Voice Quality as Digital Biomarker in Bipolar Disorder: A Systematic Review☆

*,†,‡Giovanni Briganti, and §,¶,‖,**Jérôme R. Lechien, *§Mons, †Liège, ‡Brussels, ¶Baudour, Belgium, and ‖**Paris, France

Summary: Background. Voice analysis has emerged as a potential biomarker for mood state detection and monitoring in bipolar disorder (BD). The systematic review aimed to summarize the evidence for voice analysis applications in BD, examining (1) the predictive validity of voice quality outcomes for mood state detection, and (2) the correlation between voice parameters and clinical symptom scales.

Methods. A PubMed, Scopus, and Cochrane Library search was carried out by two investigators for publications investigating voice quality in BD according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statements. Studies were assessed using the modified methodological index for non-randomized studies (MINORS).

Results. Of the 400 identified publications, 16 studies met the inclusion accounting for 575 BD patients. Machine learning approaches were implemented in 87.5% of studies, with classification accuracies ranging from 70.9% to 96.9%. Manic state detection showed the strongest predictive validity [area under the curve (AUC) up to 0.89], while depression detection demonstrated moderate performance (AUC: 0.66-0.78). Individual-specific models outperformed population-level approaches (correlation coefficients: 0.78 versus 0.44). Voice quality showed significant correlations with standardized clinical scales, particularly Young Mania Rating Scale and Hamilton Depression Rating Scale (normalized root mean square errors: 1.985 and 3.945, respectively). Prosodic features were examined in 81.25% of studies, with pitch consistently elevated during manic episodes. MINORS varied from 10 to 14, with notable limitations in sample size calculations and blinding procedures.

Conclusions. Voice quality is a promising biomarker in BD, particularly for manic state detection and individualized monitoring. While controlled settings showed strong performance, naturalistic applications yielded more modest results. Future research should focus on standardizing protocols across different environments and conducting large-scale longitudinal studies with robust methodological controls.

Key Words: BD—Voice—Otolaryngology—Otorhinolaryngology—Laryngeal—Larynx—Acoustic—Biomarker—Mood monitoring—Machine learning.

INTRODUCTION

Bipolar disorder (BD) is a chronic psychiatric condition characterized by fluctuating mood states, including manic, hypomanic, and depressive episodes. Mood states often result in substantial impairment and elevated risk for comorbid conditions, including suicide, cardiovascular disease, and metabolic disorders.[1,2] While pharmacological treatments remain central in managing BD,[3,4] the ability to reliably monitor and detect mood state changes could profoundly impact clinical outcomes.[5,6] Traditional methods for assessing mood states rely on clinical evaluations, which, although effective, are limited by subjective bias, accessibility, and patient adherence.[7,8] As a result, there is a critical need for objective biomarkers that can provide continuous, accessible, and accurate information on patients' mental states.

In that way, voice quality analysis can be considered a promising noninvasive biomarker for monitoring mood states in BD.[9] Voice features, including prosody, pitch, frequency, and spectral characteristics, are hypothesized to reflect the neurophysiological changes associated with mood fluctuations in BD. Vocal biomarkers may offer a novel approach to objectively differentiate mood states, track changes over time, and even predict mood episodes. Recent advances in digital health technologies, particularly artificial intelligence-based mobile and wearable devices, can improve the capture of voice data in real-world settings, thus expanding the potential for continuous, ecologically valid monitoring.[10,11]

Previous studies have examined various voice features as potential indicators of mood states in BD, often finding associations between voice characteristics and specific mood episodes. However, these studies differ widely in methodology, voice features analyzed, and the technology employed.[12–14]

The objective of this systematic review was to summarize the current evidence on the use of voice quality as a biomarker for mood states in BD.

Accepted for publication January 2, 2025.
☆ This research received no external funding.

From the *Unit of Computational Medicine and Neuropsychiatry, Faculty of Medicine, Pharmacy and Biomedical Sciences, University of Mons (UMONS), Mons, Belgium; †Department of Clinical Sciences, Faculty of Medicine, University of Liège, Liège, Belgium; ‡Faculty of Medicine, Université Libre de Bruxelles, Brussels, Belgium; §Department of Surgery, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium; ¶Division of Laryngology and Bronchoesophagology, Department of Otolaryngology Head Neck Surgery, EpiCURA Hospital, Baudour, Belgium; ‖Department of Otolaryngology-Head and Neck Surgery, Foch Hospital, School of Medicine, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France; and the **Department of Otolaryngology, Elsan Hospital, Paris, France.

Address correspondence and reprint requests to Jérôme R. Lechien, Department of Surgery, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium. E-mail: Jerome.Lechien@umons.ac.be

Journal of Voice, Vol xx, No xx, pp. xxx–xxx
0892-1997

https://doi.org/10.1016/j.jvoice.2025.01.002

MATERIALS AND METHODS

Framework for the data extraction

Two independent investigators (a psychiatrist-biostatistician and a laryngologist) conducted the systematic review and data collection according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)[15] checklist for systematic reviews. The criteria related to the inclusion and exclusion of studies were based on the population, intervention, comparison, outcome, timing, and setting framework.[16] Data extraction was conducted systematically to capture relevant information from each included study, following a predefined set of criteria. Data extraction was conducted independently by two reviewers to ensure accuracy and consistency, with any discrepancies resolved through discussion.

Patient population

This systematic review includes prospective, retrospective, controlled, or uncontrolled studies investigating the use of voice quality as a biomarker in patients diagnosed with BD. Eligible studies predominantly focused on adult or adolescent populations diagnosed with bipolar-I or bipolar-II disorder according to standardized diagnostic criteria[17] (eg, Diagnostic and Statistical Manual of Mental Disorders, International Classification of Diseases (ICD) guidelines). For all studies, the sample populations had to be specified, including details on demographics, BD subtype, and any reported comorbid conditions. Studies involving mixed diagnostic groups without specific analyses for BD were excluded.

