Discriminating between Patients with Unipolar disorder, Bipolar Disorder and Healthy Control Individuals based on Voice Features Collected from Naturalistic Smartphone Calls

Faurholt-Jepsen, Maria; Rohani, Darius Adam; Busk, Jonas; Tønning, Morten Lindberg; Vinberg, Maj; Bardram, Jakob Eyvind; Kessing, Lars Vedel

Published in:
Acta Psychiatrica Scandinavica

Link to article, DOI:
10.1111/acps.13391

Publication date:
2022

Document Version
Peer reviewed version

DR MARIA FAURHOLT-JEPSEN (Orcid ID : 0000-0002-0462-6444)
DR MAJ VINBERG (Orcid ID : 0000-0002-5982-1335)
PROFESSOR LARS VEDEL KESSING (Orcid ID : 0000-0001-9377-9436)

Article type : Original Article

Discriminating between Patients with Unipolar disorder, Bipolar Disorder and Healthy Control Individuals based on Voice Features Collected from Naturalistic Smartphone Calls

Running title: Voice analyses in affective disorder

Authors

Maria Faurholt-Jepsen[1, *, a], Darius Adam Rohani[2, a], Jonas Busk[3], Morten Lindberg Tønning[1], Maj Vinberg[1,4], Jakob Eyvind Bardram[2], Lars Vedel Kessing[1]

[1] *Copenhagen Affective Disorder Research Center (CADIC), Psychiatric Center Copenhagen, Rigshospitalet, Copenhagen, Denmark*

[2] *Department of Health Technology, Technical University of Denmark, Lyngby, Denmark*

[3] *Department of Energy Conversion and Storage, Technical University of Denmark, Lyngby, Denmark*

[4] *Psychiatric Center North Zealand, Denmark*

[*] *Corresponding author:*
Maria Faurholt-Jepsen
Copenhagen Affective Disorder Research Center (CADIC), Psychiatric Center Copenhagen
Rigshospitalet
Blegdamsvej 9

DK- 2100 Copenhagen
Denmark
Telephone: + 45 3864 7073
E-mail: maria@faurholt-jepsen.dk

[a] *Shared first author*

Acknowledgements

The authors would like to thank the patients for participating in the studies, the nurses and PhD students involved in the studies.

Keywords

Voice analysis, Classification, Random Forest, Unipolar disorder, Bipolar Disorder, openSMILE

Abstract

Background

It is of crucial importance to be able to discriminate unipolar disorder (UD) from bipolar disorder (BD), as treatments, as well as course of illness, differ between the two disorders. Aims: to investigate whether voice features from naturalistic phone calls could discriminate between 1) UD, BD, and healthy control individuals (HC); 2) different states within UD.

Methods

Voice features were collected daily during naturalistic phone calls for up to 972 days. A total of 48 patients with UD, 121 patients with BD, and 38 HC were included. A total of 115483 voice data entries were collected (UD (n= 16454), BD (n= 78733), and HC (n =20296)). Patients evaluated symptoms daily using a smartphone-based system, making it possible to define illness states within UD and BD. Data were analyzed using random forest algorithms.

Results

Compared to BD, UD was classified with a specificity of 0.84 (SD 0.07) /AUC of 0.58 (SD 0.07) and compared to HC with a sensitivity of 0.74 (SD 0.10)/ AUC=0.74 (SD 0.06). Compared to BD during euthymia, UD during euthymia was classified with a specificity of 0.79 (SD 0.05)/ AUC=0.43 (SD 0.16).

Compared to BD during depression, UD during depression was classified with a specificity of 0.81 (SD 0.09)/ AUC=0.48 (SD 0.12). Within UD, compared to euthymia, depression was classified with a specificity of 0.70 (SD 0.31)/ AUC=0.65 (SD 0.11). In all models the user-dependent models outperformed the user-independent models.

Conclusions

The results from the present study are promising, but as reflected by the low AUCs, does not support that voice features collected during naturalistic phone calls at the current state of art can be implemented in clinical practice as a supplementary and assisting tool. Further studies are needed.

Significant outcomes

- The present study investigated the use of voice features collected during naturalistic phone calls in a large sample of patients with unipolar disorder, patients with bipolar disorder, and healthy individuals
- There was a low sensitivitity for discriminating between patients with unipolar disorder and patients with bipolar disorder using voice features
- Voice features rather specifically discriminated between unipolar disorder and bipolar disorder, and rather sensitively discriminated between unipolar disorder and healthy individuals
- Patients with unipolar disorder during euthymia or depression was classified with high specificity compared to patients with bipolar disorder during euthymia or depression
- The results from the present study, as reflected by the low AUCs, did not confirm that voice features at the current state of art can be implemented in clinical practice as a supplementary tool

Limitations

- It is possible that other configurations of voice feature extraction than the one used in the present study could be feasible while keeping or improving the classification
- We did not have access to voice features from communication using other smartphone-based platforms
- The trade-off between sensitivity and specificity in the present study was reflected by the low AUCs

Introduction

Unipolar disorder (UD) and bipolar disorder (BD) are characterized by recurrent affective episodes with significant alterations in core features of mood, activity and sleep [1]. It is of crucial importance to be able to discriminate UD from BD, as pharmacological and psychological treatments, as well as the course of illness, differ between the two disorders [2–4]. Patients with UD are according to guidelines treated with antidepressants [5] and/or individual psychotherapy [6], whereas patients with BD may develop hypomania/mania, mixed episodes and mood destabilisation during treatment with antidepressants [7] and should initially be treated with lithium [8,9] and group-based psychoeducation [10]. When patients present in a remitted or depressive state it may be difficult for clinicians to reveal whether they suffer from UD or BD. Furthermore, patients and relatives may not recall prior (hypo)manic episodes. In this way, the diagnosis of BD could be overlooked. Studies have found that a diagnosis of UD over time may change to BD [11,12]. Clinical evaluations of prior (hypo)manic episodes are based on the patient's subjective experience and are potentially influenced by (depressive) recall bias or other recall distortions [13]. Consequently, it would be helpful for clinicians to add an objective measure that can assist in the discrimination between the two disorders considering the current state of illness.

