Cognitive Apps

MIT and **Harvard University** research
supporting Cognitive Apps' technology

https://cogapps.com

Received: 5 September 2019 | Revised: 31 December 2019 | Accepted: 17 January 2020

DOI: 10.1002/lio2.354

Laryngoscope Investigative Otolaryngology

REVIEW

Automated assessment of psychiatric disorders using speech: A systematic review

Daniel M. Low MSc[1,2] | Kate H. Bentley PhD[3,4] | Satrajit S. Ghosh PhD[1,4,5]

[1]Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, Boston, Massachusetts

[2]Department of Brain and Cognitive Sciences, MIT, Cambridge, Massachusetts

[3]Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts

[4]McGovern Institute for Brain Research, MIT, Cambridge, Massachusetts

[5]Department of Otolaryngology, Head and Neck Surgery, Harvard Medical School, Boston, Massachusetts

Correspondence
Daniel M. Low and Satrajit S. Ghosh, 46-4033F, 43 Vassar Street, Cambridge, MA 02139.
Email: dlow@mit.edu (D. M. L.) and satra@mit.edu (S. S. G.)

Funding information
Gift to the McGovern Institute for Brain Research at MIT; MIT-Philips Research Award for Clinicians; National Institute of Health, Grant/Award Number: 5T32DC000038-28; S.S.G was partially supported by 5P41EB019936

Abstract

Objective: There are many barriers to accessing mental health assessments including cost and stigma. Even when individuals receive professional care, assessments are intermittent and may be limited partly due to the episodic nature of psychiatric symptoms. Therefore, machine-learning technology using speech samples obtained in the clinic or remotely could one day be a biomarker to improve diagnosis and treatment. To date, reviews have only focused on using acoustic features from speech to detect depression and schizophrenia. Here, we present the first systematic review of studies using speech for automated assessments across a broader range of psychiatric disorders.

Methods: We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines. We included studies from the last 10 years using speech to identify the presence or severity of disorders within the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). For each study, we describe sample size, clinical evaluation method, speech-eliciting tasks, machine learning methodology, performance, and other relevant findings.

Results: 1395 studies were screened of which 127 studies met the inclusion criteria. The majority of studies were on depression, schizophrenia, and bipolar disorder, and the remaining on post-traumatic stress disorder, anxiety disorders, and eating disorders. 63% of studies built machine learning predictive models, and the remaining 37% performed null-hypothesis testing only. We provide an online database with our search results and synthesize how acoustic features appear in each disorder.

Conclusion: Speech processing technology could aid mental health assessments, but there are many obstacles to overcome, especially the need for comprehensive transdiagnostic and longitudinal studies. Given the diverse types of data sets, feature extraction, computational methodologies, and evaluation criteria, we provide guidelines for both acquiring data and building machine learning models with a focus on testing hypotheses, open science, reproducibility, and generalizability.

Level of Evidence: 3a

KEYWORDS
machine learning, mental health, psychiatry, speech, voice

1 | INTRODUCTION

Mental health disorders in the United States affect 25% of adults, 18% of adolescents, and 13% of children.[1,2] These disorders have a larger economic impact than cancer, cardiovascular diseases, diabetes, and respiratory diseases, but societies and governments spend much less on mental disorders than these other disorders.[3] Current approaches to the assessment and monitoring of psychiatric conditions rely primarily on intermittent reports from affected individuals or their caregivers. These reports are often subjective and include patients' retrospective recall biases (eg, to downplay or overestimate symptoms), cognitive limitations (eg, memory of episodes and environment, causal inference), and social stigma. There is an urgency to objectively diagnose, monitor over time, and provide evidence-based interventions for individuals with mental illnesses, particularly those who are unable to access traditional psychological or psychiatric services due to geographical, financial, or practical barriers. Only 41% of US adults suffering from a mental health condition access mental health services in a given year.[4] This systematic and objective assessment would facilitate remote assessments and better personalization of care and thereby improve clinical services across the medical practice.

One promising avenue toward improving the objectivity of psychiatric assessment and access to services is to leverage the increase in health-related data collection using sensors (eg, wearables, smartphones, cameras) alongside improvements in machine learning technology (see Figure 1 for an overview). Wearables, including watches, rings, and clothes that measure biological and behavioral indices such as temperature, skin conductance, movement, and heart rate, can be potential indicators of anxiety and depression, and used to provide biofeedback.[5,6] Features obtained from video recordings have been used to detect depression[7] and bipolar disorder.[8-10] Technologies such as MultiSense[11] can be used to measure facial expressions, body gestures, smile-frown dynamics, and eye contact. Many smartphones can measure ambient light, moisture, pressure, gait, location, and acceleration, steps taken, some of which are used to detect psychiatric disorders.[12,13] Features extracted from handwriting have shown to indicate anxiety and stress.[14] Neuroimaging data have also been provided many promising results,[15-17] but this article focuses on non-neuroimaging sensors such as voice. Finally, text obtained either from transcribed audio recordings, blogs, or social media has been used to detect many psychiatric disorders including psychotic, depressive, and anxiety disorders from morphological, syntactic, semantic, and discursive features (for reviews, see References 18-20). These technologies thus provide opportunities for better assessment of mental health.

1.1 | The promise of technological assessment of mental health

Using machine learning technology to analyze data obtained from sensors for the assessment of psychiatric disorders has the potential to: (a) screen for at-risk individuals before they access the mental health care system; (b) complement clinicians' assessments once individuals seek care; and (c) help monitor symptoms once patients leave the clinic or in between consultations. Each of these goals will be discussed in turn.

First, these technologies can help address several barriers that prevent individuals from accessing mental health diagnoses and treatments in the first place. A factor analysis of primary care surveys[21] found that the main perceived barriers to potential psychological treatment are cost, stigma, lack of motivation, fear of unsettling feelings, negative view of therapy, mismatch between therapy and needs, time constraints, accessibility restrictions, and availability of services. Some subpopulations facing especially challenging barriers to care include individuals with physical handicaps[22] or those involved in wars or humanitarian reliefs where distress is more likely.[23] Individuals outside the mental health care system could still assess their mental health remotely with such technologies and then access online resources, telemedicine options, and smartphone apps (eg, cognitive behavioral therapy strategies according to their severity).[24] These technologies may be able to, using longitudinal data across individuals, select personalized treatment alternatives by learning the success rate of different treatments given specific symptomatology.[25] More immediately, such technologies could be used for screening in schools, universities, armed forces, and primary-care settings.

Second, these technologies may improve assessment within the clinic given certain obstacles clinicians face. Once individuals access the mental health care system, qualitative clinical evaluations face the obstacle of diagnosing disorders that may be episodic and may have high comorbidity rates. This makes it harder to separate overlapping symptoms into underlying discrete diagnoses. This obstacle is evidenced by the presence of low inter-rater reliability[26] and test-retest reliability[27] in certain psychiatric diagnoses including major depressive disorder (MDD) that have a low kappa score. It is a complex engineering problem to create a model to detect a specific disorder when many patients present more than one disorder or symptoms are intermittent. For instance, more than 50% of cases of post-traumatic stress disorder (PTSD) co-occur with depressive, anxiety, or substance use disorders.[23] Furthermore, suicidal thoughts and behaviors can also be a goal in predictive models[28] (for a review, see Reference 29) and are present across many disorders. Comorbidity is one reason why the National Institute of Mental Health has developed the Research Domain

Supervised machine learning gradually learns a function that maps an input (e.g., speech features) to a known output (e.g., depression score from PHQ-9). It is a classification problem when the outputs are categorical (e.g., +PTSD or -PTSD) and a regression problem when outputs are continuous values (e.g., PTSD Checklist score). Machine learning is particularly useful in finding structure in big, complex, and multidimensional data, where humans cannot visualize the structure on their own. Certain algorithms learn by repeatedly seeing examples of inputs and outputs; the algorithm first predicts the output of an input, then sees the label, and over iterations corrects its configuration if mistaken in such a way that it would have guessed correctly. These labels can be obtained from clinical diagnosis or self-assessments. *Feeding a bad label will create a bad model.* This creates an open problem to be evaluated when gathering data for a given disorder: clinical diagnoses tend to be considered the gold-standard but certain disorders may have a low inter-rater reliability [26] and be more episodic and thus self-assessments may be preferred. Finally, model performance is measured by learning on a training set of samples (sets of inputs and outputs) and testing on new samples not used for learning.

Unsupervised machine learning is used when the labels (e.g., depression, control) for each data point (e.g., speech features of a single participant) are unknown. Therefore, unsupervised algorithms find structure in the data by, for example, clustering similar data points together. This can be useful for observing which features relate samples to each other. Moreover, such unsupervised approaches (e.g., PCA, tSNE, UMAP) may help uncover participants with similar symptoms and on that basis may reveal disorder subtypes.

Testing performance. Data can be split into a training and test set (e.g., splitting the data into a random 80-20% split), and the training set can also be split into a training and development set. This way, different models and configurations could be tried on the development to choose the one that increases performance to later evaluate on the test set not used in training. However, an issue in small datasets, which are common in the medical field, is that these random subsets (both the development or the test sets) may not be representative of the general population. Therefore the performance on these small subsets cannot be expected to generalize to the general population. A better estimate than using a single, small subset is taking multiple subsamples through k-fold cross validation, which is splitting the training data into k segments (e.g., k=10) and iteratively training the model on k-1 segments and validating on the left-out segment and then averaging performance. When there are very few participants (e.g., n < 40), leave-one-out cross validation is often used, which maximizes the amount of data seen during training at the cost of increasing the variance of the predictive model (see Figure 5 for an alternative).

Overfitting. When studies only report performance on small development sets, it is likely their models will not perform as well on unseen data because they will likely *overfit*. Overfitting consists in learning model configurations that increase performance on the development set or folds without being able to generalize performance to an unseen test set, which is the main purpose of building a predictive model. In the figure below, the data distribution is generated by a line with some noise. The polynomial model (orange) can learn to fit the input samples perfectly; however, it is more likely that the linear model (black) will make better predictions on future unseen samples. In other words, overfitting is finding an illusory pattern in the training set that does not exist in the general population. This is why larger, representative samples are needed with held-out test sets. It is also why testing should be done only once; otherwise trying different model configurations will overfit to the test set as well. For further reading, see [34,127].