Voice quality outcomes and modeling approaches

This review considered studies analyzing voice features in BD patients, particularly those exploring the relationships between specific vocal characteristics and mood states (eg, euthymia, mania, and depression). Studies utilized either controlled speech tasks or naturalistic voice samples obtained in clinical or real-world settings. The voice quality outcomes included subjective evaluations (eg, voice handicap index), perceptual subjective assessment [eg, Grade of dysphonia, Roughness, Breathiness, Asthenia, Strain, consensus auditory perceptual evaluation of voice], videolaryngostroboscopy ratings, aerodynamic, and acoustic measurements. Prosody and fluency outcomes were considered as well.

Outcomes

The primary outcomes included the association and predictive validity of voice quality outcomes with/on mood states in BD. The investigators focused on metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC) for models or algorithms designed to classify or differentiate between mood states (eg, euthymic, manic, and depressive) in BD patients. Secondary outcomes consisted of correlations between specific voice quality and clinical symptom scales, such as the Hamilton Depression Rating Scale[18] (HDRS) and Young Mania Rating Scale[19] (YMRS), with additional evaluation of reliability and cross-validation within studies to assess the robustness and generalizability of findings. Comparative analyses and modeling techniques were also reviewed, detailing any comparative analyses conducted between mood states, such as manic versus depressive or euthymic versus hypomanic, or with control groups. In addition, the authors needed to specify the modeling techniques used (eg, machine learning or statistical models), and any validation approaches (eg, cross-validation and test-retest reliability) to assess voice features.

Timing and settings

Included studies varied in their settings, spanning from controlled laboratory environments to naturalistic, real-world data collection through smartphone applications, telephonic interviews, or other mobile health platforms. The studies considering repeated measurements over days, weeks, or months to monitor mood changes longitudinally were considered. Single-session assessments without concurrent clinical mood evaluation or without specification of timing and settings were excluded.

Search strategy

The search was conducted through PubMed, Scopus, and Cochrane databases to identify studies evaluating the association between BD and voice quality through digital health devices. Studies published from 2000 to October 2024 were considered. A combination of search terms, including "BD," "voice," "speech," "acoustic features," and "mood states," was used to ensure broad coverage. The data extraction categories were designed to ensure consistency across studies and to support a comprehensive synthesis of the findings. Each selected study was reviewed in full, and the following criteria were extracted: design; country; setting (eg, clinical, real-world, or laboratory); demographics; BD subtype such as bipolar I or II; diagnosis criteria; voice outcomes; and results.

Bias analysis

The methodological quality of each included study was assessed using the validated methodological index for non-randomized studies (MINORS) tool.[20] The MINORS tool consists of 12 items that evaluate the methodological quality of non-randomized studies, with an ideal score of 16 for noncomparative studies and 24 for comparative studies. Each item was scored as 0 (absent), 1 (partially reported or inadequate), or 2 (fully reported and adequate). Key items included the clarity of the study aim, consecutive inclusion of patients, prospective data collection, and quality of endpoints. For studies evaluating changes in mood state over time, a minimum follow-up period of three months was considered adequate, while a loss-to-follow-up rate under 5% was considered acceptable. Sample size calculations, if present, were noted, indicating study rigor.

RESULTS

Of the 400 identified papers, 16 met the inclusion criteria[9,12–14,21–32] (Figure 1). There were nine prospective studies,[12–14,24,26–30] and seven cross-sectional studies,[9,21–23,25,31,32] respectively (Table 1) Six studies included a control group.[22,25,27,28,31,32] Demographics, study design, and outcomes are reported in Table 1. Demographics are available in Table 2. The primary population consisted of patients with BD (n = 575) who were compared with the following populations: major depressive disorder (n = 171), schizophrenia spectrum disorders (n = 112), anxiety disorder (n = 24), post traumatic stress disorder (n = 23), unipolar depression/disorder (n = 48), mixed disorders (n = 30), and healthy individuals (n = 100). Regarding BD subtypes and episodes, specific characterizations were provided in several studies: one focused on BD-I,[24] another included both BD-I and BD-II in various episodes,[21] one specifically studied mixed versus nonmixed episodes,[23] and another focused exclusively on manic episodes.[31] One study specified using ICD-10 criteria for BD classification.[12] The mean age was 48.1 years. The youngest population was in a child/adolescent study with mean age 12.7 years (SD = 3.2),[14] while the oldest population had a mean age of 51.1 years (SD = 12.5).[30] One study reported only an age range (23-69 years)[28] and two studies did not report age data.[9,29] According to the available data, there were 683 females and 538 males, respectively (Table 2). Most evaluations of voice quality were carried out in the hospital office. Most authors combined several voice quality outcomes. At least one acoustic measurement was used in all studies, including F0, spectral signal and formants, percent jitter, percent shimmer, spectral analyses, and harmonic-to-noise ratio (Table 2). A notable trend in the methodological approaches was the increasing implementation of technology-based data collection methods in more recent studies, with six studies utilizing smartphone or mobile sensing technologies for voice data collection.[12,24,26–29]

