voicesense

CASE STUDY

DEPRESSION SEVERITY

Clinical trial (Hamburg, Germany)

DEPRESSION SEVERITY

OBJECTIVE

The objective of the clinical study was to examine the relationship between depression severity states among depressed patients and between the acoustic vocal patterns of the patients, using the Voicesense vocal mobile device for mental health monitoring.

METHODS

163 depressed patients at different severity levels of depression participated in the trial.

The subjects participated in assessment sessions, in which they went through a psychiatric assessment using the widely known PHQ-9 tool for depression assessment, and a vocal recording sampling for the acoustic vocal analysis, using the Voicesense vocal mobile device for mental health monitoring. The PHQ-9 Depression tool scores differentiated between five depression states – Minimal, Mild, Moderate, Moderate-severe and Severe.

The Voicesense vocal mobile device for mental health monitoring collected the subjects' audio data (after secured login) while answering general questions presented by the application to the subject. At least two minutes of the subject's voice were collected in each session. The recorded audio was sent by the application to central processing in Voicesense secured cloud servers.

Voicesense central vocal analysis software applies an acoustic analysis, focusing on prosodic features of the speech. The analysis is language independent and content free (no understanding of what has being said). It calculates over 200 raw voice parameters per recording, consisting of a wide range of acoustic feature segmentation, including lengths, ranges, slopes, frequencies, values and shapes of pitch extracted parameters, amplitude extracted parameters and silences extracted parameters within the speech recording. Thousands of datapoints are calculated and averaged to reflect the individual's personal speech patterns in the given recording. The raw parameters are then calibrated and normalized to overcome possible biasing effects within the specific recording as a result of amplitude differences, pitch differences, speech type differences (conversation or monologue), gender differences and age differences.

The calibrated and normalized parameters were then analyzed using machine learning models in order to select and weight the vocal parameters that best correlate with Depression. The process used the common K-fold cross-validation method that randomly splits the dataset into training and test sub-samples and repeats the process for multiple iterations in order to reach a stable and reliable predictive model equation. The software output was a unified vocal-based score of depression.

Each subject participated in at least one assessment session (psychiatric and vocal), while part of the subjects participated in two or three assessment sessions with at least one-month interval between sessions. Overall, there were 292 valid assessment sessions that included both psychiatric and vocal evaluations.

The statistical fit between the vocal depression scores and the psychiatric depression scores (PHQ-9) were evaluated using Pearson correlation, Anova variance analysis and positive and negative predictive values (confusion matrix), separately for the training sub-sample, for the test sub-sample and for the entire sample.

PARTICIPANTS

Subjects were recruited in an outpatient neuropsychiatric treatment center in Hamburg, Germany. A total number of 163 subjects were included in the study. The sample consisted of 47 males and 116 and females. The subjects varied across ages ranging from 15 to 82 years old. Subjects varied across education levels, marital status and psychological treatment status as well.

Data collection took place between November 2018 and September 2019. A total number of 292 session samples (psychiatric and vocal) were collected.

PATIENT'S SELECTION

Inclusion Criteria included neurological or psychiatric diagnosis made by a specialist, no terminal or life-threatening illnesses, and fluent use of the German language.

Exclusion criteria included psychosis, dementia, speech or language disorders in neurological diseases, addiction history, a suicide attempt recently or in the last 12 months, or insufficient language skills.

RESULTS

Based on the PHQ-9 psychiatric depression scores of the subjects, the 292 session assessments varied across all five depressive severity states—minimal, mild, moderate, moderate-sever and severe. The corresponding 292 vocal depression assessment scores were found to be strongly and significantly correlated to the PHQ-9 assessments.
The overall Pearson correlations between the Vocal depression score and the PHQ-9 depression score were highly significant, as follows:

Overall sample (N=292):

r= 0.41; **p< 0.0001**

Training sample (N=192):

r= 0.46; **p< 0.0001**

Test sample (N=100):

r= 0.30; **p< 0.002**

The overall differentiation between the vocal depression scores of the different PHQ-9 depression categories was found as highly significant (single factor Anova test; F=14.5672; df=291; **p<0.0001)**.

Depression (PHQ-9) score	**Minimal (1)**	**Mild (2)**	**Moderate (3)**	**Moderate-Severe (4)**	**Severe (5)**
Minimal (1)	1	0.0032 **	9.89E-05 ***	8.41E-09 ***	1.6E-06 ***
Mild (2)		1	0.1155	7.26E-05 ***	0.0004 ***
Mild (3)			1	0.0145 *	0.0149 *
Moderate -Severe (4)				1	0.5238
Severe (5)					1

* p< 0.05 ** p< 0.01 *** P< 0.0001

Table 1. T-test probability matrix of Vocal depression scores by PHQ-9 severity category

The significant results of the study were consistent for both genders, for different age groups, different education levels, different marital statuses and for different psychological treatment statuses.