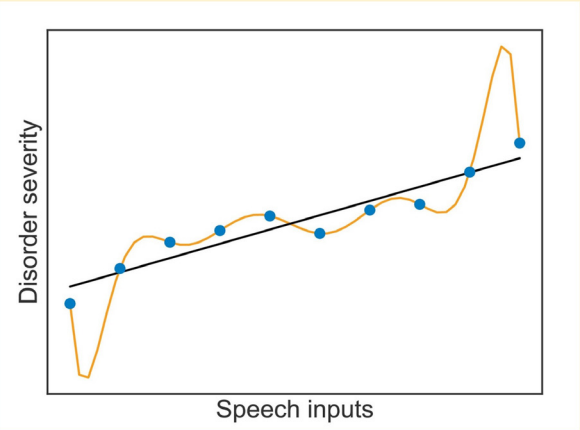

FIGURE 1 How machine learning works

Criteria with the goal of deconstructing diagnoses with biomarkers—from genetic to behavioral—to predict and improve response to treatments.[30] Therefore, algorithms trained on behavioral descriptors could provide likelihood estimates for different disorders to aid clinicians in differential diagnosis (eg, determining whether a patient meets criteria for unipolar depression or bipolar disorder[31]), help detect risk for chronic psychiatric disorders,[32] psychiatric episodes,[33] or suicidal behavior[29]; and over time learn to predict the best treatment given multimodal

(genetic, brain-imaging, behavioral) data.[34] Therefore, complementing clinical interviews with machine learning models trained on the recordings of these interviews could improve outcomes, save clinicians' time, reduce health care costs, and make treatment planning more efficient.

Finally, this technology may improve mental health care by facilitating more regular and real-time monitoring of symptoms. For instance, even if individuals are able to see a clinician in-person, they may not return; therefore, remote monitoring would allow individuals, caregivers, or clinicians to observe and assess mental health and decide if it is time to seek help. Furthermore, once chronic patients are in a regular visiting schedule, symptoms may fluctuate in between visits. Sensors and just-in-time adaptive interventions might ultimately be able to detect urgent episodes or warning signs and deploy online resources or computerized therapy before problems escalate.[35] With monitoring via these real-time methods, patients and clinicians have the potential to more reliably observe behavior, perform early detection of episodes, request unscheduled evaluations, and/or change the course of treatment in a personalized way.

These promises are far from being fulfilled. Most studies of such applications to date do not use large, representative samples that are needed to assess disorders in out-of-sample individuals. Clinical data sets tend to be small, and models trained on limited observations of a certain type of data (eg, recorded in a silent room, Caucasian speakers, adults) may not even extrapolate to data that seems to be similar. Furthermore, algorithms are susceptible to learning biases inherent in the data used to train them (eg, incorrectly assigning lower disorder severity to African Americans because less of them have the disorder in the training set).[36-39] Critically, many high-performing algorithms (eg, deep neural networks, proprietary models) are "black boxes," since it is currently not understood how these models combine features to output the severity of a disorder. This creates a lack of trust since they have been shown to be fooled by adversarial attacks (ie, perceptually small manipulations in the inputs that create incorrect outputs).[40] This is why a recent European Union regulation requires a right to obtain an explanation of life-affecting decisions from automated algorithms[41,42] such as clinical assessments, and DARPA has released an Explainable Artificial Intelligence program to tackle these challenges[43] ("Explain and interpret models to reduce bias and improve scientific understanding" guideline in the Section 4).

1.2 | Speech as an automated biomarker for mental health

Most of us speak effortlessly without realizing the complexity of coordination that this act entails. Speaking is not just moving the mouth. It is an orchestration of human communication expressing thought, intent, and emotion in a carefully choreographed performance. This motor coordination involves over 100 muscles and is supported by a large network of brain regions processing auditory, somatosensory, and visual input, language perception and production.[44] Thus, spoken communication is a window into the mind, and opens the strong potential for the plethora of technologies to capture and process speech to evaluate mental health.

Speech patterns have been known to provide indicators of mental disorders: in 1921, Emil Kraepelin stated that depressed patients' voices tended to have lower pitch, more monotonous speech, lower sound intensity, and lower speech rate as well as more hesitations, stuttering, and whispering.[45] In comparison to other behavioral descriptors (eg, skin conductance, acceleration), speech has a number of advantages: it is hard to hide symptoms, it directly expresses emotion and thought through its language content, it indirectly reflects neural modulation through motor and acoustic variation, it may generalize across languages (due to similar vocal anatomy) which is especially useful for low-resource languages when natural language processing technology is not available, and it is relatively effortless to obtain using smartphones, tablets, and computers instead of more costly wearables or invasive neuroimaging methods, especially considering many clinical interviews are already recorded. Furthermore, it is a type of data that will be increasingly available given the improvements in speech recognition and shown through virtual assistants such as Amazon Alexa, Apple Siri, Google Voice Search, speech to text applications for electronic health records, and voice biometrics for security, military, and education.

Table 1 provides an overview of the different approaches to assess mental health and their relative advantages and disadvantages (see also Reference 48). In this review, we focus on studies that compare whether acoustic features differ in psychiatric populations through null-hypothesis testing and predictive models which use acoustic features to detect the presence or severity of a psychiatric disorder in an individual. Both types of models are built using automatically extracted acoustic features. Null-hypothesis models isolate variables deemed important above a relatively arbitrary *P* value, and can be incongruent with the variables that maximize predictions in new settings.[49] Significant differences are usually considered more useful for scientific inference than prediction. Predictive studies train models on a subset of the data and test performance on the rest of the data not used for training and therefore give insight into how well they may generalize to new individuals.

The goals of this review are to both provide a state of the art on computationally detecting mental health disorders from acoustic speech features and synthesize best practices to achieve this goal. There are reviews on using speech to detect specific psychiatric disorders such as depression[29,50,51] and schizophrenia[52,53]; however, this is the first systematic review on a broad range of psychiatric disorders. The reasons for performing this review are to address the lack of a clear picture of the utility of speech signals in detecting and differentiating mental health disorders, as well as the relative efficacy of the signal across disorders; highlight the variability of acoustic features that may be useful in assessing psychiatric disorders; discuss confounders that affect such assessment and how are they controlled; provide a practical guide to a set of experimental tasks to elicit speech; and report which methodologies are being used to improve generalization of models to new individuals. We wish to show how having access to speech data could improve mental health care, which we will discuss provides a new bridge between psychiatry and laryngology.

TABLE 1 Advantages and disadvantages of different types of psychiatric assessments

Measurement	Advantages	Disadvantages
Clinician assessments using perceptual-rated questionnaires	• Clinician experience • Tests have populational norms • Clinicians can ask for further assessments • Clinicians can offer treatment pathway • Questionnaire items are interpretable	• Costs of clinic and clinician • Time-consuming • Requires extensive training • Normally assessed sporadically in clinic • Questionnaires often use ordinal and vague variables (eg, never, sometimes) • Prone to clinician's biases: expertise, culture and race,[46] research question • Patient's memory distortions[47] • Patients' perceived barriers to pursuing treatment (see main text) • Inter-rater reliability can be low • Cannot capture complex features
Self-assessments	• Potentially free • Less time-consuming than clinician assessments • No clinical training required • Can be administered anywhere • Tests have populational norms	• More narrow than clinical evaluation • Biased by patient's voluntary responses • Generally cannot offer personalized treatment • Cannot capture complex features • Assessments have to be created and validated based on observation of symptoms
Automated computational assessments based on sensors	• Potentially free • Potentially instantaneous • Can be done remotely, continuously, and naturalistically (app prompts) • Can incorporate larger and more varied samples than clinic samples • Avoids human biases and single rater • Can capture multimodal features (audio, video, text, accelerometer) • Ratio and continuous variables • Can capture complex features due to linear and nonlinear multivariate models, and find new structure in data • Allows scalability because models can be fast and automated	• Most models have not been validated through clinical trials thus far • Needs large amounts of data • Many sources of variation in the signal, and their relative contributions are poorly understood • Models can be affected by biases in data (eg, race, age, noise) • Difficult to incorporate expert priors (eg, body language, clinical history) into models • Assessment does not automatically lead to treatment or intervention options

Therefore, our specific aims for this systematic review are to (a) synthesize results from publications covering null-hypothesis testing and predictive models which use automatically extracted acoustic features to detect psychiatric disorders, (b) characterize psychiatric disorders based on the acoustic features that are significantly different in comparison to neurotypical populations, (c) link these altered acoustic features to observed symptoms or behaviors, and (d) considering the challenges of this field, offer guidelines for acquiring data and building machine learning models to achieve higher reproducibility and generalizability. Thus, we hope to facilitate the application of these methods to improve assessment and treatment of psychiatric disorders.