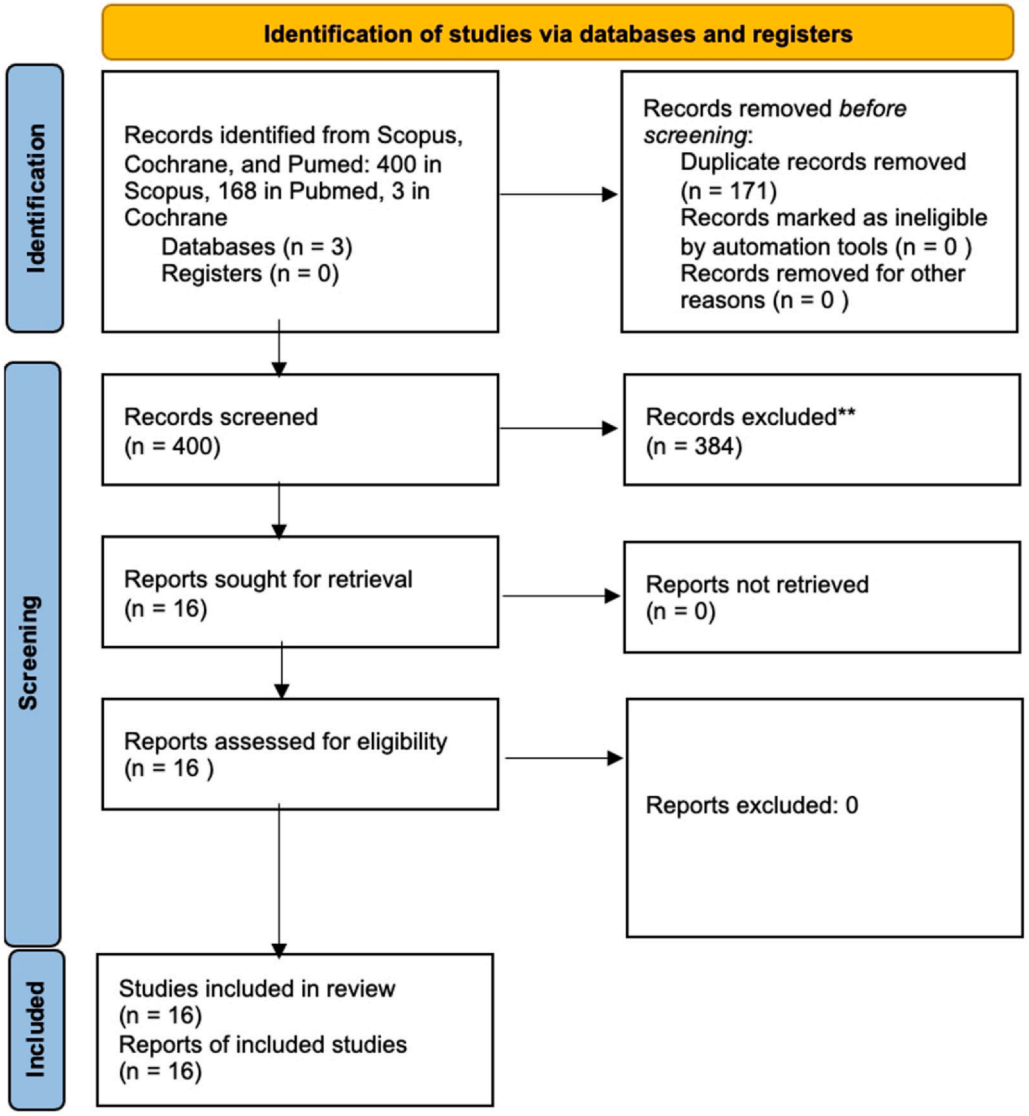

FIGURE 1. PRISMA flowchart. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

TABLE 1.
Study Design and Outcomes

References	Design	Setting/Device	*N*	F/M	Age (m, SD)	Outcomes	Results
Kaczmarek-Majer et al[12]	Prospective Uncontrolled	Smartphone	51 BD	28/23	38.2 (12.5)	Prediction of BD mood for prosodic, spectral VQ, intensity, and F0 variability	70.9%-71.4% for mood scales
Larsen et al[13]	Prospective Uncontrolled	Clin. Eval.	40 BD 34 MDD 30 others	76/28	36.5 (13.1)	Prediction of BD mood state of F0, jit, shim, and HNR	RR = 1.53 (P = 0.0138), Aggregated 2.00 (P = 0.0068)
Luo et al[14]	Prospective Uncontrolled	Clin. Eval.	70 BD 80 MDD	75/75	12.7 (3.2)	Determination of MDD/BD with spectral, and prosody (F0, formants, and jit)	AUC (MDD): 0.97 AUC (BD): 1.0
Cansel et al[22]	Cross-sectional Controlled	Clin. Eval.	15 BD 15 SSD 24 AD 25 MDD 25 HC	13/2 6/9 20/4 22/3 17/8	35.6 (11.0) 36.5 (13.1) 35.2 (11.3) 37.5 (11.0) 29.8 (8.1)	Determination of disorder with pitch stability, voice amplitude, pause duration, and articulation rate	Accuracy: 96.9% SVM: 96.5%
Weiner et al[23]	Cross-sectional Uncontrolled	Clin. Eval.	56 BD	41/15	41.1 (13.1)	Differentiation between mixed-nonmixed BD for spectral, MFCCs, and pitch modulation	Accuracy: 0.83
Faurholt-Jepsen et al[27]	Prospective Controlled	Smartphone	121 BD 48 UD 38 HC	73/48 29/19 17/21	35.7 (12.4) 45.6 (14.9) 31.7 (10.9)	Differentiation between UD, BD, and HC with pitch dynamics, voice energy, spectral variability, and articulation rate	AUC UD vs HC: 0.74 AUC BD vs UD: 0.58
Birnbaum et al[25]	Cross-sectional Controlled	Clin. Eval.	21 BD 41 SSD 27 HC	14/7 12/29 15/12	25.3 (4.2) 23.7 (4.0) 28.5 (5.2)	Differentiation between BD and SKI with jit, shim, intensity, and speech rate	AUC: 0.73
Faurholt et al[26]	Prospective Controlled	Smartphone	121 BD 21 UR 38 HC	73/48 11/10 17/21	35.7 (12.4) 32.3 (10.6) 31.7 (10.9)	Prediction of mood states with pitch, formant transitions, and speech regularity	Sensitivity: 0.79 Specificity: 0.54
Farrús et al[28]	Prospective Uncontrolled	Home Eval.	13 BD	NS	23-69	Jit, shim, speech rate, and F0	Best HDRS RMSE: 3.945 MRS RMSE: 1.985
Pan et al[21]	Cross-sectional Uncontrolled	Clin. Eval.	65 BD I-II	35/30	38.5 (11.4)	Mood state differentiation with prosody (F0, intensity), jit, and shim	Accuracy: 78%
Place et al[29]	Prospective Uncontrolled	Smartphone	30 BD 20 MDD 23 PTSD	24/49	NS	Prediction mood with vocal effort, call frequency, and F0 changes	Depressed mood AUC = 0.74 Social AUC = 0.83
Arevian et al[30]	Prospective Uncontrolled	Clin. Eval.	20 BD 15 SSD 12 MDD	21/26	51.1 (12.5)	Prediction state between BD, SKI, and MDD with word choice, pause length, F0 variability, and harmonics	Individualized model r = 0.78, Population r = 0.44
Zhang et al[31]	Cross-sectional Controlled	Clin. Eval.	30 BD 30 HC	16/14 14/16	41.4 (11.4) 36.3 (13.7)	Differentiation between maniac/nonmaniac states with formants (F1-F6), linear prediction coefficient (LPC)	F1 and F2 higher in mania > nonmania
Faurholt-Jepsen et al[24]	Prospective Uncontrolled	Smartphone	28 BD	18/10	30.3 (9.3)	Prediction of mood changes with F0, intensity variation, and speech rate	AUC = 0.89 for mania
Guidi et al[32]	Cross-sectional Controlled	Clin. Eval.	9 BD 10 HC	3/6 NS	41.1 (9.7) 30 (5)	Prediction of mood changes with LTAS with F0 correction for mood state differentiation	LTAS significant in hypomania vs euthymia
