Skirrow C, et al. Developing digital tools for remote clinical research: how to evaluate the validity and practicality of active assessments in field settings. *J Med Internet Res* 2021; 23: e26004.
30. Yang Y, Fairbairn C and Cohn JF. Detecting depression severity from vocal prosody. *IEEE Trans Affect Comput* 2013; 4: 142–150.
31. Mundt JC, Snyder PJ, Cannizzaro MS, et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguist* 2007; 20: 50–64.
32. Wirz-Justice A. Diurnal variation of depressive symptoms. *Dialogues Clin Neurosci* 2008; 10: 337–343.
33. Kroenke K, Spitzer RL and Williams JBW. The PHQ-9. *J Gen Intern Med* 2001; 16: 606–613.
34. Spitzer RL, Kroenke K, Williams JBW, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med* 2006; 166: 1092–1097.
35. Sheehan KH and Sheehan DV. Assessing treatment effects in clinical trials with the Discan metric of the Sheehan disability scale. *Int Clin Psychopharmacol* 2008; 23: 70–83.
36. McKay D, Przeworski A and Neill O'S. Chapter 14 – Emerging technologies for clinical practice. In: Luiselli JK and Fischer AJ (eds) *Computer-assisted and web-based innovations in psychology, special education, and health [internet]*. San Diego: Academic Press, 2016 [cited 2022 Aug 27], pp.365–378. https://www.sciencedirect.com/science/article/pii/B9780128020753000140
37. Rabasco A and Andover M. 3.05 – ecological momentary assessment. In: Asmundson GJG (ed) *Comprehensive clinical psychology (second edition) [internet]*. Oxford, UK: Elsevier, 2022 [cited 2022 Aug 27], pp.83–90. https://www.sciencedirect.com/science/article/pii/B9780128186978001898
38. Honnibal M and Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing [Internet], https://sentometrics-research.com/publication/72/ (2017).
39. Chen D and Manning C. A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) [internet]. Doha, Qatar: Association for Computational Linguistics, 2014, pp.740–750, https://aclanthology.org/D14-1082
40. Boersma P and Weenink D. Praat: doing phonetics by computer [Internet]. 5.1.44. http://www.praat.org/ (2010).
41. Jadoul Y, Thompson B and de Boer B. Introducing parselmouth: a Python interface to praat. *J Phon* 2018; 71: 1–15.

42. Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information [Internet]. arXiv; 2017 [cited 2023 Mar 3], http://arxiv.org/abs/1607.04606
43. Pennington J, Socher R and Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) [internet]. Doha, Qatar: Association for Computational Linguistics, 2014 [cited 2023 Mar 3], pp.1532–1543, https://aclanthology.org/D14-1162
44. Brysbaert M and New B. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav Res Methods* 2009; 41: 977–990.
45. Kuperman V, Stadthagen-Gonzalez H and Brysbaert M. Age-of-acquisition ratings for 30,000 English words. *Behav Res* 2012; 44: 978–990.
46. Stadthagen-Gonzalez H and Davis CJ. The Bristol norms for age of acquisition, imageability, and familiarity. *Behav Res Methods* 2006; 38: 598–605.
47. Warriner AB, Kuperman V and Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav Res* 2013; 45: 1191–1207.
48. Mota NB, Vasconcelos NAP, Lemos N, et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE* 2012; 7: e34928.
49. R Core Team. *R: a language and environment for statistical computing [Internet]*. Vienna, Austria: R Foundation for Statistical Computing, 2021, https://www.R-project.org/
50. Bakdash JZ and Marusich LR. Repeated measures correlation. *Front Psychol [Internet]* 2017; 8: 456. https://www.frontiersin.org/article/10.3389/fpsyg.2017.00456. [cited 2022 Jul 3].
51. König A, Tröger J, Mallick E, et al. Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. *BMC Psychiatry* 2022; 22: 830.
52. Beck AT. *Depression: clinical, experimental, and theoretical aspects*. New York: Hoeber Medical Division, Harper & Row, 1967.
53. Beck AT, Rush AJ, Shaw BF, et al. *Cognitive therapy of depression*. New York: Guilford Press, 1979.
54. Cannizzaro M, Harel B, Reilly N, et al. Voice acoustical measurement of the severity of major depression. *Brain Cogn* 2004; 56: 30–35.
55. Hollien H. Vocal indicators of psychological stress. *Forensic Psychol Psychiatr* 1980; 347: 47–71.
56. Farrús M, Hernando J and Ejarque P. Jitter and shimmer measurements for speaker recognition. In: 8th Annual conference of the international speech communication association, 2007 August 27–31, Antwerp (Belgium): ISCA, 2007, pp.778–781. International Speech Communication Association (ISCA).
57. Yamamoto M, Takamiya A, Sawada K, et al. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLOS ONE* 2020; 15: e0238726.
58. Williamson JR, Young D, Nierenberg AA, et al. Tracking depression severity from audio and video based on speech articulatory coordination. *Comput Speech Lang* 2019; 55: 40–56.