2 | METHODS

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines.[54]

2.1 | Eligibility criteria and literature search

The search aimed at identifying articles from the last 10 years that include computational methods for predicting psychiatric disorders by analyzing speech from individuals' recordings through machine learning methods. The following studies were excluded: (a) studies with children or about developmental disorders; (b) case studies; (c) studies that only used perceptual evaluations of speech; (d) studies without control groups or comparison along a diagnostic scale's severity; (e) unpublished or non-peer reviewed theses; and (f) if disorder had with many eligible studies (>40), we excluded studies published before 2018 with under four citations plus one citation per year of antiquity (ie, included 2017 articles with four citations, 2016 articles with five citations). Google Scholar was used as it indexes journals as well as conference articles, a common type of publication in computational speech analysis. Articles ranging from 2009 to the present were searched between May 16 and August 12, 2019 by finding keywords in the title in the following manner: "allintitle:(<disorder> + acoustic

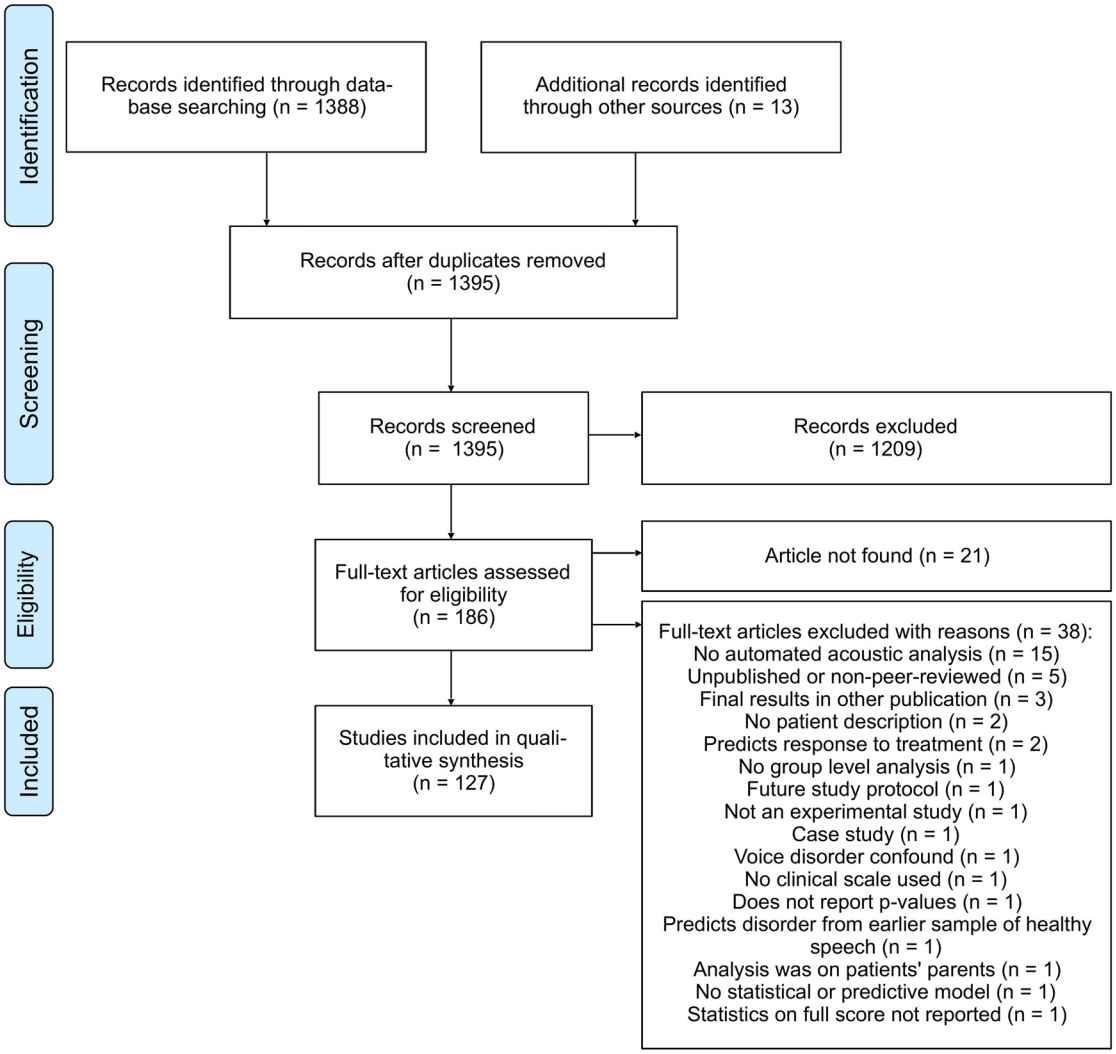

FIGURE 2 PRISMA flow diagram of study inclusion and exclusion criteria for the systematic review

TABLE 2 Summary of systematic review results

Disorder	Articles % (N)	Median sample size (range)	Clinical assessment % (N)	Predictive models % (N)
Depression	49.6 (63)	123 (11-1688)	38 (24)	87 (55)
PTSD	7.9 (10)	41 (10-253)	70 (7)	80 (8)
Schizophrenia	18.1 (23)	44 (18-195)	86 (20)	13 (3)
Anxiety	4.7 (6)	45 (20-104)	50 (3)	0 (0)
Bipolar	16.5 (21)	39 (5-89)	90 (19)	66 (14)
Bulimia	0.8 (1)	22 (-)	100 (1)	0 (0)
Anorexia	1.6 (2)	107 (66-148)	100 (2)	0 (0)
OCD	0.8 (1)	35 (-)	100 (1)	0 (0)

Note: The distribution of the 127 studies that matched the inclusion criteria is described in the Articles column. Within each disorder, the following characteristics are described: median sample size (case group plus control group), proportion of clinical diagnosis vs self-report measures, and proportion of predictive vs null-hypothesis testing studies.
Abbreviations: OCD, obsessive-compulsive disorder; PTSD, post-traumatic stress disorder.

OR acoustical OR speech OR voice OR vocal OR audio OR pitch OR prosody OR vowel)," where <disorder> was replaced by DSM-5[55] disorders (iteratively searched for each disorder and related terms due to character limitations in Google Scholar searches). We excluded neurological or neurocognitive disorders (eg, Neurocognitive Disorder Due to Parkinson's Disease), neurodevelopmental disorders (eg, autism),

noncognitive or body-centered disorders (eg, sleep, catatonic, somatic, sexual, elimination), and substance use disorders, which allows us to approximately reduce the scope of the review to adolescent and adult psychiatric disorders. This resulted in the following search terms with associated names: "post-traumatic stress" OR PTSD OR posttraumatic stress; bipolar OR mania OR manic OR cyclothymic, anxiety OR anxious OR mutism OR phobia OR panic OR agoraphobia; "obsessive-compulsive" OR obsessive-compulsive disorder (OCD) OR dysmorphic OR hoarding OR trichotillomania; dissociative OR depersonalization; "eating disorder" OR anorexia OR bulimia OR "binge-eating" OR pica OR rumination; "personality disorder" OR "paranoid personality" OR schizoid OR antisocial OR borderline OR histrionic OR narcissistic OR avoidant OR "dependent personality"; "mood disorder" OR "mood dysregulation"; schizophrenia OR schizophrenic OR schizotypy OR schizotypal OR psychosis OR psychotic OR delusion OR delusional OR paranoia OR paranoid OR alogia; depression OR depressed OR depressive OR dysthymia OR MDD. Reviews were searched twice by adding the term "+ review." General keywords were also included in the review search such as "mental health" OR psychiatry OR psychiatric OR "affective disorder" OR "psychological disorder" OR "mental illness."

2.2 | Data extraction

Screening was performed by the first author (D.M.L.) by reading the title and abstract. From each article, the following features were synthesized if available: disorders, sample size, presence of control group, age, clinically assessed or self-assessed, clinical scales used for diagnosis, tasks to obtain speech, predictive model, highest performance or statistical significance, type of validation or test set, and other relevant findings (especially if stated which features were predictive).

3 | RESULTS

A total of 127 studies were included in the review (see Figure 2). See Table 2 for a general description on the search results. Full synthesized search results are available online (https://tinyurl.com/y6ojfq56), which can be updated with new studies on a blank row by adding comments for every column, with the table at the time of publication also available (https://tinyurl.com/tu58te3). Review articles and data sets without models were included for easy reference, but were not counted in Table 2.

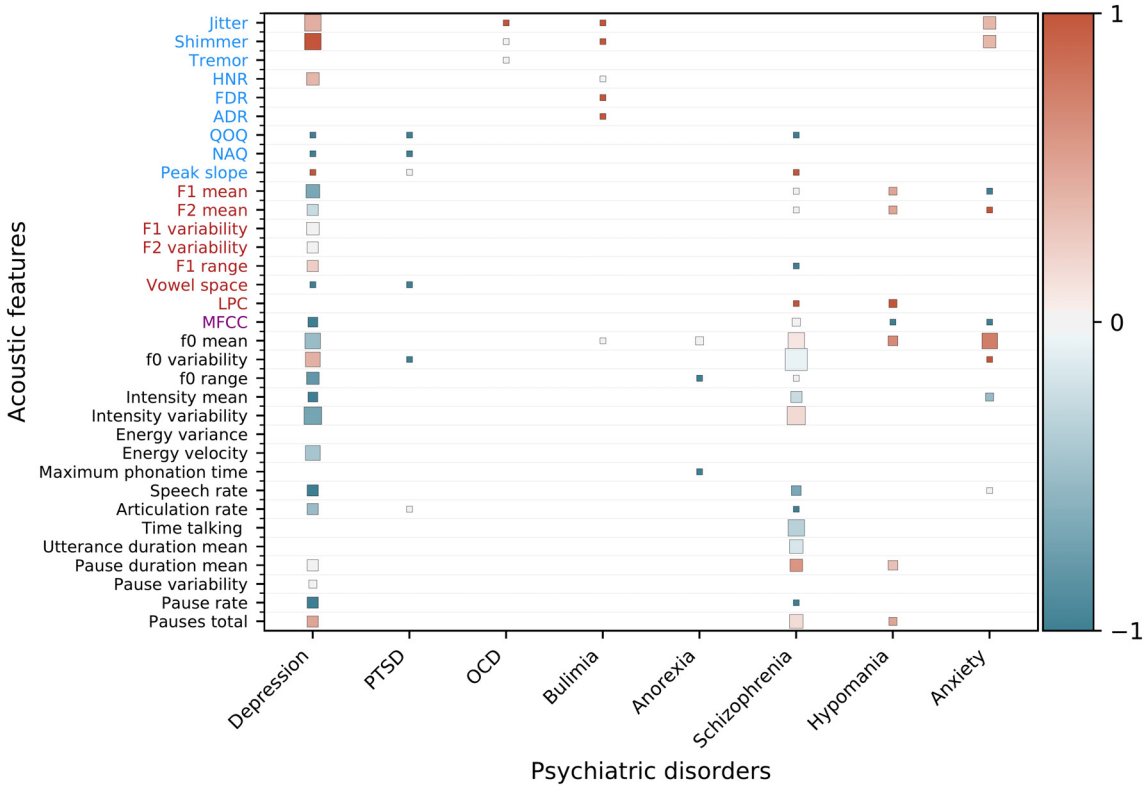

FIGURE 3 Synthesis of null-hypothesis testing studies across psychiatric disorders. Acoustic features are color-coded on the y-axis into source features from the vocal folds (blue), filter features from the vocal tract (red), spectral features (purple), and prosodic or melodic features (black).[56] Features that are significantly higher in a psychiatric population than healthy controls or that correlate positively with the severity of a disorder receive a score of 1 (red), features that are lower or correlate negatively receive a score of −1 (blue), and nonsignificant or contradicting findings receive a score of 0 (gray). The mean is computed for features with multiple results. The cell size is weighed by the amount of studies. Features not studied in a disorder are blank. Anxiety, social or general anxiety disorder; OCD, obsessive-compulsive disorder; PTSD, post-traumatic stress disorder