Speech patterns have been shown to provide indicators of mental disorders. In 1921, Emil Kraepelin described that patients with depression tended to have a lower pitch, lower speech rate and more monotonous speech [14]. More recently, speech pause times have been suggested as pragmatically useful objective pathophysiologic markers in depression [15,16]. Digital phenotyping refers to approaches in which personal data gathered from mobile devices and sensors is analyzed to provide health information on physiological functions, or behavioral indicators, such as the user's speech [17]. Software for ecologically extracting data on voice features from naturalistic phone calls has been developed [18]. Smartphones comprise an available platform for detailed remote real-time monitoring of patient-reported symptoms such as mood through Ecological Momentary Assessments (EMAs) [19,20].

Recent systematic reviews and original studies concerning automated assessment of psychiatric disorders using speech suggested that speech processing technology could aid mental health assessments [21], and will be a key component in the search for objective markers for depression and monitoring of depression severity in the future [22–25]. A study including patients with schizophrenia, BD and UD found that speech variability measures collected from computers generally did not differ between the three groups, but differed compared to controls in average speech pause length [26].

Based on voice features collected during naturalistic phone calls clinicians would potentially get accurate and objective real-time data on the patients' states. This could provide opportunities for diagnosis and symptoms monitoring during long-term outside clinical settings and give possibilities for an individual intervention strategy between outpatient visits.

To the best of our knowledge, no study has investigated whether voice features collected from naturalistic phone calls can discriminate between UD, BD, and healthy control individuals (HC). Smartphone-based voice technology collected during naturalistic settings and between outpatient visits could potentially aid clinicians in differentiating between UD and BD and in identifying and targeting symptoms and upcoming affective states in the disorders. The authors have previously investigated differences in voice features according to state within BD [27].

Objectives

Using voice features collected from naturalistic phone calls the aims of the present study were to investigate whether these data 1A) could discriminate between patients with UD, patients with BD, and HC; 1B) could discriminate between patients with UD during euthymia and patients with BD during euthymia; 1C) could discriminate between patients with UD during depression and patients with BD during depression; 2) within patients with UD could discriminate between a) depression and euthymia; b) periods with decreased activity and neutral activity; c) periods with insomnia and periods without, and d) periods with combined decreased mood and decreased activity and periods without.

We hypothesized that voice features would be able to discriminate between patients with UD, patients with BD and HC with a sensitivity and specificity around 0.80 in all cases. Further, we hypothesized, that within patients with UD voice features would be able to discriminate between affective states with a sensitivity and specificity around 0.80 in all cases.

Material and Methods

Study design, settings, and participants

The present study included data collected as part of two studies - the RADMIS trials [28,29] and the currently ongoing Bipolar Illness Onset cohort study (the BIO study) [30], which were conducted during the period from 2017 to 2020. All participants underwent The Schedules of Clinical Assessment in Neuropsychiatry (SCAN) interview [31] to confirm the clinical diagnosis of unipolar disorder or bipolar disorder (or the lack of).

The RADMIS trials: Patients with a diagnosis of UD or BD who were hospitalized due to an affective episode and subsequently discharged from one of five psychiatric centers at the Mental Health Services, Capital Region of Denmark, Denmark in the period from May 2017 to August 2019 were invited to participate in the trial. Inclusion criteria were age above 18 years, UD or BD diagnosis according to the ICD-10 anddischarge from a psychiatric hospital in The Capital Region of Denmark following an affective episode (depression, mania or mixed episode). Exclusion criteria were pregnancy and a lack of Danish language skills. In the RADMIS trials, in addition to standard treatment, patients with UD or BD were randomized with a balanced allocation ratio to either 1) daily use of a smartphone-based monitoring system (the Monsenso system – se details below) (the intervention group) or to 2) normal use of smartphones (the control group) during a six months follow-up period.

The BIO study: From the BIO study two groups of participants were included: Patients with newly diagnosed BD, and HC. *Patients with BD:* Inclusion criteria were a newly diagnosis of a single manic episode or bipolar disorder according to the ICD-10 and ages between 15 to 70 years. *HC:* HC were recruited among blood donors, aged between 15 to 70 years, from the Blood Bank at Rigshospitalet, Copenhagen. All participants in the BIO study were offered to use the Monsenso system on a daily basis during the study period.

Patient-reported smartphone-based data

The Monsenso system comprised a smartphone-based monitoring system that was installed on the patients' own smartphones (both iPhone and Android smartphones). The smartphone-based monitoring system was developed by the authors and used by the patients on a daily basis to collect fine-grained real-time recordings of mood, activity, and sleep duration [32]. In patients with UD, mood was evaluated with scores on a five-point scale from depressed to euthymic (-3, -2, -1, -0.5, 0). In patients with BD, mood was evaluated with scores on a 9-point scale from depressed to manic (-3, -2, -1, -0.5, 0, 0.5, 1, 2 ,3). Euthymic

mood was defined as a mood score of -0.5, 0 or 0.5. Depressive mood was defined as mood score < -0.5. In patients with UD, daily activity levels were rated on a scale from (-3, -2, -1, 0, 1, 2, 3), with 0 representing normal activity level. Sleep duration was calculated based on daily reports of bedtime and wake-up time. Insomnia was defined as total sleep duration < 360 min. Combined decreased mood and activity were defined as mood < 0.5 and activity < 0.

Voice features

Voice features were collected from the participants' phone calls during their everyday life using the open-source Speech and Music Interpretation by Large-space Extraction (openSMILE v. 2.1.0, Emo-Large) toolkit (availale for Android smartphones) [18,33]. The toolkit is a feature extractor for signal processing and machine learning applications, and it is designed for real-time processing. The toolkit processed voice samples from each incoming and outgoing phone call on the participants' smartphones to extract acoustic features. No recordings of the raw speech or sound data were available. Thus, once the acoustic features were extracted locally on the phone, the voice recording was discarded. The Emo-Large configuration was a predefined set consisting of 6552 features, e.g., pitch, loudness, and energy, represented through various 1st level descriptive statistics including means, regression coefficients, and percentiles. The set has been found to be particularly relevant for classifying emotions [34].

Clinical assessments

Clinical evaluations of the severity of depressive and manic symptoms were conducted by trained medical doctors using the Hamilton Depression Rating Scale 17-items (HDRS) [35] and the Young Mania Rating Scale (YMRS) [36].

Statistical analyses

In patients with UD, smartphone-based data for any specific day during the study period were included in the analyses if both voice features and patient-reported smartphone-based data on mood, activity or sleep were available for the same day.