Following the significant relationships that were found in the study between the vocal depression score and the depression states, and given the practical intended use of the vocal analysis, the study went on and tried to evaluate the practical accuracy, or predictive power, that can be expected by using the vocal analysis for tracking and screening for depression.
The predictive power of the vocal model, or its statistical fit to the PHQ-9 depression reference scores, was evaluated using binary confusion matrix analysis.

The PHQ-9 scores were grouped into two categories: 'Low Depression' (PHQ scores 1,2 and 3) and High Depression' (PHQ scores 4 and 5). 185

subjects were labeled as 'Low Depression' (63.4% of the sample) and 107 subjects were labeled as 'High Depression' (36.6% of the sample).
The Vocal depression scores were also grouped into two categories: 'Low depression risk' (vocal scores 1-6) and 'High depression risk' (vocal scores 7-10). 175 subjects were labeled 'Low depression risk' (59.9% of the sample) and 117 subjects were labeled 'High depression risk' (40.1% of the sample).

As explained in the Methods paragraph above, the Vocal depression model was developed using a training sub-sample (on which the model was trained) and a test sub-sample (on which the model was tested).
Hence, the predictive power of the model would be best evaluated by the confusion matrix results for the Test sub-sample. The differences between the Training and the Test sub-samples also give an indication of the expected stability of the model.

Tables 2 and 3 show the confusion matrix results for the Training sub-sample (N=192), and tables 4 and 5 show the confusion matrix results for the Test sub-sample (N=100).

As described in the methods section above, the training and test sub-samples were randomly allocated 10 times, so the displayed results in tables 2-5 show the average results over all 10 iterations.

Training Sub-Sample

Training sample (N=100)		**PHQ-9 Depression**		
		Low	High	Total
Vocal Depression	Low risk	89.2	24.7	113.9
	High risk	31.3	46.8	78.1
	Total	120.5	71.5	192

Table 2. PHQ-9 Depression and Vocal depression labeled matrix (Training sub-sample)

Training sub-sample		
Accuracy	% correct classifications	**70.8%**
Sensitivity (recall)	True positive rate	**65.5%**
Specificity	True negative rate	**74.0%**
Precision	Positive predictive value	**59.9%**
FPR	False positive rate	**26.0%**
FNR	False negative rate	**34.5%**

Table 3. Confusion matrix attributes for the Training sub-sample

Training Sub-Sample

Training sample (N=100)		**PHQ-9 Depression**		
		Low	High	Total
Vocal Depression	Low risk	46.8	14.3	61.1
	High risk	17.7	21.2	38.9
	Total	64.5	35.5	100

Table 4. PHQ-9 Depression and Vocal depression labeled matrix (Test sub-sample)

Test sub-sample		
Accuracy	% correct classifications	**68.0%**
Sensitivity (recall)	True positive rate	**59.7%**
Specificity	True negative rate	**72.6%**
Precision	Positive predictive value	**54.5%**
FPR	False positive rate	**27.4%**
FNR	False negative rate	**40.3%**

Table 5. Confusion matrix attributes for the Test sub-sample

As can be seen, the confusion matrix results for the Test sub-sample (table 5) are quite similar to the results of the Training sub-sample (table 3).
The overall accuracy of the model is 70.8% in the training sub-sample and 68% in the Test sub-sample. This means that the Vocal depression model is rather stable and that it is expected to provide an overall accuracy close to 70% when used for Depression screening and tracking.

CONCLUSIONS

The study's results demonstrated strong and significant relationship between the depression state of patients as measured by psychiatric evaluations (PHQ-9) and between the Vocal depression scores as measured by the vocal Voicesense analysis system.

As can be seen, the confusion matrix results for the Test sub-sample (table 5) are quite similar to the results of the Training sub-sample (table 3). The overall accuracy of the model is 70.8% in the training sub-sample and 68% in the Test sub-sample. This means that the Vocal depression model is rather stable and that it is expected to provide an overall accuracy close to 70% when used for Depression screening and tracking.

The results received from the confusion matrix analysis suggest that the predictive power of the system is expected to provide about 70% overall accuracy. About 60% of the patients marked as 'High depression' by the PHQ, are expected to be classified accordingly by the vocal system (sensitivity). About 73% of the patients marked as 'Low depression' by the PHQ are expected to be classified accordingly by the vocal system (specificity). About 55% of the patients marked as high risk by the vocal system, are expected to be classified accordingly by the PHQ-9 (precision).

The results obtained in the study were all based upon one vocal sample of the patient (two minutes in length).The system operational mode for patient tracking would collect on-going vocal samples of the patients on a daily/weekly basis. The analysis results would be averaged over time (e.g. weekly reports) and would therefore provide even better accuracies than presented in the study.

Moreover, the results obtained in the study identified general vocal patterns common to all depressed patients. In the system's operational mode for patient tracking, the system would use the patient's typical vocal patterns as individual base-line reference, rather than the generic depression vocal patterns, and would therefore provide even better accuracies than presented in the study.