3.1 | Key acoustic features across disorders

Figure 3 provides a synthesis of the field, showing which features have been studied thus far through null-hypothesis testing. We synthesize the acoustic features that have been shown to be statistically lower or higher in a psychiatric disorder in comparison to healthy controls or those that correlate with a diagnostic scale (see Figure 4 for glossary). Each cell represents the sign of a statistical test (eg, psychiatric group significantly higher than control group) or correlation (eg, negative correlation with diagnostic scale) across potentially multiple studies for a given acoustic feature and disorder. Acoustic features that are significantly higher in psychiatric population than healthy controls or that correlate positively with the severity of a disorder receive a score of 1, ones that are lower or correlate negatively receive a score of −1, and nonsignificant findings receive a score of 0. Then the overall score for each cell is obtained by taking the mean. The cell size reflects the amount of studies supporting this effect. To improve visibility, features that appear in only one study were excluded for the top-studied disorders (ie, depression, schizophrenia, and hypomanic states within bipolar disorder). Results that correlate a feature with a subsymptom of a clinical scale were excluded (ie, only total scores from scales were used).[1]

Some aspects were simplified: for instance, both social anxiety disorder and generalized anxiety disorder were grouped under the anxiety column. Energy and intensity features were grouped under intensity. Within bipolar disorder studies, features that characterized depression with regards to euthymia were placed in depression disorder column, while both hypomania vs euthymia and bipolar vs control were placed in the hypomania column. Also worth noting is that the type of task used could change results, even within the same study,[58] such as extracting features from sustained vowels or voiced speech.[59]

4 | DISCUSSION

The majority of studies found in this review that use automated speech feature extraction to assess mental health conditions focused on MDD, bipolar disorder, and schizophrenia. The reason is likely due to the fact that Audio/Visual Emotion Challenge and Workshop (AVEC) machine learning competitions have been carried out for MDD-PTSD[60-63] and bipolar disorder,[64] where the goal is to detect disorder severity using audio and video features. The open-access research data sets used in these competitions such as the Distress Analysis Interview Corpus[65] were then used by many other studies after the competition. Out of the 127 studies included in this review, 32% used AVEC data sets. The creation of apps to collect data such as MONARCA[13,33] or PRIORI[66] apps for hypomanic and depressive state detection in bipolar disorder has also helped promote studies. Schizophrenia, on the other hand, has been studied less from the machine learning field since 87% of studies performed null-hypothesis testing only.

Regarding performance in predictive models, few studies used held-out test sets. It is unlikely that performance will generalize as reported if studies did not evaluate performance on a representative held-out test set (which are most studies in this review) and instead used some form of cross-validation (which is the case in most reviewed studies), which is likely overfitting (see Figures 1 and 5). This limited generalizability and overfitting are observed for instance in the drop in performance from development to test set in submissions to the AVEC challenges.[8,9,64,67] For results that used held-out test sets, which are more likely to generalize if they are representative, scores range from close to chance to higher scores including Afshan et al[68] (F1-score = 0.95) which most likely benefited from having a large sample size (N depressed = 735, N controls = 953) and all participants being the same sex (female). At the same time, Kächele et al[69] obtained one of the highest performances in AVEC 2014 (ie, mean absolute error = 7.08), simply using provided audio baseline features and a random forest classifier (the highest performance combined audio and visual features).[70] Therefore, performance is a function of sample size, preprocessing, feature selection, and model, which will all depend on the specific data being used (ie, different algorithms on different datasets will make different speed-accuracy-complexity tradeoffs and therefore there is no universally best model; see "no free lunch theorem").[71]

When analyzing feature importance for a given disorder, within predictive models, many types of features have shown to be predictive and their relevance seems to be influenced by how different algorithms capture information from different types of data. Within null-hypothesis testing studies, the most studied acoustic features were f0 mean, f0 variability, intensity variability, jitter, shimmer, and total time talking. Next, we discuss how acoustic features may relate to observable symptoms.

4.1 | Linking acoustic features to symptoms of psychiatric disorders

4.1.1 | Major depressive disorder

A decrease in f0 and f0 range in depressed individuals has repeatedly been observed,[56,59,72] which is a classic finding (for a review on speech patterns in depression, see Reference 29), and reflects the monotonous speech often seen in depression. Meanwhile, other acoustic features such as jitter, shimmer, and f0 variability tend to increase with depression severity and psychomotor retardation (ie, slowing of thought, physical movement, and reaction times) which affects motor control precision and laryngeal muscle tension.[56,59,72]

4.1.2 | Post-traumatic stress disorder

Marmar et al[73] interpreted the features that helped achieved high performance in detecting PTSD, which indicated that speech from individuals with PTSD is more monotonous, slower, and flatter. Similarly, other studies found reduced tonality in the vowel space (ie, F1 and F2 2D space for the vowels /a/, /i/, /u/)[74] and f0 variability.[75]

Acoustic feature	Description
Source features	*Features reflecting airflow from the lungs through the glottis (i.e. glottal features) or vocal fold vibrations (i.e., voice quality features), which is the sound source later filtered by the vocal tract following the source-filter theory of speech production.*[29]
Jitter [%]	Deviations in individual consecutive f0 period lengths, which indicates irregular closure and asymmetric vocal-fold vibrations.
Shimmer [%]	Difference of the peak amplitudes of consecutive f0 periods, which indicates irregularities in voice intensity.
Tremor [Hz]	Frequency of the most intense low-frequency fundamental frequency-modulating component in a specified analysis range.
Harmonics-to-noise ratio (HNR) [dB]	Ratio between f0 and noise components, which indirectly correlates with perceived aspiration. This may be due to reducing laryngeal muscle tension resulting in a more open, turbulent glottis.[56]
Frequency disturbance ratio (FDR) [%]	Relative mean value of the frequency disturbance from 5 to 5 periods (five points average)[60].
Amplitude Disturbance ratio (ADR) [%]	Relative mean amplitude value over a set of windows.[108]
Quasi-open quotient (QOQ)	Ratio of the vocal folds' opening time. Functional dysphonias often reduce QOQ range. Speaking loudly requires more effort with a low QOQ and sounds more stalled.
Normalized amplitude quotient (NAQ)	Ratio between peak-to-peak pulse amplitude and the negative peak of the differentiated flow glottogram and normalized with respect to the period time. It can be an estimate of glottal adduction.
Peak slope	Slope of the regression line that is fit to log10 of the maxima of each frame.[116]
Filter features	*The resonant properties of the vocal and nasal tracts filter the sound source from the vocal folds: the filter attenuates certain frequencies and strengthens others by the shape of the vocal and nasal tracts.*
F1 mean [Hz]	First peak in the spectrum (especially of voiced utterances such as vowels) that results from a resonance of the human vocal tract.
F2 mean [Hz]	Second peak in the spectrum (especially of voiced utterances such as vowels) that results from a resonance of the human vocal tract.
F1 variability [Hz]	Measures of dispersion of F1 (variance, standard deviation).
F2 variability [Hz]	Measures of dispersion of F2 (variance, standard deviation).
F1 range [Hz]	Difference between the lowest and highest F1 values.
Vowel space	F1 and F2 2D space for the vowels /a/, /i/, /u/.[74]
Linear predictive coding (LPC) coefficients	Coefficients that best predict the values of the next time point of the audio signal using the values from the previous n time points, which is used to reconstruct filter properties.
Spectral features	*Features characterizing the spectrum of speech, which is the frequency distribution of the speech signal at a specific time.*[29]
Mel-frequency cepstral coefficients (MFCCs)	The coefficients derived by computing a spectrum of the log-magnitude Mel-spectrum of the audio segment. The lower coefficients represent the vocal tract filter and the higher coefficients represent periodic vocal fold sources.
Prosodic features	*Changes over longer segments of time, which is perceived in the rhythm, stress, and intonation of speech.*[29]
f0 mean [Hz]	Fundamental frequency: lowest frequency of the speech signal, perceived as pitch (mean, median).
f0 variability [Hz]	Measures of dispersion of f0 (variance, standard deviation).
f0 range [Hz]	Difference between the lowest and highest f0 values.
Intensity [dB]	Defined as the acoustic intensity (i.e., power carried by sound per unit area in a direction perpendicular to that area) in decibels relative to a reference value, perceived as loudness.
Intensity variability [dB]	Measures of dispersion of intensity (variance, standard deviation).
Energy velocity	Measured as the mean-squared central difference across frames and may correlate with motor coordination[65].
Maximum phonation time [s]	The mean of three attempts of the following measure is taken: the maximum time during which phonation of a vowel (usually /a/) is sustained as long as possible with an upright position, deep breath, and a comfortable pitch and loudness[66].
Speech rate	Number of speech utterances per second over the duration of the speech sample (including pauses).
Articulation rate	Number of speech units per second over the duration of the speech sample (excluding pauses).
Time talking [s]	Sum of the duration of all speech segments.
Utterance duration mean [s]	Mean duration of utterance length.
Pause duration mean [s]	Mean duration of pause length.
Pause variability [s]	Measures of dispersion of pause duration (variance, standard deviation).
Pause rate [s]	Total length of pauses divided by the total length of speech (including pauses).
Pauses total [s]	Total duration of pauses.