Vanello et al[9]	Cross-sectional Uncontrolled	Clin. Eval.	6 BD	1/5	NS	Prediction of mood change with F0, jitter, and F0 STD	F0: hypomania > mania Intra-subject variations for jit

Table 1 *Abbreviations:* AD, anxiety disorder; BD, bipolar disorder; F31, ICD-10 code for bipolar affective disorder; GAD, generalized anxiety disorder; GTCC, gammatone cepstral coefficient; HC, healthy controls; HNR, harmonic-to-noise ratio; ICD-10, International Classification of Diseases, 10th revision; Jit, percent jitter; LTAS, long-term average spectrum; MDD, major depressive disorder; MFCC, Mel-frequency cepstral coefficients; NS, not specified; PTSD, post-traumatic stress disorder; SD, standard deviation; Shim, percent shimmer; SMI, severe mental illness; SSD, schizophrenia spectrum disorders; UD/UR, unipolar depression/disorder; VQ, voice quality; WPT, wavelet packet transform.

TABLE 2.
Summary of Demographics, Populations, and Outcomes

Outcomes	*N*
Gender	
Females	683
Males	538
Mean age (years)	48.1
Populations	
Bipolar disorders	575
Major depressive disorder	171
Schizophrenia spectrum disorders	112
Anxiety disorder	24
Post-traumatic stress disorder	23
Unipolar depression/disorder	48
Mixed/unspecified disorders	30
Healthy individuals	100
Setting/devices	
Clinical evaluations (hospital)	10
Smartphone	5
Home evaluation (without smartphone)	1
Outcomes	
F0 (standard deviation, variability, and mean)	11
Spectral signal and formants	9
Percent jitter	6
Voice intensity	5
Speech rate	5
Percent shimmer	4
Pause duration (connected speech)	2
Harmonic-to-noise ratio	1
Word choice	1
Mel-frequency cepstral coefficients	1

Table 2 *Abbreviation: N,* number.

Voice quality and modeling approaches

Prosodic features were examined in 81.25% of studies ($n = 13$).[12–14,21–28,31,32] These studies consistently reported elevated pitch during manic states, with significant differences in fundamental frequency (F0) compared with euthymic states. F0 was a key voice feature across multiple studies: one study reported that both male and female participants exhibited higher F0 during manic/hypomanic states and lower F0 during depressive states compared with euthymia, though specific gender-differentiated values were not provided.[24] This pattern was partially supported by another study[9] who found higher F0 during hypomania (131 ± 12 Hz vs 119 ± 12 Hz in euthymia) in their mixed-gender sample. However, F0 changes during depression showed more variability, with some subjects showing increased and others decreased F0 compared with euthymia, possibly due to comorbid anxiety. Despite F0 being a commonly studied feature, most studies did not differentiate results by gender despite known baseline F0 differences between males and females. Additionally, no studies specifically examined F0 patterns in children or adolescents, limiting our understanding of age-related differences in voice features across mood states. F0 findings specific to the two aforementioned studies that investigated the F0 changes in males and females are thoroughly reported in Table 3: no study investigated children. Temporal features, including pause duration and speech rate, were analyzed in 43.75% of studies ($n = 7$),[9,22,26,27,29,30,32] with slower speech patterns typically associated with depressive episodes. Spectral characteristics were examined in 75% of studies ($n = 12$), with a particular focus on spectral tilt, harmonicity, and Mel-frequency cepstral coefficients.[9,12–14,21,23,25–28,31,32]