Aim 1 concerned the discrimination between patients with UD, patients with BD, and HC based on the use of voice features collected from naturalistic phone calls. *Aim 2* concerned the use of voice data from

patients with UD to classify the symptom class labels within mood, activity, and sleep collected daily from smartphones, and a combination of mood and activity.

For all analyses, we built Random Forest (RF) classifiers to discriminate between groups/classes [37]. The RF classifier is an ensemble method that combines several decision tree classifiers into a single classifier (the 'forest'). We chose the RF model as it is generally able to handle a large number of features while being robust to overfitting. All classifications were kept binary, e.g., analyzing the population class patients with UD versus patients with BD, or the symptom class `depressed' against `euthymic' (for further specifications of RF see supplements section a). For aim 2, voice data for days without a corresponding patient-reported smartphone-based data entry of either mood, activity, or sleep were removed. We ran RF models on the resulting data set through a 5-fold participant-based cross-validation (for further specifications of the cross-validation scheme and hyper paramteters see supplements section).

Analyses for aim 2 were separated in two model types. First, a user-independent model that - as for aim 1 - combined all patients in the same model. The model uses information from previous participants to classify symptoms of `new', unknown, patients. Second, a user-dependent model where we built a personalized model for each patient. In this case the results represent a mean (M) and standard deviation (SD) across all the patient models.

We observed significant class imbalance in the data for all aims (e.g. fewer cases of symptoms of `depression', and `insomnia' compared with neutral). Therefore, we applied a resampling process on the training training data to balance the two classes. We did a combination of SMOTE oversampling [38] of the minority class to represent 33% of the cases, followed by random under sampling of the majority class until the sample size was identical to the minority class. The combination of oversampling and under sampling with SMOTE has previously been shown effective to counter class imbalance [39]. Without a resampling scheme, the RF classifier would favor the overrepresented class, which we also observed in our data. However, resampling was only performed on the training data, to keep the test set class distribution representative for the collected data. In the cases where class distribution was less than 33% skewed, we only ran random under sampling.

Data were imported to and processed in Python (version 3.8) with packages sklearn (v. 0.23.2), imblearn (v. 0.7.0), and pandas (v. 1.1.4). For all built-in functions (e.g., sklearn's implementation of the RF classifier) we used the default parameters, unless otherwise stated.

Model performance

To evaluate the RF classifier performance, we applied several standard metrics for binary classification computed on a test set held out data and compared the results to a majority vote baseline model. The metrics include a) `accuracy', defined as the number of correct classifications of the positive and negative cases divided by the total number of cases; b) `F1-score', a measure that estimates the model's ability to identify the positive (1) class correctly. It is a balanced measure of classifiers `precision' and `recall' and defined as the true positives divided by the true positives and the average between false positives and false negatives; c) `sensitivity' (SE), defined as true positives divided with positives; d) `specificity' (SP) defined as true negatives over all negatives; e) `area under the characteristic curve' (AUC). All classification metrics were computed within each cross-validation fold to yield a mean (M) and standard deviation (SD) value across all 5 folds. In the personalized model we further averaged across all patients.

For aim 1, we ran a randomized permutation model [40] to test whether voice data from the three populations were statistically significantly different from each other. We randomly shuffled the class label for each participant and re-ran the entire RF classification. This was repeated 200 times to generate a non-parametric null-distribution of AUC scores (**Figure 1**). Statistically significance was determined if the RF test AUC statistics with true class labels exceeds the null distribution with a significance level of alpha = 0.05. Such permutation test is computational heavy but has several advantages from traditional parametric tests (e.g., ANOVA) as it is independent of normality assumptions about the underlying data, and thus helps to statistically quantify whether the voice data from the two compared populations were drawn from the same distribution.

For aim 2, we developed a majority vote model as a baseline. Unlike the RF model, the majority vote did not include voice data. Simply, the most frequently observed class label in the training data, was used to classify test data. In cases where there was an equal class distribution, the test data was classified at random.

Results

Participant flow and background characteristics

Overall, a total of 242 participants were included in the present study. Some of the participants did not provide voice data (n= 35/242) leaving a total of 207 participants for analyses (UD, n=48; BD, n= 121; HC, n= 38). A total of 115.483 voice data recordings across UD (n= 16.454), BD (n= 78.733), and HC (n= 20.296) were available. The participants provided on average 157 (SD 174.5) days with voice data, ranging between 1 and 972 days. The participants had a mean age of 37.3 (SD 13.6) years, ranging from 18 to 78 years. A total of 57% (n=117) were women with no statistically significant difference in proportions across the three populations. A total of 63% (n=30) of patients with UD had a HDRS score ≥ 13, and 30% (n=15) had a HDRS score > 17. A total of 41% (n=49) of patients with BD had a HDRS score ≥ 13, and 19% (n=23) had a HDRS score > 17. There was a statistically significant difference in age, proportion of students, and HDRS score between patients with UD and patients with BD. Further background characteristics are presented in Table 1.

Voice features for classification of diagnostic groups

Table 2 present the results for the classification of patients with UD (16454 observations), patients with BD (78731 observations), and HC (20296 observations) based on voice features collected from naturalistic phone calls.

The sensitivity and specificity for classifying patients with UD versus HC were 0.74 (SD 0.10) and 0.56 (SD 0.06), respectively and with an AUC of 0.74 (SD 0.06). The sensitivity and specificity for classifying patients with UD versus patients with BD was 0.27 (SD 0.14) and 0.84 (SD 0.07), respectively and with an AUC of 0.58 (SD 0.07). The sensitivity and specificity for classifying patients with UD during a euthymic state versus patients with BD during a euthymic state were 0.18 (SD 0.19) and 0.79 (SD 0.05), respectively and with an AUC of 0.43 (SD 0.16). The sensitivity and specificity for classifying patients with UD during a depressive state versus patients with BD during a depressive state were 0.16 (SD 0.09) and 0.81 (SD 0.09), respectively and with an AUC of 0.48 (SD 0.12).

Figures 1 A & B present the generated null-distribution of AUC from permuted class labels together with the AUC presented in Table 2. The lighter area shows the critical level for a one-tail test with a significance level of 0.05. The observed AUC for patients with UD versus HC (4/200, p=0.02) differed statistically

significantly. Thus, patients with UD and HC were sampled from two distinct populations that were statistically significantly different. The observed AUC for patients with UD versus patients with BD (28/200, p=0.14) did not statistically significantly differ from each other.