FIGURE 4 Glossary of acoustic features. Classification based on References 29 and 56. For further discussion, see the Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[57] and Section 4.3.3

4.1.3 | Schizophrenia

Several studies found total time talking[76,77] and speech rate[78,79] to be significantly lower, while mean pause duration to be higher[76,77,80] in schizophrenia, which are measures poverty of speech and alogia, classical negative symptoms in schizophrenia.[53] Flat affect, another negative symptom, could be expressed by lower f0 mean and variability but results were mixed with many null results in line with a meta-analysis that showed weaker effects for atypicalities in pitch variability than ones in duration.[53]

4.1.4 | Bipolar disorder

Significant increases in tonality were observed including increases in median f0,[81,82] and mean F1 and F2.[83] Furthermore, a significantly higher number of longer pauses were observed in depressive states than in euthymic or hypomanic states,[81] and speech pauses became longer as patients transitioned into depressive states.[13] Changes in speech in depression and mania are well known clinically as they are captured by the psychomotor retardation item in the Hamilton Depression Rating[84] and the speech rate item on the Young Mania Rating Scale[85]; however, speech rate was not identified in the reviewed studies, and therefore, remains promising for future studies.

4.1.5 | Anxiety disorders

Many studies found a significant increase in mean f0 in social anxiety disorder[58,86-88] and generalized anxiety disorder,[89] with some studies finding null results.[88,90] This was the highest score for mean f0 across disorders (see Figure 3). Jitter and shimmer were also significantly higher in anxious patients.[89,90] Only one study[91] met inclusion criteria for OCD and found that the voices of individuals with OCD had significantly more jitter than healthy controls. Beyond automated feature extraction, a clinical evaluation of showed OCD voice to be significantly more hoarse and breathy and have a lower speech rate.

4.1.6 | Bulimia nervosa

Significantly higher source features such as jitter, shimmer, frequency disturbance ratio, and amplitude disturbance ratio have been found in bulimia nervosa along with a variety of laryngeal alterations due to vomiting (for reviews, see References 92-94). For instance, dysphonia has been observed[95] as well as pharyngeal reflux in singers who also presented vocal fold edema and polypoid changes.[96] Lesions such as laryngitis posterior, pharyngitis, and hematomas in the vocal folds could be caused by chronic irritation due to the presence of chyme during self-induced vomiting.[97] Therefore, these impairments may alter these source features originated in the vocal folds.

4.1.7 | Anorexia nervosa

Contradicting findings were found for mean f0[97,98] in anorexia nervosa. However, when analyzing only participants who presented the disorder before the menarche, it was found that they had significantly higher mean f0.[98] Furthermore, inappropriate larynx structure was observed in older patients along with a weak, asthenic voice, and some hyperfunctional dysphonia.[97] Therefore, it may be f0 is altered if anorexia affects puberty development due to being present before menarche or if anorexia has been present for enough time to cause weakness. More research is needed to clarify these findings.

4.2 | Guidelines for acquiring data

Studies tend to propose models for detecting disorders, however they vary greatly in sample size, demographics, confounds that were controlled, diagnosis criteria, speech-eliciting task and recording environment. Therefore, the detection will be biased to whichever criteria they used to acquire data. We summarize and discuss different strategies found in the reviewed articles to record speech, avoid confounds, and elicit relevant signals in speech for psychiatric-trait detection while safeguarding privacy (for further discussion, see Reference 99).

4.2.1 | Report comorbidities

Most studies reviewed included individuals with psychiatric comorbidities.[88] Scherer et al (2013) included subjects who presented a high correlation (Pearson's $r > 0.8$) between their depression PHQ-9 score and their PTSD PCL-C score.[116] However, most studies did not report the comorbidities. Few studies specifically addressed this problem and tried, for instance, to differentiate unipolar depression from bipolar disorder[100] and dysthymia from generalized anxiety disorder.[101] To better understand what is being classified, multiple diagnostic questionnaires should be used to detect comorbidities. Future research could also compare models built for populations with and without comorbidities.

4.2.2 | Detect symptoms or problems instead of disorders

There are very few publications focusing on predicting symptoms or problems. However, it would be desirable to link acoustic features to specific symptoms or problems that may be shared across disorders by detecting specific subitems within diagnostic questionnaires in line with the NIMH RDoC described in Section 1 (eg, Reference 102; for further discussion, see Reference 103). In psychiatry, there is a current trend to move from symptoms, which assume an underlying latent disease or disorder, to problems (eg, less sleep, lower energy), which may be related to underlying biological mechanisms.

4.2.3 | Consider additional confounds when selecting control group

A control group must not match diagnostic criteria by either being evaluated as neurotypical or within a different pathological group (eg, MDD in comparison to PTSD) and present statistically equivalent values for potential confounding variables that would affect speech. A strategy to improve classification is to discard intermediate scoring participants.[66] The most common controlled confounding variables in this review were age, sex (including sex-dependent classifiers), native language, and comorbidities, especially other psychiatric disorders, as well as traumatic brain injury, speech and respiratory disorders, cleft lip and palate, and substance abuse. Critically, most studies did not actually test the null hypothesis for confounds. Few studies controlled other variables that may affect speech patterns including race, education, and medication. Some psychiatric medications have shown to produce dry mouth, tremors,[104] and dyskinesia,[105] which impact speech. Therefore, medicated participants might be excluded (though this would affect sample generalizability) or medication should be reported. Other less studied variables that may affect speech and may vary within a group, thus inserting noise, are height, weight, dialectal variant, energetic state at the time of speech elicitation, and intimacy.[29] If these variables are statistically different between case and control groups, they can be better matched through techniques such as propensity score matching.[106]

4.2.4 | Self-report assessments may not match clinical diagnosis

Even though clinical evaluations by a psychiatrist or clinical psychologist are generally considered the "gold-standard" for diagnosis in comparison to self-report measures, this would be costly and inter-rater reliability can still be quite low.[27] Most studies in this review had a clinician evaluate participants instead of using self-report questionnaires. However, the widely used AVEC data sets for depression and PTSD[60-63] and bipolar disorder[64] used self-report measures. When using self-report measures, the goal of the study must be reframed from predicting diagnosis to predicting self-report questionnaire scores, which may not always match clinical diagnosis. Finally, selecting open access questionnaires are better for reproducibility purposes, since new studies could incorporate them.

4.2.5 | Use power analysis to determine sample size for null-hypothesis testing

The median psychiatric group size was 30 participants. However, closer to 74 participants per group would be needed to reach 95% power for reliable effect size estimates in null-hypothesis testing studies.[53] Machine learning models need a large enough test set to be representative of the general population (see Figure 1). Furthermore, predictive models such as deep neural networks are surprisingly effective but need a much larger amount of data than simple linear classifiers such as support vector machines because they need to adjust millions of weights to map input to output. An approximate rule of thumb is that a model needs 10 times more training samples than degrees of freedom to prevent overfitting[107]; however, this depends on the amount of input features, their distributions across disorders, and the underlying, often unknown, size of the effect.

4.2.6 | Use multiple tasks

Choosing the right task is important given a feature may correlate with a diagnostic scale using one task but not using another, and even change correlational direction using symptom subitems of a scale.[56] Using the classification found in Parola et al (in press)[53], we synthesize examples and advantages of different types of tasks in Table 3. Producing sustained vowels is optimal for measuring source features (eg, shimmer, jitter) since finding voiced sections in continuous speech is difficult.[114] Maximum phonation time of a sustained vowel with comfortable loudness measured by a stopwatch negatively correlated with years having anorexia[97] and can be caused by the weakening of respiratory muscles, decrease in subglottal pressure, and excessive tension of laryngeal muscles. Reducing laryngeal control could cause monotonous speech which is a classic sign of psychomotor retardation in depression.[56] Less ecologically valid methods can nevertheless provide more control over evoked emotions such as reading positive, negative, and neutral narratives since every participant elicits acoustic patterns constrained by reading the same text.[115] For instance, Scherer et al[116] found that positive and neutral questions differentiated a PTSD from a control group better than negative ones (for a meta-analysis on reactions to positive and negative stimuli, see Reference 117). Alghowinem et al[118] showed how this type of pattern changes according to what features are extracted, the polarity of the question, and what time segment is used to train models, showing that the first seconds performed better than using the whole recording. More ecological free speech responses can be followed by generic follow-up questions such as "How did this change your life?" so more data are acquired. Interestingly, interviews may be done by virtual humans or avatars which reduces costs, may increase comfortableness for some participants, and this can help re-enact dramatic scenarios the same way across users,[119-121] which has been used in the AVEC challenges for depression classification.[8]

4.2.7 | Use one microphone per speaker in interviews

When recording an interview between the participant and an interviewer, the main issue is being able to extract only speech segments belong to the participant to train models. To do this, one must separate speakers, a process known as diarization. This is much easier if there is a microphone next to each speaker. Headsets or lapel mics tend to be ideal, but may make certain participants more uncomfortable then desk microphones. However, differences in recording setup and distances between speaker and microphone can cause confounds.[122] Two smartphones can be

TABLE 3 Advantages of different types of speech-eliciting tasks

Task and examples		Advantages
Constrained	Sustained vowel[108] • Maximum phonation time[97]	Optimal for measuring source and respiration features • Captures muscle weakness and aspects of motor control
	Repeating "PATAKA"	Tests diadochokinetic rate,[109] captures speech sequencing, and is a proxy for lung capacity
	Counting[64]	More control over acoustic patterns using a common vocabulary
	Reading[110] • Emotion-evoking sentences • Rainbow passage • The Grandfather passage	• More control over evoked emotions • Contains every sound in English and is representative of normal speech[111] • Paragraph used to assess communication disorders[112]
Free speech	Monologue: describing, retelling happy, or traumatic memory[86]	More ecologically valid than reading
	Dialogue: • Semi-structured interviews[73] • Phone conversations[33,66,113]	Social dynamics (turn taking, intimacy) • Already done in many clinics • By not recording other caller, no need for diarization. Smartphones provide accelerometer data[13]

placed next to each speaker with an acoustic barrier in between to better detect speakers (see preprocessing section below). Finally, when saving the audio file, it is a good strategy to include all the information of the sample (participant ID, group, task, date) on the file name or in a separate file linked to a file ID. To avoid discrepancies between interviewers' file naming (eg, upper/lower case, spaces) and what is included in the recording, file names could be generated automatically.