Regarding software tools, OpenSMILE® was utilized in 25% of studies ($n = 4$),[12,24,26,27] while Praat® was employed in 18.75% of studies ($n = 3$).[9,28,30] Custom algorithms were developed in 31.25% of studies ($n = 5$).[21,23,25,31,32] Machine learning approaches were implemented in 87.5% of studies ($n = 14$), with support vector machines (SVM) being the most common classifier. Classification accuracies using SVM ranged from 96.48%[22] to 70.9%,[12] while k-nearest neighbors achieved 96.943% accuracy in one study.[22] Random Forest classifiers were employed in three studies,[24,26,27] demonstrating consistent performance in longitudinal monitoring applications. Cross-validation was reported in 56.25% of studies ($n = 9$), with fivefold cross-validation being the most common approach.[12,21–24,26–29]

Pitch analysis was reported in 75% of studies ($n = 12$), revealing consistent patterns across different research settings. Studies demonstrated elevated pitch during manic episodes,[9,24,31] significant correlations between pitch variability and mood severity,[12] and higher fundamental frequency in hypomanic states.[32] Voice quality measures, specifically jitter and shimmer, were examined in 50% of studies ($n = 8$),[9,13,14,25,28–30,32] with significant associations between these parameters and mood states reported in six studies, particularly in differentiating between manic and depressive states.

Predictive validity of voice quality biomarkers

The predictive validity of voice biomarkers for mood state detection and diagnosis in BD was a primary outcome among the reviewed studies. Diagnostic accuracy varied substantially depending on the specific classification task

TABLE 3.
Fundamental Frequency (F0) Changes Across Mood States in Bipolar Disorder by Gender and Study

State	Gender	F0 Finding	Study
Mania/hypomania	Male	Higher F0 compared with euthymia	Faurholt-Jepsen 2016
Mania/hypomania	Female	Higher F0 compared with euthymia	Faurholt-Jepsen 2016
Depression	Male	Lower F0 compared with euthymia	Faurholt-Jepsen 2016
Depression	Female	Lower F0 compared with euthymia	Faurholt-Jepsen 2016
Hypomania	Mixed gender	Higher F0 (131 ± 12 Hz) vs euthymia (119 ± 12 Hz)	Vanello 2012
Depression	Mixed gender	Variable results—both increases and decreases observed	Vanello 2012

TABLE 4.
MINORS Analysis

Reference	Clearly Stated Aim	Consecutive Patients	Prospective Data Collection	Endpoints Appropriate to Study	Unbiased Endpoint Assessment	Follow-Up Adequate Period	<5% Lost to Follow-Up	Study Size Population Calculation	Adequate Control Group	Contemporary Groups	Baseline Group Equivalence	Adequate Stat AnalysesS	Total MINORS core
Kaczmarek-Majer et al[12]	2	1	2	2	1	1	1	0	2	2	2	2	13
Larsen et al[13]	2	2	2	2	2	2	1	0	1	2	2	1	14
Luo et al[14]	2	2	0	2	2	N/A	N/A	0	2	2	1	1	13
Cansel et al[22]	2	1	0	2	1	N/A	N/A	0	1	1	1	1	12
Weiner et al[23]	2	2	0	2	1	N/A	N/A	0	2	2	1	1	13
Faurholt-Jepsen et al[27]	2	2	2	2	1	2	1	0	2	2	2	2	14
Birnbaum et al[25]	2	2	2	2	1	2	1	0	1	1	2	1	12
Faurholt et al[26]	2	2	2	2	1	2	1	0	2	2	2	2	14
Farrús et al[28]	2	2	0	2	1	N/A	N/A	0	N/A	N/A	1	1	10
Pan et al[21]	2	2	2	2	2	N/A	N/A	0	2	2	2	2	14
Place et al[29]	2	2	2	2	1	2	1	0	1	1	2	2	12
Arevian et al[30]	2	2	2	2	1	1	1	0	1	1	2	1	13
Zhang et al[31]	2	2	0	2	2	N/A	N/A	0	2	2	1	2	13
Faurholt-Jepsen et al[24]	2	2	2	2	1	2	N/A	0	N/A	N/A	2	2	12
Guidi et al[32]	2	2	0	2	1	N/A	N/A	0	1	1	1	1	12
Vanello et al[9]	2	2	0	2	1	N/A	N/A	0	N/A	N/A	1	1	10

Table 4 *Abbreviations:* NA, not available.

and methodological approach. In distinguishing BD from healthy controls, classification performance showed promising results, with AUC values ranging from 0.74 to 0.80.[26,27] The differentiation between BD and other psychiatric conditions demonstrated mixed results, with a strong performance in distinguishing BD from schizophrenia (AUC 0.73)[25] and notably high accuracy in young populations for differentiating BD from major depressive disorder (accuracy 95.6%).[14]

Mood state prediction within BD showed varying levels of accuracy across studies (Table 1). The highest predictive performance was observed for manic states, with one study reporting an AUC of 0.89,[24] while another achieved 78% accuracy in mood state differentiation.[21] Mixed state detection demonstrated strong predictive validity, with an accuracy of 83%.[23] Depression detection showed moderate performance, with AUC values typically ranging from 0.66 to 0.78 across studies.[24,26]