Voice features for classifications of affective states within unipolar disorder

A total of 33 patients with UD provided both voice features and patient-reported smartphone-based data with a range of 3 to 220 days. Table 3 present the results for the classification of different states in patients with UD. In all the models presented in Table 3, the personalized user-dependent models outperformed the general user-independent models. Therefore, the results from the user-dependent models are presented below. Figure 2 presents the associations between patient-reported mood and clinical ratings of depressive symptoms according to the total HDRS ($r = -0.59$, $p < 0.001$) and the HDRS subitem 1 (mood, $r = -0.64$, $p < 0.001$).

Depression (2838 observations) versus euthymia (4504 observations): The sensitivity and specificity for classifying depression versus euthymia were 0.49 (SD 0.31) and 0.70 (SD 0.31), respectively and with an AUC of 0.65 (SD 0.11). *Increased activity (3570 observations) versus neutral activity (1708 observations):* The sensitivity and specificity for classifying increased activity versus neutral activity were 0.61 (SD 0.28) and 0.52 (SD 0.30), respectively and with an AUC of 0.62 (SD 0.11). *Decreased activity (2372 observations) versus neutral activity:* The sensitivity and specificity for classifying decreased activity versus neutral activity were 0.63 (SD 0.29) and 0.49 (SD 0.29), respectively and with an AUC of 0.60 (SD 0.12). *Insomnia (1129 observations) versus periods without (6381 observations):* The sensitivity and specificity for classifying insomnia versus periods without was 0.30 (SD 0.25) and 0.79 (SD 0.21), respectively and with an AUC of 0.60 (SD 0.13). *Combined depression and decreased activity (1310 observations) versus periods without (5675 observations)*: The sensitivity and specificity for classifying combined depression and decreased activity versus periods without was 0.53 (SD 0.18) and 0.69 (SD 0.14), respectively and with an AUC of 0.64 (SD 0.13). We investigated the error for each user-independent model to explore any possible biases across patients. Through several patient specific covariates (e.g., sex and age) we observed a bias in illness duration for the depression versus euthymia model. The classification error had patient samples with significantly higher illness duration ($M = 23.95$, $SD = 16.94$), than the correct classifications ($M = 14.93$, $SD = 11.67$; $t(3700) = 18.24$, $p < 0.001$).

The Receiver Operating Characteristic (ROC) curves for each model are presented in **Figure 3.** The ROC curve was generated from the aggregated result of all model class estimates and the corresponding true class in each cross-validation fold, as well as each patient in the user-dependent classifiers, are presented in **Figure 3**. As can be seen, the ROC curve for the combined model including decreased mood and decreased activity versus periods without was the closest to random, while the model including decreased mood versus euthymia performed best.

Discussion

The present large and innovative study used voice features collected from naturalistic phone calls for classifications of patients with UD, patients with BD and HC, and for state classifications within patients with UD. In contrast to our hypotheses voice features discriminated between patients with UD and patients with BD with low sensitivity and AUC. In accordance with our hypotheses, voice features discriminated between patients with UD and patients with BD with a rather high specificity. Notably, the AUCs based on voice features did not statistically significantly differ from patients with unipolar disorder and patients with bipolar disorder. Voice features discriminated with rather high sensitivity between patients with UD and HC, but in contrast with our hypotheses with low specificity. Patients with UD during euthymia were classified with a low sensitivity and rather specifically compared to patients with BD during euthymia. The same results were found when looking at patients with UD during depression compared to patients with BD during depression. Looking within patients with UD, compared to euthymia, in contrast to our hypotheses depression was classified with a modest specificity. Insomnia was classified with a rather high specificity compared to periods without. In all analyses within patients with UD, the user-dependent models outperformed the user-independent models suggesting that changes in voice features are highly individual. In clinical practice, it is often difficult to differentiate UD from BD when patients are in a remitted or mild/moderate depressive state. Patients may not recall or may even deny prior (hypo)manic episodes and clinicians may not be sufficiently observant on the prior course of illness. Thus, a subset of patients diagnosed with UD in clinical practice in fact suffers from BD [2,41,42]. Other studies have shown that one third of patients with BD do not get a correct BD diagnosis until at least 10 years from illness onset [43–45]; the most common incorrect diagnosis being UD [46]. Furthermore, most patients with BD seek treatment for depression, and not for (hypo)mania [47] adding to the frequency of misdiagnoses. The authors have previously suggested that objective measures of psychomotor activity may add to discriminating between UD and BD [48]. The results from the present study suggest that real-time monitoring and analyses of voice features may provide opportunities to rather specifically diagnose and differentiate between UD and BD, but at a cost with a low sensitivity. Thus, the tradeoff between the sensitivity and specificity (as reflected by the low AUCs) should be considered in future studies and considerations on whether a high sensitivity could be important even though it could be at a cost of lower specificity – or the other way around - and thereby the risk of false-positive or false-negative classifications of patients or of affective states. Although, it is important to state that the variability between participants was high. For some patients the model performed excellent in discriminating symptoms based on the recorded voice features (AUC values in the

order of 0.9) while others were as good as randomly guessing. Such observation could indicate that specific symptoms (e.g., insomnia) might have a larger effect on the recorded voice for some than for others.

Advantages and limitations

Several clinical as well as methodological limitations to the present study should be mentioned. Case-control studies carry an inherent risk of bias at different levels, such as selection bias, information bias and confounding, necessitating strict methodological requirements and thorough considerations of the study design and analyses. The present study was the first to include a large sample of both well-charatcerised patients with UD, patients with BD, and HC with repeated and fine-grained data during a rather long follow-up period. The available voice features were collected unobtrusively as natural speech samples during naturalistic settings reducing the Hawthorne effect (the risk of change in behavior simply due to being monitored (in this case a change in how they speak)) [49], which could possible not have been the case if the voice features were collected during laboratory settings. Collecting speech samples during free living introduces a risk of collecting speech samples from individuals borrowing the patients' smartphone during the study period. In addition, the patients might not have used their smartphone for phone calls during more severe illness states and thereby not providing data during these times, which could have been possible if the samples were collected in a laboratory.