4.2.8 | Strengthen privacy

Consent forms should be signed with clear indication as to whether participants' data can be shared with other research teams through request or publicly. However, even though this is authorized, it creates a risk for the participant having audio (or video) recordings of potentially vulnerable information such as what is shared in a clinical interview. Therefore, speech features such as those extracted by openSMILE[123] can be shared instead of their raw audio data. It is not possible to reconstruct the raw audio signal from these features; however, the participant can still be identified by matching the extracted features to new features extracted from a different recording of the same participant. Another approach to improve privacy is to filter recordings in real time or use bone conduction microphones, which can allow the extraction of acoustic data without being linguistically interpretable.[124] The disadvantage is that text data cannot be obtained later on. When text data are obtained from regular recordings, then it can be manually annotated to replace identifying information.[65] Another approach consists of capturing audio on the participant's device, extracting encrypted acoustic features from which the raw audio cannot be reconstructed, and then sending the features to a secure server to later download.[33] Distributed training[125] is a more novel approach where the actual model is trained on the participant's device, and then the learned weights or configuration—but not the data—is returned to the researcher, a training style which can now be done with tools such as TensorFlow.js.[126]

4.3 | Guidelines for machine learning models

An ideal scenario would entail having a model that can detect a disorder in a new person even if this person is recorded in a novel environment with a different age, accent, language, background noise, recording equipment, comorbidity, and tasks than the one provided in the training data—the process of generalization. However, the current state of the field is to try to detect a disorder in a new person given a model that is trained on other examples collected in similar settings, which is a much more limited form of generalization. The following guidelines are intended to help avoid overfitting and improve generalization.

4.3.1 | Preregister the model building protocol

Within null-hypothesis testing, given the large amount of acoustic features that can be extracted from a short time window of audio, the more features there are, it will be more likely to find a feature that significantly correlates with the disorder. A critical downside is that these features may correlate with a disorder in one dataset but not another. Within predictive models, multiple configurations can be tried out to increase performance; however, without confirming performance on an unseen test set, it is likely these results have overfitted the training data similar to p-hacking in null-hypothesis testing.[127] Therefore, preregistering the features hypothesized to correlate or be statistically different between groups and preregistering a protocol that specifies how models will be built and tested reduces the possibility of increasing bias via analytical choices.[128] In the preregistration process, exploratory research (where flexibility is encouraged to uncover new hypotheses) is distinguished from confirmatory research (where flexibility is denied to avoid confirmation and hindsight biases).[127] Furthermore, in general, it is important to always include nonsignificant results because they are important to judge the relative effect of positive results in other studies; this is guaranteed in preregistrations as they are stated as future tests. Though some studies in this review

test hypotheses,[74,88] we have not encountered preregistrations, which is becoming a common practice in other scientific fields.

4.3.2 | Preprocessing

Voice activity detection (VAD) can be used to obtain audio segments containing speech to discard silences and noise. Furthermore, continuous speech segments shorter than 1 second can be ignored as in Ringeval et al[8] (the exact value can be tested for performance). If two speakers are present (eg, in a clinical interview), the interviewer's segments may be discarded through automatic diarization. Diarization can be performed automatically through open-source packages (eg, Kaldi) or paid diarization systems such as rev.ai or Google Cloud Speech-to-Text. When two microphones are used, to avoid segments with interviewer's voice, VAD can be performed on both channels, and only participant's channel voiced segments with higher VAD score than the corresponding interviewer channel segments can be retained.[73] To avoid overfitting, preprocessing (eg, performing dimensionality reduction or feature selection) should be done separately on train and test sets.

4.3.3 | Feature extraction

Acoustic feature extraction can be done with open-source packages such as openSMILE, covarep, pyAudioAnalysis, openEAR, and Praat. As seen in Figure 3, similar features (eg, total pause and pause percentage) are extracted in different studies. Since even the same features can be extracted in slightly different ways, standardizing feature extraction provides the benefit of comparing results across. Examples include the extended GeMAPS[57] which was developed to determine a minimal but powerful set of features (88) that index voice changes during affective processes or brute-force approaches such as INTERSPEECH 2013 ComParE competition feature set[129] that uses many more features (6373) but may reduce generalization (by providing more features to overfit the specific training set used) and thus extracting more features may require larger sample sizes. We encourage testing statistical significance of a feature between populations even in predictive studies to advance the understanding of how these features may characterize a disorder. Furthermore, extracting text features is possible with automatic speech recognition (ASR) using open source platforms such as Mozilla DeepSpeech or commercial platforms such as rev.ai or Google Cloud Speech-to-Text. Text features include ngrams (ie, counts of words or phrases), semantic coherence (ie, stability of meaning from sentence to sentence), syntactic complexity, sentiment polarity, average sentence lengths, psychological domains (eg, cognitive processes, social processes),[130] and word embeddings (ie, word meaning). ASR can be improved by human annotators with tools such as ELAN.[131]

4.3.4 | Perform bootstrapping with small samples

Evaluating on a held-out test set only once is reasonable if the test set is representative of the population of interest. When using small data sets that are common in the medical field, a 20% test set or a 5-fold cross validation of a 100-person study will result in a biased estimate of how the model will generalize since is unlikely the test sets or folds are representative of the entire population being tested. Instead, repeated bootstrapping with repetition of, for example, 60 samples, can be performed relatively fast on small data sets (see Figure 5). This will result in a distribution of performance scores of which the mean or median can be taken as final performance metric. This approach may not be feasible with more complex models, such as deep neural networks, due to their higher computational complexity. In such cases, k-fold cross-validation may be used to provide a more expedient validation. Such complex models also require more data, and as data set size increases, there is less need for bootstrapped resampling.

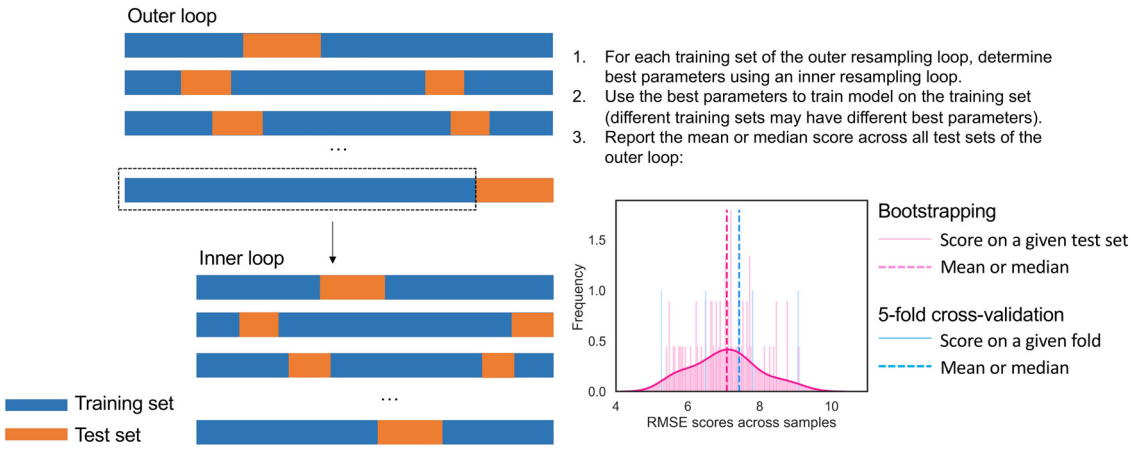

FIGURE 5 Nested bootstrapping for more robust performance estimation on small datasets and hyperparameter tuning. Example uses RMSE as performance metric on 60 bootstrapping samples and 5-fold cross-validation. K-fold cross-validation assumes large sample sizes and on small datasets may return a biased estimate of the underlying performance distribution. RMSE, root mean squared error

4.3.5 | Perform nested resampling for hyperparameter tuning

As described in Figure 5, to avoid overfitting, perform hyperparameter tuning (eg, grid or randomized search) over each training set of the resampling method be it with replacement (bootstrapping) or without replacement (k-fold cross-validation). These methods can be combined to reduce computational complexity (eg, the inner loop can be done with k-fold cross-validation).

4.3.6 | Test performance statistically through a permutation test

It is not sufficient to achieve a higher score than chance, but this score must have a reasonable effect size and be statistically significant to be of clinical use. For instance, one study[66] indicates that performance is significantly higher than baseline with a paired *t* test. However, to increase confidence in generalizing results even more, a permutation test can be performed where models are trained on randomly sorted labels to evaluate how much the model can learn from noise and the inherent biases in the data, which often surpasses chance. This can also be performed following the bootstrapping procedure. Then a paired-test can be used to test significance between the permuted and nonpermuted distributions of scores.

4.3.7 | Report multiple metrics and consider class imbalance

When dealing with unbalanced classes (more cases of healthy than the psychiatric population), for classification tasks it is important to not use accuracy since it will be biased toward true negatives, and instead use F1-score. It is also relevant to report F1-score for both positive and negative classes. Additional important metrics to report are precision, recall, and area under the receiver operating characteristic curve (ROC AUC). For visualization, when classification is binary and very unbalanced, precision-recall curve plots can provide a more accurate description than ROC plots.[132] For regression tasks, common metrics are root mean squared error, mean absolute error, and the coefficient of determination (r^2). An alternative metric is the concordance correlation coefficient, which is not altered by changes in scale and location and includes information on precision and accuracy.[8] Overall, it is useful to report multiple metrics (eg, the ones mentioned in this section) since it is difficult to compare studies that report different metrics.

4.3.8 | Explain and interpret models to reduce bias and improve scientific understanding

Fortunately, the field of interpretable machine learning is trying to systematize what interpretability means and how it can be measured.[133,134] For a practical perspective, there is a book covering many useful tools[135] and packages that compare multiple explainability methods.[136] Explaining which features are predictive of a given disorders not only allows debugging but also allows the field to create new hypotheses and better understand the disorders to ultimately create generative models. Feature importance can be done by additive feature attribution methods,[137] feature selection methods, correlating each feature with the diagnostic severity, testing whether a feature is significantly different between groups, retraining with important features alone to measure if they are sufficient for optimal model performance, or all of the above. Other types of explanations include counterfactual and adversarial examples.[135] Some studies in this review presented excellent quantitative descriptions of a feature in the psychiatric and control group but lacked statistical analysis. Furthermore, as many articles have done, it is useful to attempt to link changes in acoustic features to psychiatric behaviors or symptoms (eg, low f0 variability with flat affect), and use these links for testing hypotheses. There is an ongoing debate around whether complex, difficult to interpret models that perform well should be sacrificed for lower performing but simpler to interpret models.[138] From our point of view, in current medicine, we would not want to discard complex diagnostic tools (eg, biopsy) for simpler but less effective ones (eg, lump palpation). Similarly, the precise mechanisms of many drugs are not well understood, but they are used because they have been proven to work through clinical trials. Therefore, validating complex models on large, representative samples is key since they may have biases as described in Section 1.