Longitudinal monitoring applications revealed important insights into the temporal stability of voice biomarkers. Individual-specific models consistently outperformed population-level approaches, as demonstrated by correlation coefficients of 0.78 versus 0.44, respectively.[30] Home monitoring systems showed promising predictive capability with RMSE values of 3.94 for depression scales and 1.99 for mania rating scales.[28] Recent smartphone-based studies reported predictive accuracies ranging from 70.9% to 71.4% for mood scales,[12] while another study demonstrated relative risk ratios of 1.53 ($P = 0.0138$) for single samples and 2.00 ($P = 0.0068$) for aggregated measures.[13] The highest overall classification accuracies were reported in controlled clinical settings, with two studies achieving accuracies exceeding 96% using different machine learning approaches.[22] However, naturalistic settings typically showed more modest predictive performance, with sensitivity values ranging from 0.27 to 0.79 and specificity from 0.54 to 0.84 across various classification tasks.[26] This pattern suggests that while voice biomarkers show strong potential in controlled environments, their predictive validity may be more variable in real-world applications. Gender-specific analyses revealed differential predictive performance, with one study reporting distinct patterns of voice feature correlations with mood severity by gender.[12] Age-stratified analyses were less common, though one study in youth populations demonstrated particularly high predictive accuracy,[14] suggesting potential age-related variations in the utility of voice biomarkers. Real-time symptom tracking capabilities showed moderate-to-strong predictive validity, with AUC values of 0.74 for depressed mood detection and 0.83 for social behavior prediction.[29] These findings suggest that voice biomarkers may be particularly valuable for continuous monitoring applications, especially when combined with other behavioral indicators.

Correlation with symptom scales and clinical validation

Clinical validation primarily involved correlations with established mood rating scales, particularly the HDRS, YMRS, and other standardized psychiatric assessments.

Longitudinal monitoring studies revealed significant correlations between voice quality and mood scale scores. One study reported strong correlations between voice quality, YMRS scores, and manic state detection (AUC = 0.89).[24] Another study demonstrated moderate correlations with both HDRS and YMRS, achieving root mean square errors of 3.945 and 1.985, respectively, in predicting scale scores through voice analysis.[28] Individual-specific models showed stronger clinical correlations compared with population-level approaches, with correlation coefficients of 0.78 versus 0.44, respectively.[30]

Prosodic features demonstrated consistent correlations with clinical ratings. Pitch (F0) parameters showed significant associations with YMRS scores, particularly during manic episodes.[9,31] One study found gender-specific correlations between prosodic features and both HDRS and YMRS scores, with predictive accuracies of 70.9%-71.4% for these clinical scales.[12] Spectral features, particularly in the analysis of voice quality measures, showed significant correlations with symptom severity, as demonstrated by relative risk ratios of 1.53 for single samples and 2.00 for aggregated measures.[13]

Clinical validation studies using multiple psychiatric scales showed varying degrees of concordance. The BASIS-24 and SF-12 scales were used alongside voice analysis in one study, showing moderate correlations with global provider assessments.[30] Another study employed the M3 Checklist for clinical validation, demonstrating significant associations between vocal biomarkers and mental health severity measures.[13] The correlation between voice features and Clinical Global Impression scores was examined in several studies, with one study reporting significant associations between formant frequencies and clinical severity ratings.[31]

The temporal stability of these correlations was assessed in longitudinal studies, with follow-up periods ranging from 4 weeks to 208 days. Two studies demonstrated consistent correlations between voice features and mood ratings across multiple time points, though the strength of associations varied.[26,27] Home monitoring applications showed promising correlations with clinical scales, suggesting potential utility in remote assessment applications.[28]

In mixed states, one study achieved 83% accuracy in distinguishing mixed from nonmixed states when correlating voice features with clinical assessments.[23] The validation of voice parameters against structured clinical interviews was reported in several studies, with one study demonstrating correlations between voice features and SCID-based diagnoses for depression modules.[29]

Timing and settings for voice quality biomarkers in BDs

The duration of data collection varied substantially across studies, with longitudinal studies ($n = 6$) implementing monitoring periods from 4 weeks[13] to 208 days.[12] The remaining cross-sectional studies ($n = 10$) typically involved single or multiple recording sessions within controlled environments. Clinical settings were used in five studies, with recordings conducted during structured psychiatric interviews or standardized assessment sessions.[9,21,25,31,32] Controlled environments facilitated high-quality recordings but potentially limited ecological validity. Naturalistic settings, particularly through smartphone-based data collection, were employed in six studies.[12,13,24,26,27,29] These studies captured voice samples during routine phone calls or through dedicated voice diary entries, offering greater ecological validity but introducing challenges in controlling recording quality and environmental noise. The MONARCA® system[24] and BDmon app®[12] represented examples of systematic approaches to naturalistic voice data collection. Three studies implemented home monitoring systems,[28–30] combining elements of controlled recording conditions with natural environmental settings. Such studies employed structured voice tasks or standardized prompts while allowing participants to complete recordings in their home environment, achieving a balance between recording quality and ecological validity.

Recording frequency varied significantly across studies. Continuous monitoring approaches in smartphone-based studies typically collected multiple samples daily or weekly,[12,24,26,27] while clinical setting studies often involved less frequent but more structured recording sessions. The IVR system employed in one study implemented regular scheduled recordings, demonstrating the feasibility of systematic voice sampling in community-based settings.[30]

Time-of-day effects were specifically addressed in four studies,[12,13,27,29] with some evidence suggesting daytime variations in voice parameters, and time-stamped recordings to account for potential circadian influences on voice characteristics. The timing of recordings relative to mood episodes varied across studies, with some capturing voice samples during acute mood episodes[23,31] and others focusing on longitudinal mood fluctuations. The challenge of capturing voice samples during rapid mood transitions or mixed states was specifically addressed by one study that achieved 83% accuracy in distinguishing mixed from nonmixed states through targeted recording timing.[23]