The included populations were well-characterized according to clinical as well as research-based assessments using the SCAN interview. The affective states within patients with UD and BD were defined according to scores on daily smartphone-based patient-reported mood, activity and sleep. As a result, the changes in patient-reported mood were mapped to clinical symptom changes when observing its relation to voice features. Moreover, smartphone-based patient-reported mood was found to be associated with scores on the HDRS-17 and HDRS subitem 1. Voice data for days without a corresponding patient-reported smartphone-based data entry of either mood, activity, or sleep were removed. Thus, there may be a risk that the patient-reported smartphone-based data were not missing at random, and thus potentially we did not include voice features during the most severe affective states. In addition, in the present study there was an overrepresentation of lower smartphone-based patient-reported mood scores, which limit the predictive value and generalizability of findings in more severe cases. Thus, the association between smartphone-based patient-reported mood and clinical evaluations of the severity of depressive and manic symptoms may be overestimated or underestimated in the present study and should be further investigated in future studies.

The present study included the entire openSMILE emolarge feature set. It is possible that other configurations of the openSMILE toolkit or other feature extraction technologies, and subsequent feature selection, than the ones used in the present study could be feasible while keeping or improving the classification. However, no references exist that presented a subset of the emolarge feature set that better discriminate mood-related symptoms. Due to a lower number of extreme cases (e.g., mood ratings above or below neutral) we did not consider to perform feature selection on an isolated fraction of the data. Further, while RF is a popular model and known to perform well with a large number of input features. We encourage future work to compare different models for openSMILE-based feature classification in affective disorders and use feature engineering to select more representative information from raw voice data. The voice features were extracted during regular phone calls on Android-based smartphones only, and thus we did not have access to voice features from communication using other smartphone-based platforms. Defining and recruiting a proper control group in case-control studies is always difficult. The HC included in the present study were recruited from the Blood Bank at Rigshospitalet, Copenhagen University Hospital, Denmark, and thus may represent a 'super healthy' comparison group. Alternative methods for recruiting control groups include using advertisements or the Danish Civil Registration System. However, both methods have relatively low participation response rates and a high risk of selection bias. Taken together, we find that our control group represents the most reasonable and assessable control group for this study. Furthermore, a potential confounding effect of differences in psychopharmacological treatment between patients with UD and BD cannot be ruled out [50]. Also, other factors that are not related to the mental health status of the participants such as work status could influence the results. Future observational studies could consider investigating this aspect further. Lastly, within patients with UD insomnia was defined as a total sleep duration < 360 min. In the present study, we did not have access to information on awakenings during the night or poor sleep quality that could have provided more insights into the issues on insomnia. However, the cut off of <360 min sleep per night was chosen as a pragmatic solution, with the risk of overlooking other aspects of insomnia.

The findings that, within patients with UD, the user-dependent models outperformed the user-independent models suggested that changes in voice features are highly individual. Thus, in clinical practice it would be necessary to collect voice features for each patient during a follow-up period before it would be possible to conduct the user-dependent models. However, due to the longitudinal design of the studies included we were able to conduct this type of analyses and identify this aspect in contrast to other studies of cross-sectional nature.

Perspectives and future implications

Voice feature collection during free living may reflect an innovative, objective and unobtrusive supplementary method for discriminating between patients with UD and patients with BD. Thus, using voice features for monitoring may provide an opportunity for clinicians to differentiate more accurately between the two disorders and allow for earlier correct diagnosis and treatment. Furthermore, voice features could potentially reflect an objective and supplementary real-time method for monitoring changes in state within patients with UD and BD [27].

Conclusion

The present innovative study investigated the use of voice features collected during naturalistic phone calls in patients with UD, patients with BD, and HC. Most significantly, voice features discriminated between UD and BD with low sensitivity but rather specifically. Voice features with a modest sensitivity discriminated between UD and HC, but with low specificity. In addition, patients with UD during euthymia were classified with low sensitivity compared to patients with BD during euthymia, but with rather high specificity. The same was found when looking at patients with UD during depression compared to patients with BD during depression. Looking within patients with UD, insomnia was classified with a rather high specificity compared to periods without. In all analyses within patients with UD, the user-dependent models outperformed the independent user models suggesting that changes in voice features is highly individual. The trade-off between sensitivity and specificity in the present study was reflected by the low AUCs. Due to the lack of objective markers, the clinical diagnostic process relies on patient information and information from relatives, clinical observations and evaluations. Overall, the results from the present study are promising, but as reflected by the low AUCs, do not support that voice features collected during naturalistic phone calls at the current state of art can be implemented in clinical practice as a supplementary and assisting tool. Further studies are needed.

Disclosures

MFJ, DR, JoB, and MLT have no competing interests. MV has within the last three years been a consultant for Lundbeck, Sunovion and Janssen-Cilag. LVK has been a consultant for Lundbeck and Teva within the past three years. JB is a co-founder and shareholder in Monsenso.

Funding sources

The RADMIS trial was funding by Innovation Fund Denmark (5164-00001B9). The BIO study was funded by grants from the Mental Health Services, Capital Region of Denmark, The Danish Council for Independent Research, Medical Sciences (DFF—4183-00570), Weimans Fund, Markedmodningsfonden (the Market Development Fund, (2015-310), Gangstedfonden (A29594), Helsefonden (16-B-0063), Innovation Fund Denmark (the Innovation Fund, Denmark, 5164-00001B), Copenhagen Center for Health Technology (CACHET), EU H2020 ITN (EU project 722561), Augustinusfonden (16-0083), Lundbeck Foundation (R215-2015-4121).