4.3.9 | Release code and data through a container to improve reproducibility

To reproduce results of a machine learning study, data sets and code must be provided. However, clinical data sets do not always contain the permission to be shared, but the features extracted from raw audio could be shared under proper permission (see section 4.2.8). Providing code is the main tool to compare studies that use different methods and evaluation metrics. Finally, even if data and code are shared, they are often not reproducible since code might be incomplete or dependencies might not be specified. Therefore, we encourage using containers such as Docker and Singularity that contain code, data, packages, and a basic operating system, which can be rerun easily to re-execute original analyses or replicate analyses on other data.[139,140] Innovation might be faster if models could be tweaked in a more efficient manner.

4.3.10 | Competitions promote research but do not necessarily produce useful models

As stated before, 32% of studies in this review used AVEC data sets during or after competitions. Some of these studies present useful innovations in feature extraction and model design. However, one

challenge competitions face is that the more results a team presents, the more likely incorrect inferences are to occur since they may overfit the test set with one of their models by chance. Still, teams are often allowed to submit more than once. The issues with multiple-hypothesis testing count equally if multiple submissions occur within teams or across teams. Given the relatively small test sets that are provided in competitions for clinical problems, it is likely the winning model happens to win the second best submission just by chance, because it happened to overfit the small test set slightly better. Therefore, we cannot trust that novel models that perform slightly better in competitions will generalize because of this "crowd overfitting"[141]: the top N performing models may actually perform similarly in the population, even though one happens to work best for the competition test set. Even more concerning, the fifth best performing model in the competition may be the best performing model in the population. Some solutions to competition overfitting include performing multiple-comparison correction across submissions and prioritizing simpler models.[142,143]

4.4 | Future approaches

4.4.1 | Limitations

Given the breadth of psychiatric disorders included in this review, keywords were only searched in the title and not in the abstract. The resulting articles were then screened by reading title and/or abstract. Therefore articles where speech description was not the main focus may have been missed if "speech" or related terms were not in the title (eg, studies that measured speech among other behaviors).

4.4.2 | Understudied disorders

More studies need to be done on disorders beyond depression, schizophrenia, and bipolar disorder such as OCD, bulimia, anorexia, anxiety disorders, and personality disorders. It is likely the most studied disorders were inspired by clinician's intuitions that speech may provide a link to diagnosis. Considering machine learning can find structure in ways that are nonintuitive for humans, it is likely there will be other disorders that also carry a signal in speech.

4.4.3 | Generalization

Even though models work for one data set, we do not know if they will generalize to a new sample or similar samples that vary in age, geography, socio-economic level, recording setup, and task. Alghowinem et al[144] tested performance when training and testing on different datasets from different countries, languages, and accents (see also References 59, 145, and 146). Several studies[66,147] compared performance across different smartphones which result in different amounts of clipping, loudness, and noise, which is important to achieve device-independent predictions. Mitra et al measured the effect on performance of noise and reverberation changes between train and test sets on depression detection.[148]

4.4.4 | Longitudinal studies

The vast majority of studies were cross-sectional. It is not understood if acoustic features that are predictive cross-sectionally across individuals are also predictive longitudinally within individuals. Several bipolar disorder studies[33,149,150] covered in this review took a longitudinal approach to capture different states (eg, manic, hypomanic, euthymic, depressed), this approach is not often taken in other disorders even though symptoms can naturally oscillate and disorders remiss. A few studies on MDD[112,151] analyzed changes in symptom severity and treatment impact longitudinally through speech patterns.

4.4.5 | Multimodal learning

Even though this review focuses on speech, many studies provided multimodel models trained on audio and video recordings such as those from AVEC competitions, in which some multimodal models reported improved performance as compared to unimodal models.[70,152] Some studies combined these types of features with neurophysiological measures such as electroencephalography.[153]

4.4.6 | From disorders to diseases

Machine learning might change diagnostic criteria given personalized medicine and continuous, real-time monitoring,[154] especially considering the limitations diagnostic criteria may currently have.[26,30,103,155] By linking behavioral and biological features to symptoms instead of diagnoses, we could further understand the underlying diseases and endophenotypes that gives rise to the personalized configuration of symptoms and reduce the need of traditional disorders.[34] Furthermore, not all acoustic features are measured across studies. Therefore, using standardized feature sets (see Section 4.3.3), performing null-hypothesis testing in parallel to machine learning studies, and including data from other modalities (eg, text, video, accelerometer) is a step toward characterizing more features across disorders and understanding underlying diseases further.

4.4.7 | Risks of this technology

It is extremely important to thoroughly assess the ethical implications of this research. Insurance companies and employers could turn down applicants if they predict a psychiatric disorder is present or will develop. Friends and foes could gain insight into our private mental lives by obtaining samples of our voice or other behaviors. Even when data are shared consensually, understanding what consent actually

implies is challenging.[156] More clear information and examples should be provided on how exactly data might be used. Furthermore, channels to deactivate consent could be offered intermittently. Since most models have not actually been validated through a clinical trial, it would seem these risks do not exist at present. However, companies may still use these models as if they were valid, and given the exponential growth of technology they may achieve validity soon. Those developing technology should be aware of the multiple ways systems can fail and strategies to prevent failures, discrimination, and negative side effects.[157] As a community of scientists, technologists, and clinicians, we must participate in debates with citizens and policy makers to help prevent abuses and safeguard the advancement of a technology that could help so many.

4.5 | Breaking the barrier between psychiatry and laryngology

The barrier between psychiatry and neurology can be somewhat arbitrary and are rooted in distinct historical practices in the 20th century.[158,159] With this review article, we hope to demonstrate that psychiatry and laryngology have more in common than previously thought for several reasons. First, a substantial amount of individuals attending otolaryngology centers are suffering from mental health disorders[160] and so to provide care, mental health could be assessed for potential referrals. These disorders may be independent to their complaint or they may actually be causing alterations that make them see an otolaryngologist in the first place, as seen in the relationship between anxiety, depression, and tinnitus.[161,162] Most voice alterations presented in this review are not necessarily voice disorders, simply patterns linked to disorders. One critical exception is bulimia nervosa where voice alterations have been observed.[92-94]

Furthermore, regardless of whether the underlying cause is psychiatric or laryngeal, having a voice pathology tends to produce distress[163] as seen with dysphonia[164] and stuttering.[160] Given this and potential psychogenic disorders, laryngologists also have the challenging task of promoting psychiatric consultation and psychotherapy in a way that reduces the associated stigma, since it is currently underutilized: from 1998 to 2007, psychotherapy use in the US population decreased from 3.4% in 1998 to 3.2% in 2007 even though the distribution of mental health disorders is much higher.[165]

Another reason the two fields can overlap is that laryngology visits can include extensive voice evaluations which are often recorded. Therefore, given what has been shown in this review, it seems reasonable (and relatively low burden) to start assessing mental health during ENT visits. This would create a rich source of data to link mental health assessments with acoustic features easily obtained from recordings of voice evaluations. Even though the voice disorder will confound certain acoustic features that tend to predict psychiatric disorders, other features may remain unconfounded. With the guidelines previously presented on how to acquire data, we hope this will open new collaborations between laryngologists, psychiatrists, and machine learning specialists.

5 | CONCLUSIONS

A total of 127 studies were reviewed that measure acoustic features from speech to distinguish psychiatric from healthy individuals either through null-hypothesis testing or predictive machine learning models. We provided a synthesis of significant and nonsignificant acoustic features across disorders as well as those that correlate with a given disorder severity. We discussed guidelines on how to acquire data, prevent confounds, safeguard privacy, select speech-eliciting tasks, and improve generalization and reproducibility of machine learning models. Certain disorders have been less studied such as eating and anxiety disorders. More studies have been carried out in MDD, PTSD, and bipolar disorder thanks to open-access research data sets provided by AVEC competitions[60-64] and the DAIC data set.[65] Competitions in particular provide a common framework to compare innovations under equal data and evaluation metrics, measure performance with a held-out test set to estimate overfitting (but there is still a great need to improve crowd overfitting), and allow future studies to be done with the same dataset. Therefore, we encourage creating open data sets, if possible through competitions, as they have shown to be highly productive. Whereas productivity is healthy, reproducibility is key: since the studies in this review build computational models, data and code can easily be shared—ideally through containers—to test claims and make gradual innovations as a community. Furthermore, more studies using multiple datasets and preregistering hypotheses could help improve generalization and resolve conflicting findings regarding the significant and predictive acoustic features in each disorder. In closing, building machine learning models on speech seems a promising pathway toward improving mental-health assessments and treatments in line with preventive and personalized medicine.

ACKNOWLEDGMENTS

The authors would like to thank Robert Ajemian for his useful comments on Figure 1. D.M.L. was supported by a National Institutes of Health training grant (5T32DC000038-28). K.H.B. was partially supported by the MIT-Philips Research Award for Clinicians. The work was supported by a gift to the McGovern Institute for Brain Research at MIT. S.S.G was partially supported by 5P41EB019936.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Daniel M. Low https://orcid.org/0000-0002-8866-8667
Satrajit S. Ghosh https://orcid.org/0000-0002-5312-6729

ENDNOTE

[1] The table from which Figure 3 was built is provided online along with code to create Table 2 and all figures (https://github.com/danielmlow/review) so that these can be reproduced and updated.