Bias analysis

The methodological quality of the included studies was assessed using the MINORS criteria. Total MINORS scores ranged from 10 to 14 out of a possible 16 points, with higher scores indicating better methodological quality. Four studies achieved the highest score of 14,[13,26,27] while two studies received the lowest score of 10.[9,28] All studies clearly stated their aims, receiving maximum scores (2 points) for this criterion. Similarly, endpoint appropriateness was consistently well-addressed across all studies. However, significant methodological variations were observed in other domains. Prospective data collection was adequately performed in 56.25% of studies ($n = 9$), while the remaining studies relied on retrospective or cross-sectional data collection approaches. The blinded evaluation was a notable area of potential bias, with only 31.25% of studies ($n = 5$) achieving maximum scores for this

criterion,[13,14,21,31] while the remaining studies demonstrated partial or incomplete blinding procedures, potentially introducing assessment bias. Follow-up adequacy varied considerably among the longitudinal studies. Of the studies where follow-up was applicable, only five achieved maximum scores for follow-up period adequacy.[13,25–27,29] Loss to follow-up reporting was consistently suboptimal across studies, with most receiving only partial scores (1 point) for this criterion. None of the studies performed prospective sample size calculations, representing a universal methodological limitation. Control group similarity and baseline equivalence were well-addressed in studies, including comparison groups, with six studies achieving maximum scores for these criteria.[12,14,21,26,27,31]

Statistical analysis quality varied, with 50% of studies ($n = 8$) receiving maximum scores for appropriate statistical methodology. Bias in outcome measurements was adequately controlled in 37.5% of studies ($n = 6$), while others showed potential limitations in this domain.

DISCUSSION

This systematic review examined the utility of voice analysis in BD across 16 studies published between 2012 and 2024, revealing both promising applications and methodological challenges. The findings demonstrate potential value in voice quality as a biomarker for mood state detection and monitoring, highlighting important considerations for their clinical implementation.

The predictive validity of voice biomarkers showed considerable variation across different applications. While controlled settings achieved high classification accuracies exceeding 96%,[22] naturalistic environments typically yielded more modest results. This disparity underscores the challenge of translating laboratory findings into real-world applications. Notably, manic state detection consistently showed stronger predictive performance (AUC up to 0.89) compared with depression detection,[24] which suggests the potential state-specific utility of voice quality analysis. The correlation between voice features and clinical scales demonstrated moderate-to-strong associations, particularly in longitudinal monitoring applications. Individual-specific models consistently outperformed population-level approaches,[30] which highlights the importance of personalized baseline measurements in voice analysis applications. This finding aligns with the known heterogeneity of BD presentations and suggests that voice biomarkers may be most effective when calibrated to individual patients.

Methodological approaches varied substantially across studies, with recent trends favoring smartphone-based data collection. While this approach enhances ecological validity and facilitates continuous monitoring, it introduces challenges in standardization and quality control. The predominant use of machine learning classifiers (87.5% of studies) reflects the complexity of voice feature analysis but also raises questions about model generalizability and clinical interpretability.

The methodological quality assessment revealed important limitations in the current evidence base. The absence of prospective sample size calculations across all studies and inconsistent blinding procedures in many studies suggest areas for methodological improvement in future research. The variation in follow-up periods and high attrition rates in some longitudinal studies also highlight challenges in maintaining long-term voice monitoring protocols.

Several perspectives emerge from this review. First, the trade-off between controlled recording conditions and ecological validity remains a significant consideration in study design. Second, the optimal frequency and timing of voice sampling for reliable mood state detection requires further investigation. Third, the integration of voice analysis into clinical practice faces practical challenges regarding standardization and interpretation of results.

The primary limitation of this review is the heterogeneity across studies in inclusion criteria, populations, psychiatric and voice quality outcomes, and settings/devices. The lack of longitudinal follow-up of some studies is an additional limitation because some voice quality outcomes (eg, F0, speech rate, and intensity) should change over time and regarding the evolution of the psychiatric profile of patients. Despite methodological issues, the primary strength of this study is the summary of the potential of voice quality as a biomarker of mood state in PD. The summary of the current literature is an important step in conducting future high-quality studies considering the following key points: (1) standardizing multidimensional voice quality extraction and analysis protocols across different recording environments; (2) investigating the minimum necessary frequency and duration of voice sampling for reliable mood state detection; (3) developing robust methods for integrating voice analysis with traditional clinical assessments; and (4) conducting larger-scale longitudinal studies with proper sample size calculations and rigorous methodological controls. While voice analysis may not yet be ready for standalone clinical application in BD, its potential as an objective, noninvasive monitoring tool warrants continued investigation. The growing implementation of smartphone-based collection methods and advanced analytics suggests a promising path forward, particularly in the context of personalized medicine approaches to BD management.

CONCLUSION

Voice quality analysis is a promising biomarker of mood changes in BD. The strongest evidence supports its use in controlled settings and with individual-specific calibration. However, the current literature reports significant methodological challenges between studies, particularly regarding standardization across naturalistic environments and long-term implementation.

Data Availability Statement

Not applicable.

Declaration of Competing Interest

The authors have no financial interest in the subject under discussion. All authors have read and approved the paper.

Acknowledgments

None.

Institutional Review Board Statement

Not required.

Informed Consent Statement

Not applicable.