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

1. Goodwin, F. K. & Jamison, K. R. *Manic-Depressive illness*. vol. 1996 (New Oxford University Press, 1996).
2. Kessing, L. V. The effect of the first manic episode in affective disorder: a case register study of hospitalised episodes. *J Affect Disord* **53**, 233–239 (1999).
3. Kessing, L. V., Andersen, P. K., Mortensen, P. B. & Bolwig, T. G. Recurrence in affective disorder. I. Case register study. *Br J Psychiatry* **172**, 23–28 (1998).
4. Kessing, L. V., Hansen, M. G. & Andersen, P. K. Course of illness in depressive and bipolar disorders. Naturalistic study, 1994-1999. *Br J Psychiatry* **185**, 372–377 (2004).
5. Bauer, M. *et al.* World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, part 1: update 2013 on the acute and continuation treatment of unipolar depressive disorders. *World J Biol Psychiatry* **14**, 334–385 (2013).
6. Parikh, S. V. *et al.* Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder. *Can J Psychiatry* **61**, 524–539 (2016).
7. Pacchiarotti, I. *et al.* The International Society for Bipolar Disorders (ISBD) task force report on antidepressant use in bipolar disorders. *Am J Psychiatry* **170**, 1249–1262 (2013).
8. Yatham, L. N. *et al.* Canadian Network for Mood and Anxiety Treatments (CANMAT) and International Society for Bipolar Disorders (ISBD) 2018 guidelines for the management of patients with bipolar disorder. *Bipolar Disord* **20**, 97–170 (2018).
9. Kessing, L. V. Lithium as the drug of choice for maintenance treatment in bipolar disorder. *Acta Psychiatrica Scandinavica* **140**, 91–93 (2019).
10. Soo, S. A. *et al.* Randomized Controlled Trials of Psychoeducation Modalities in the Management of Bipolar Disorder: A Systematic Review. *J Clin Psychiatry* **79**, (2018).
11. Kessing, L. V. Diagnostic stability in depressive disorder as according to ICD-10 in clinical practice. *Psychopathology* **38**, 32–37 (2005).
12. O'Donovan, C. & Alda, M. Depression Preceding Diagnosis of Bipolar Disorder. *Front Psychiatry* **11**, (2020).
13. Moffitt, T. E. *et al.* How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychol Med* **40**, 899–909 (2010).
14. Lord, J. R. Manic-depressive Insanity and Paranoia. By Prof. Emil Kraepelin; translated by R. Mary Barclay, M.A., M.B.; edited by George M. Robertson, M.D., F.R.C.P.Edin. Edinburgh: E. & S. Livingstone, 1921. Demy 8vo. Pp. 280. Forty-nine illustrations, eighteen in colour. Price 12s. 6d. *Journal of Mental Science* **67**, 342–346 (1921).
15. Greden, J. F., Albala, A. A., Smokler, I. A., Gardner, R. & Carroll, B. J. Speech pause time: a marker of psychomotor retardation among endogenous depressives. *Biol. Psychiatry* **16**, 851–859 (1981).
16. Greden, J. F. & Carroll, B. J. Decrease in speech pause times with treatment of endogenous depression. *Biol. Psychiatry* **15**, 575–587 (1980).
17. Insel, T. R. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA* **318**, 1215–1216 (2017).
18. Eyben, F., Wöllmer, M. & Schuller, B. openSMILe- The Munich Versatile and Fast Open.Source Audio Feature Extractor. in *Prooceedings of ACM Multimedia* vol. 2010 (2010).
19. Colombo, D. *et al.* Current State and Future Directions of Technology-Based Ecological Momentary Assessment and Intervention for Major Depressive Disorder: A Systematic Review. *J Clin Med* **8**, (2019).
20. Trull, T. J. & Ebner-Priemer, U. Ambulatory Assessment. *Annu Rev Clin Psychol* **9**, 151–176 (2013).

21. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig Otolaryngol* **5**, 96–116 (2020).
22. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. *Speech Communication* **71**, 10–49 (2015).
23. Horwitz, R. *et al.* On the relative importance of vocal source, system, and prosody in human depression. in *2013 IEEE International Conference on Body Sensor Networks* 1–6 (2013). doi:10.1109/BSN.2013.6575522.
24. Kiss, G. & Vicsi, K. Mono- and multi-lingual depression prediction based on speech processing. *Int J Speech Technol* **20**, 919–935 (2017).
25. Quatieri, T. & Malyska, N. Vocal-source biomarkers for depression: A link to psychomotor activity. *Proceedings of Interspeech* **2**, 1059–1062 (2012).
26. Cohen, A. S., McGovern, J. E., Dinzeo, T. J. & Covington, M. A. Speech deficits in serious mental illness: a cognitive resource issue? *Schizophr Res* **160**, 173–179 (2014).
27. Faurholt-Jepsen, M. *et al.* Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry* **6**, e856 (2016).
28. Faurholt-Jepsen, M. *et al.* Reducing the rate of psychiatric Re-ADMISsions in Bipolar Disorder using smartphones The RADMIS trial. *Acta Psychiatr Scand* (2020) doi:10.1111/acps.13274.
29. Tønning, M. L. *et al.* The effect of smartphone-based monitoring and treatment on the rate and duration of psychiatric readmission in patients with unipolar depressive disorder: The RADMIS randomized controlled trial. *J Affect Disord* **282**, 354–363 (2020).
30. Kessing, L. V. *et al.* The Bipolar Illness Onset study: research protocol for the BIO cohort study. *BMJ Open* **7**, e015462 (2017).
31. Wing, J. K. *et al.* SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Arch. Gen. Psychiatry* **47**, 589–593 (1990).
32. Bardram, J. E. *et al.* Designing mobile health technology for bipolar disorder: a field trial of the monarca system. in *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* pages 2627-2636 (2013).
33. Schuller, B. *et al.* The INTERSPEECH 2010 paralinguistic challenge. in 2794–2797 (2010).
34. Pfister, T. & Robinson, P. Speech Emotion Classification and Public Speaking Skill Assessment. in *Human Behavior Understanding* (eds. Salah, A. A., Gevers, T., Sebe, N. & Vinciarelli, A.) 151–162 (Springer, 2010). doi:10.1007/978-3-642-14715-9_15.
35. Hamilton, M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* **6**, 278–296 (1967).
36. Young, R. C., Biggs, J. T., Ziegler, V. E. & Meyer, D. A. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry* **133**, 429–435 (1978).
37. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
38. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2002).
39. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. & Herrera, F. Big data preprocessing: methods and prospects. *Big Data Analytics* **1**, 9 (2016).
40. Berry, K. J., Mielke, P. W. & Mielke, H. W. The Fisher-Pitman permutation test: an attractive alternative to the F test. *Psychol Rep* **90**, 495–502 (2002).
41. Hirschfeld, R. M. A. *et al.* Screening for bipolar disorder in the community. *J Clin Psychiatry* **64**, 53–59 (2003).
42. Calabrese, J. R. *et al.* Predictors of bipolar disorder risk among patients currently treated for major depression. *MedGenMed* **8**, 38 (2006).