Author Contributions

Giovanni Briganti: Design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Jerome R. Lechien: Design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work; final approval of the version to be published; agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Phillips ML, Kupfer DJ. Bipolar disorder diagnosis: challenges and future directions. *Lancet.* 2013;381:1663–1671.
2. Carvalho AF, Firth J, Vieta E. Bipolar disorder. *N Engl J Med.* 2020;383:58–66. https://doi.org/10.1056/NEJMra1906193.
3. Dias VV, Balanzá-Martinez V, Soeiro-de-Souza MG, et al. Pharmacological approaches in bipolar disorders and the impact on cognition: a critical overview. *Acta Psychiatr Scand.* 2012;126:315–331. https://doi.org/10.1111/j.1600-0447.2012.01910.x.
4. Hui Poon S, Sim K, J. Baldessarini R. Pharmacological approaches for treatment-resistant bipolar disorder. *Curr Neuropharmacol.* 2015;13:592–604. https://doi.org/10.2174/1570159×13666150630171954.
5. Valenza G, Nardelli M, Lanata A, et al. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. *IEEE J Biomed Health Inform.* 2013;18:1625–1635.
6. Grünerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J Biomed Health Inform.* 2014;19:140–148.
7. Baldassano CF. Assessment tools for screening and monitoring bipolar disorder. *Bipolar Disord.* 2005;7(s1):8–15. https://doi.org/10.1111/j.1399-5618.2005.00189.x.
8. Miller CJ, Johnson SL, Eisner L. Assessment tools for adult bipolar disorder. *Clin Psychol Sci Pract.* 2009;16:188.
9. Vanello N, Guidi A, Gentili C, et al. Speech analysis for mood state characterization in bipolar patients. *Annu Int.* 2012;2012:2104–2107. https://doi.org/10.1109/EMBC.2012.6346375.
10. Mouchabac S, Conejero I, Lakhlifi C, et al. Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases. *Dialogues Clin Neurosci.* 2021;23:52–61. https://doi.org/10.1080/19585969.2022.2042165.
11. Maatoug R, Oudin A, Adrien V, et al. Digital phenotype of mood disorders: a conceptual and critical review. *Front Psychiatry.* 2022;13:895860.
12. Kaczmarek-Majer K, Dominiak M, Antosik AZ, et al. Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: a prospective study. Acta Psychiatr Scand. n/a(n/a). doi:10.1111/acps.13735.
13. Larsen E, Murton O, Song X, et al. Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study. *Front Psychiatry.* 2024;15:1342835 . https://doi.org/10.3389/fpsyt.2024.1342835.
14. Luo J, Wu Y, Liu M, et al. Differentiation between depression and bipolar disorder in child and adolescents by voice features. *Child Adolesc Psychiatry Ment Health.* 2024;18:19. https://doi.org/10.1186/s13034-024-00708-0.
15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372. Available at: https://www.bmj.com/content/372/bmj.n71.short. Accessed November 10, 2024.
16. Samson D, Schoelles KM. Developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. Methods Guide Med Test Rev Internet. Published online 2012. Available at: https://www.ncbi.nlm.nih.gov/sites/books/NBK98235/. Accessed November 10, 2024.
17. First MB, Yousif LH, Clarke DE, et al. DSM-5-TR: overview of what's new and what's changed. *World Psychiatry.* 2022;21:218.
18. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23:56–62.
19. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry.* 1978;133:429–435.
20. Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ J Surg.* 2003;73:712–716. https://doi.org/10.1046/j.1445-2197.2003.02748.x.
21. Pan W, Deng F, Wang X, et al. Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls. *Front Psychiatry.* 2023;14:1079448. https://doi.org/10.3389/fpsyt.2023.1079448.
22. Cansel N, Faruk Alcin Ö, Furkan Yılmaz Ö, et al. A new artificial intelligence-based clinical decision support system for diagnosis of major psychiatric diseases based on voice analysis. *Psychiatr Danub.* 2023;35:489–499. https://doi.org/10.24869/psyd.2023.489.
23. Weiner L, Guidi A, Doignon-Camus N, et al. Vocal features obtained through automated methods in verbal fluency tasks can aid the identification of mixed episodes in bipolar disorder. *Transl Psychiatry.* 2021;11:415. https://doi.org/10.1038/s41398-021-01535-z.
24. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry.* 2016;6:e856. https://doi.org/10.1038/tp.2016.123.
25. Birnbaum ML, Abrami A, Heisig S, et al. Acoustic and facial features from clinical interviews for machine learning-based psychiatric diagnosis: algorithm development. *JMIR Ment Health.* 2022;9:e24699. https://doi.org/10.2196/24699.
26. Faurholt-Jepsen M, Rohani DA, Busk J, et al. Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states. *Int J Bipolar Disord.* 2021;9:38. https://doi.org/10.1186/s40345-021-00243-3.

27. Faurholt-Jepsen M, Rohani DA, Busk J, et al. Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls. *Acta Psychiatr Scand.* 2022;145:255–267. https://doi.org/10.1111/acps.13391.
28. Farrús M, Codina-Filbà J, Escudero J. Acoustic and prosodic information for home monitoring of bipolar disorder. *Health Inform J.* 2021;27:1460458220972755. https://doi.org/10.1177/1460458220972755.
29. Place S, Blanch-Hartigan D, Rubin C, et al. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *J Med Internet Res.* 2017;19:e75. https://doi.org/10.2196/jmir.6678.
30. Arevian AC, Bone D, Malandrakis N, et al. Clinical state tracking in serious mental illness through computational analysis of speech. *PloS One.* 2020;15:e0225695. https://doi.org/10.1371/journal.pone.0225695.
31. Zhang J, Pan Z, Gui C, et al. Analysis on speech signal features of manic patients. *J Psychiatr Res.* 2018;98:59–63. https://doi.org/10.1016/j.jpsychires.2017.12.012.
32. Guidi A, Schoentgen J, Bertschy G, et al. Voice quality in patients suffering from bipolar disease. *Annu Int.* 2015;2015:6106–6109. https://doi.org/10.1109/EMBC.2015.7319785.