43. Suppes, T. *et al.* The Stanley Foundation Bipolar Treatment Outcome Network. II. Demographics and illness characteristics of the first 261 patients. *J Affect Disord* **67**, 45–59 (2001).
44. Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A. & Hirschfeld, R. M. The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *J Affect Disord* **31**, 281–294 (1994).
45. Hirschfeld, R. M. A., Lewis, L. & Vornik, L. A. Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *J Clin Psychiatry* **64**, 161–174 (2003).
46. Kessing, L. V. Diagnostic stability in bipolar disorder in clinical practise as according to ICD-10. *J Affect Disord* **85**, 293–299 (2005).
47. Hirschfeld, R. M. Bipolar spectrum disorder: improving its recognition and diagnosis. *J Clin Psychiatry* **62 Suppl 14**, 5–9 (2001).
48. Faurholt-Jepsen, M. *et al.* Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state. *J Affect Disord* **141**, 457–463 (2012).
49. Wickström, G. & Bendix, T. The 'Hawthorne effect'--what did the original Hawthorne studies actually show? *Scand J Work Environ Health* **26**, 363–367 (2000).
50. Bock, J. M. Medications and Vocal Function. *Otolaryngol Clin North Am* **52**, 693–702 (2019).

Table 1. Background characteristics of participants, N= 207

	Patients with Unipolar Disorder (UD)	Patients with Bipolar Disorder (BD)	Healthy control Individuals (HC)	p
N, % female	48 (60.0)	121 (60.0)	38 (45.0)	UD: BD (p= 0.95) UD: HC (p= 0.22) BD: HC (p= 0.16)
Age, years	45.6 (14.9)	35.7 (12.3)	31.7 (10.9)	UD: BD (p< 0.001) UD: HC (p< 0.001) BD: HC (p= 0.21)
Years of education	14.0 (3.0)	13.6 (4.8)	15.6 (1.6)	UD: BD (p= 0.86) UD: HC (p= 0.16) BD: HC (p= 0.03)
HDRS[a] at inclusion	14.5 (5.5)	11.1 (6.9)	0.95 (1.6)	UD: BD (p=0 .003) UD: HC (p< 0.001) BD: HC (p< 0.001)
YMRS[b] at inclusion	N/A	3.8 (4.8)	0.51 (0.98)	BD: HC (p<0.001)
Previous depressive episodes, number	6 [1 - 45]	10 [1 - 80]	N/A	
Illness duration (years)	12.2 (12.3)	14.9 (10.4)	N/A	
Psychotropic medication				
Anticonvulsant, % (n)	12 (6)	42 (51)	N/A	
Lithium, % (n)	6 (3)	49 (59)	N/A	
Antipsychotics, % (n)	31 (15)	50 (60)	N/A	
Antidepressants, % (n)	100 (48)	20 (24)	N/A	

Data are mean (SD), median [IQR] or proportions (n, %) unless otherwise stated

[a] HDRS: Hamilton Depression Rating Scale 17-items score; [b] YMRS: Young Mania Rating Scale score

Table 2. Discrimination between patients with Unipolar Disorder (UD) (n= 48), patients with Bipolar Disorder (BD) (n= 121), and Healthy Control Individuals (HC) (n= 38) based on voice features collected from smartphones, N= 207

Binary classifier (n= number of observations)	Model type	Accuracy (SD)	F1 score (SD)	Sensitivity (SD)	Specificity (SD)	AUC (SD)	B10 (SD)
UD (n= 16454) compared with HC (n = 20296)	Random Forest model	0.63 (0.07)	0.60 (0.11)	0.74 (0.10)	0.56 (0.06)	0.74 (0.06)	5.62 (4.15)
	Majority vote	0.43 (0.00)	0.29 (0.23)	0.60 (0.49)	0.40 (0.49)	0.50 (0.00)	-3.10 (0.00)
UD (n= 16454) compared with BD (n = 78731)	Random Forest model	0.73 (0.07)	0.22 (.07)	0.27 (0.14)	0.84 (0.07)	0.58 (0.07)	-0.60 (1.3)
	Majority vote	0.82	0.00 (0.00)	0.00 (0.00)	1.0 (0.00)	0.50 (.00)	0.00 (0.00)
UD, euthymia (n = 4504) compared with BD, euthymia (n= 38328)	Random Forest model	0.76 (0.04)	0.03 (0.03)	0.18 (0.19)	0.79 (0.05)	0.43 (0.16)	-0.52 (2.18)
	Majority vote	.97 (.00)	.00 (.00)	.00 (.00)	1.0 (.00)	.50 (.00)	-3.84 (0.00)
UD, depression (n= 2838) compared with BD, depression (n= 5329)	Random Forest model	0.66 (0.15)	0.14 (0.05)	0.16 (0.09)	0.81 (0.09)	0.48 (0.12)	0.44 (3.92)

Table 3. Classification within patients with Unipolar Disorder (n= 48) according to daily smartphone-based patient-reported mood, activity, and sleep

Binary classifier (n= number of observations)	Model type	Accuracy (SD)	F1 score (SD)	Sensitivity (SD)	Specificity (SD)	AUC (SD)	B10 (SD)
MOOD							
Depression (n= 2838) versus Euthymia (n= 4504)	Random Forest model User independent	0.45 (0.11)	0.32 (0.14)	0.44 (0.19)	0.50 (0.22)	0.41 (0.10)	-1.12 (0.88)
	Random Forest model User dependent	0.78 (0.13)	0.50 (0.29)	0.49 (0.31)	0.70 (0.31)	0.65 (0.11)	3.44 (8.16)
	Majority vote User independent	0.59 (0.0)	0.00 (0.0)	0.00 (0.0)	1.0 (0.0)	0.50 (0.0)	-3.81 (0.0)
	Majority vote User dependent	0.91 (0.00)	0.25 (0.40)	0.29 (0.45)	0.71 (0.45)	0.48 (0.02)	-2.92 (0.0)
ACTIVITY							
Decreased (n= 2372) versus Neutral (n= 1708)	Random Forest model User independent	0.51 (0.09)	0.54 (0.11)	0.58 (0.19)	0.43 (0.22)	0.48 (0.11)	-0.72 (1.07)
	Random Forest model User dependent	0.69 (0.13)	0.62 (0.27)	0.63 (0.29)	0.49 (0.29)	0.60 (0.12)	2.65 (9.42)
	Majority vote User independent	0.67 (0.0)	0.69 (0.11)	1.0 (0.0)	0.0 (0.0)	0.50 (0.0)	-3.69 (0.0)
	Majority vote User dependent	0.65 (0.0)	0.56 (0.39)	0.68 (0.47)	0.32 (0.47)	0.47 (0.02)	-3.25 (0.0)
SLEEP							
Insomnia (n= 1129) versus normal sleep (n= 6381)	Random Forest model User independent	0.68 (0.09)	0.16 (0.09)	0.45 (0.15)	0.72 (0.10)	0.60 (0.10)	-0.87 (1.76)
	Random Forest model User Dependent	0.74 (0.18)	0.31 (0.24)	0.30 (0.25)	0.79 (0.21)	0.60 (0.13)	0.22 (3.26)
	Majority vote User independent	0.98 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.50 (0.0)	-2.85 (1.28)

	Majority vote User dependent	0.83 (0.0)	0.17 (0.32)	0.21 (0.41)	0.79 (0.41)	0.48 (0.02)	-3.17 (0.0)
	COMBINED LOW MOOD AND LOW ACTIVITY						
Combined low mood and low activity (n= 1310) versus rest (n= 5675)	Random Forest model User independent	0.60 (0.12)	0.18 (0.12)	0.28 (0.18)	0.67 (0.20)	0.46 (0.10)	-1.15 (0.76)
	Random Forest model User dependent	0.68 (0.14)	0.45 (0.18)	0.53 (0.18)	0.69 (0.14)	0.64 (0.13)	2.82 (5.56)
	Majority vote User independent	0.75 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (0.0)	0.50 (0.0)	-3.31 (0.0)

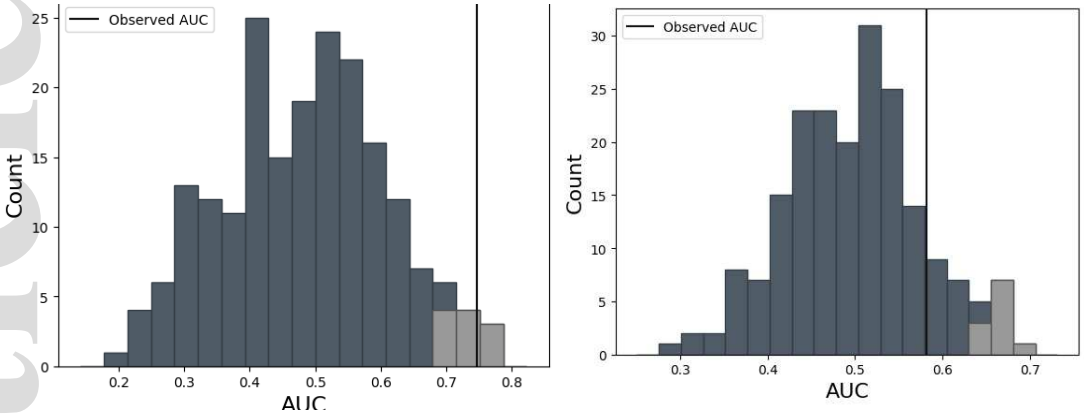
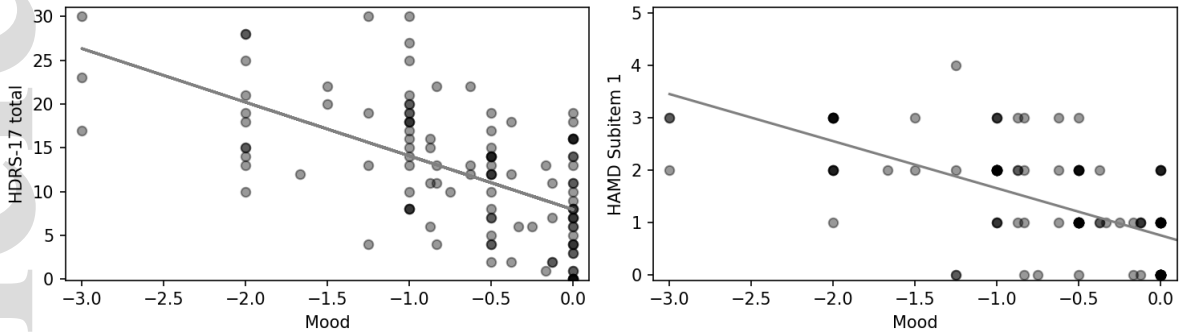
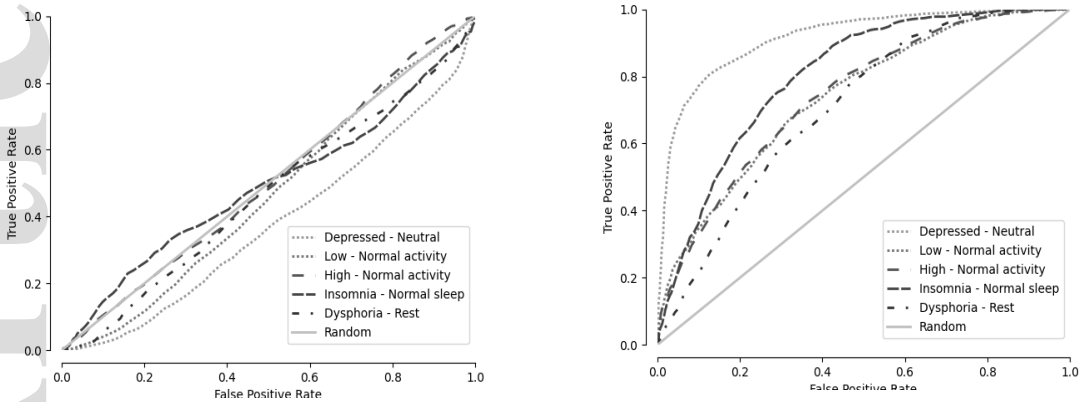


Figure 1 A & B. A generated null-distribution of AUC values from permutation test where the class label (e.g. patients with unipolar disorder and healthy control individuals) are randomly shuffled 200 times and an AUC for each permutation is plotted. The light grey bars represent the critical areas with the 5% largest values. The vertical line represents the observed AUC values from the true class label. **A:** Generated null-distribution for the Random Forest classification of patients with unipolar disorder against healthy control individuals. **B:** Generated null-distribution for the Random Forest classification of patients with unipolar disorder against patients with bipolar disorder.



A B

Figure 2. The association between **A:** Hamilton Depression Rating Scale 17-items score or **B:** Hamilton Depression Rating Scale score subitem 1 (mood) (B) and smartphone-based patient-reported mood in patients with unipolar disorder. The grey line indicates the linear least-square fit for each combination.



A B

Figure 3. The ROC curve for the classifications of different states based on voice features in patients with unipolar disorder. **A)** The user-independent models; **B)** The user-dependent models. Dysphoria defined as combined decreased mood and decreased activity.