REFERENCES

1. Merikangas KR, He J-P, Burstein M, et al. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). *J Am Acad Child Adolesc Psychiatry*. 2010;49(10):980-989.
2. Substance Abuse and Mental Health Services Administration. *Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health (HHS Publication No. SMA 18-5068, NSDUH Series H-53)*. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2018.
3. Trautmann S, Rehm J, Wittchen H. The economic costs of mental disorders. *EMBO Rep*. 2016;17(9):1245-1249.
4. Substance Abuse and Mental Health Services Administration. *Results from the 2014 National Survey on Drug Use and Health: Mental Health Findings, NSDUH Series H-50, HHS Publication No.(SMA) 15-4927*. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2015.
5. Goessl VC, Curtiss JE, Hofmann SG. The effect of heart rate variability biofeedback training on stress and anxiety: a meta-analysis. *Psychol Med*. 2017;47(15):2578-2586.
6. Miranda D, Calderón M, Favela J. Anxiety detection using wearable monitoring. In Proceedings of the 5th Mexican Conference on Human-Computer Interaction. Oaxaca, Mexico: 2014.
7. Williamson JR, Godoy E, Cha M, et al. Detecting depression using vocal, facial and semantic communication cues. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC '16), New York, NY: ACM; 2016:11-18.
8. Ringeval F, Schuller B, Valstar M, et al. AVEC 2019 workshop and challenge: state-of-mind, depression with AI, and cross-cultural affect recognition. Proceedings of the 2019 on Audio/Visual Emotion Challenge and Workshop. ACM; Nice, France: 2019.
9. Yang L, Li Y, Chen H, Jiang D, Oveneke MC, Sahli H. Bipolar disorder recognition with histogram features of arousal and body gestures. Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC '18). NYC, USA: ACM; 2018:15-21.
10. Syed ZS, Sidorov K, Marshall D. Automated screening for bipolar disorder from audio/visual modalities. Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC '18). NYC, USA: ACM; 2018:39-45.
11. Scherer S, Morency LP, Rizzo A. Multisense and SimSensei—a multimodal research platform for real-time assessment of distress indicators. In: 2012 Conference, Arlington, VA, October 19.
12. Gravenhorst F, Muaremi A, Bardram J, et al. Mobile phones as medical devices in mental disorder treatment: an overview. *Pers Ubiquit Comput*. 2015;19(2):335-353.
13. Maxhuni A, Muñoz-Meléndez A, Osmani V, Perez H, Mayora O, Morales EF. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive Mob Comput*. 2016;31:50-66.
14. Likforman-Sulem L, Esposito A, Faundez-Zanuy M, Clémençon S, Cordasco G. EMOTHAW: a novel database for emotional state recognition from handwriting and drawing. *IEEE Trans Hum Machine Syst*. 2017;47(2):273-284.
15. Ghosh SS, Baker JT. Will neuroimaging produce a clinical tool for psychiatry? *Psychiatr Ann*. 2019;49(5):209-214.
16. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. *Neuroimage Clin*. 2016;10:115-123.
17. Librenza-Garcia D, Kotzian BJ, Yang J, et al. The impact of machine learning techniques in the study of bipolar disorder: a systematic review. *Neurosci Biobehav Rev*. 2017;80:538-554.
18. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci*. 2017;18:43-49.
19. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. *Nat Lang Eng*. 2017;23(5):649-685.
20. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. *Int J Methods Psychiatr Res*. 2016;25(2):86-100.
21. Mohr DC, Ho J, Duffecy J, et al. Perceived barriers to psychological treatments and their relationship to depression. *J Clin Psychol*. 2010;66(4):394-409.
22. Turner RJ, Jay Turner R, Lloyd DA, Taylor J. Physical disability and mental health: an epidemiology of psychiatric and substance disorders. *Rehabil Psychol*. 2006;51(3):214-223.
23. Shalev A, Liberzon I, Marmar C. Post-traumatic stress disorder. *N Engl J Med*. 2017;376(25):2459-2469.
24. Rathbone AL, Clarry L, Prescott J. Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: systematic review. *J Med Internet Res*. 2017;19(11):e399.
25. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250.
26. Freedman R, Lewis DA, Michels R, et al. The initial field trials of DSM-5: new blooms and old thorns. *Am J Psychiatry*. 2013;170(1):1-5.
27. Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry*. 2013;170(1):59-70.
28. Gideon J, Schatten HT, Mc Innis MG, Provost EM. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In: The 20th Annual Conference of the International Speech Communication Association INTERSPEECH; Sep. 15-19, Graz, Austria: 2019.
29. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun*. 2015;71:10-49.
30. Insel TR. The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. *Am J Psychiatry*. 2014;171(4):395-397.
31. Huang K, Wu C, Su M, Kuo Y. Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model. *IEEE Trans Affect Comput*. 2018;9:563-577.
32. Bedi G, Carrillo F, Cecchi GA, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr*. 2015;1:15030.
33. Faurholt-Jepsen M, Busk J, Frost M, et al. Voice analysis as an objective state marker in bipolar disorder. *Transl Psychiatry*. 2016;6:e856.
34. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230.
35. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2018;52(6):446-462.
36. Koh PW, Liang P. Understanding black-box predictions via influence functions. Proceedings of the 34th International Conference on Machine Learning – ICML'17. Vol 70. Sydney, Australia: JMLR.org; 2017:1885-1894.
37. Kleinberg J, Mullainathan S. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. 2019.
38. Foster KR, Koprowski R, Skufca JD. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomed Eng Online*. 2014;13:94.
39. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.

40. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access*. 2018;6:14410-14430.
41. Regulation P. Regulation (EU) 2016/679 of the European Parliament and of the council. *Regulation*. 2016;679:2016.
42. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag*. 2017;38(3): 50-57.
43. Gunning D. Explainable Artificial Intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web 2017;2. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf. Accessed December 25, 2019.
44. Denes PB, Pinson EN. *The Speech Chain: The Physics and Biology of Spoken Language*. Murray Hill, NJ: Bell Telephone Laboratories; 1963.
45. Kraepelin E. Manic depressive insanity and paranoia. *J Nerv Ment Dis*. 1921;53(4):350.
46. Snowden LR. Bias in mental health assessment and intervention: theory and evidence. *Am J Public Health*. 2003;93(2):239-243.
47. Ely JW, Graber ML, Croskerry P. Checklists to reduce diagnostic errors. *Acad Med*. 2011;86(3):307-313.
48. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. *Curr Opin Psychiatry*. 2014;27(3):203-209.
49. Bzdok D, Ioannidis JPA. Exploration, inference, and prediction in neuroscience and biomedicine. *Trends Neurosci*. 2019;42(4): 251-262.
50. Morales M, Scherer S, Levitan R. A cross-modal review of indicators for depression detection systems. Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality; 2017:1-12.
51. Tokuno S. Pathophysiological voice analysis for diagnosis and monitoring of depression. In: Kim Y-K, ed. *Understanding Depression. Clinical Manifestations, Diagnosis and Treatment*. Vol 2. Singapore: Springer; 2018:83-95.
52. Cohen AS, Mitchell KR, Elvevåg B. What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. *Schizophr Res*. 2014;159(2-3):533-538.
53. Parola A, Simonsen A, Bliksted V, Fusaroli R. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. *Schizophr Res*. In press.
54. Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009;6(7):e1000097.
55. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. Washington, DC: American Psychiatric Publishing; 2013.
56. Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J. On the relative importance of vocal source, system, and prosody in human depression. In: 2013 IEEE International Conference on Body Sensor Networks; 2013:1-6.
57. Eyben F, Scherer KR, Schuller BW, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans Affect Comput*. 2016;7(2):190-202.
58. Gilboa-Schechtman E, Galili L, Sahar Y, Amir O. Being "in" or "out" of the game: subjective and acoustic reactions to exclusion and popularity in social anxiety. *Front Hum Neurosci*. 2014;8:147.
59. Kiss G, Vicsi K. Mono- and multi-lingual depression prediction based on speech processing. *Int J Speech Technol*. 2017;20(4):919-935.
60. Valstar M, Schuller B, Smith K, et al. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge - AVEC '13. Barcelona, Spain: 2013.
61. Valstar M, Schuller B, Smith K, et al. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14. Orlando, USA: 2014.
62. Valstar M, Pantic M, Gratch J, et al. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16. Amsterdam, USA: 2016.
63. Ringeval F, Schuller B, Valstar M, et al. Real-life depression, and affect recognition workshop and challenge. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17). Mountain View, USA: ACM; 2017:3-9.
64. Ringeval F, Schuller B, Valstar M, et al. AVEC 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition. Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC'18). Beijing, China: 2018:3-13.
65. Gratch J, Artstein R, Lucas GM, et al. The distress analysis interview corpus of human and computer interviews. In: LREC. Citeseer; 2014: 3123-3128.
66. Gideon J, Provost EM, McInnis M. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. *Proc IEEE Int Conf Acoust Speech Signal Process*. Shanghai, China: March 20-25, 2016:2359-2363.
67. Xing X, Cai B, Zhao Y, Li S, He Z, Fan W. Multi-modality hierarchical recall based on GBDTs for bipolar disorder classification. Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC'18). Beijing, China: 2018:31-37.
68. Afshan A, Guo J, Park SJ, Ravi V, Flint J, Alwan A. Effectiveness of voice quality features in detecting depression. *Interspeech*. Hyderabad, India: 2018;1676-1680.
69. Kächele M, Schels M, Schwenker F. Inferring depression and affect from application dependent meta knowledge. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC '14). Orlando, NYC: 2014:41-48.
70. Williamson JR, Quatieri TF, Helfer BS. Vocal and facial biomarkers of depression based on motor incoordination and timing. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; Orlando, NYC: 2014.
71. Ho YC, Pepyne DL. Simple explanation of the no-free-lunch theorem and its implications. *J Optimiz Theory Appl*. 2002;115(3): 549-570.
72. Quatieri TF, Malyska N. Vocal-source biomarkers for depression: a link to psychomotor activity. In: Thirteenth Annual Conference of the International Speech Communication Association; Portland, USA, Sept. 9-13: 2012.
73. Marmar CR, Brown AD, Qian M, et al. Speech-based markers for posttraumatic stress disorder in US veterans. *Depress Anxiety*. 2019; 36(7):607-616.
74. Scherer S, Lucas GM, Gratch J, Skip Rizzo A, Morency L. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Trans Affect Comput*. 2016;7(1):59-73.
75. Xu R, Mei G, Zhang G, et al. A voice-based automated system for PTSD screening and monitoring. *Stud Health Technol Inform*. 2012; 173:552-558.
76. Kliper R, Vaizman Y, Weinshall D, Portuguese S. Evidence for depression and schizophrenia in speech prosody. In: Third ISCA Workshop on Experimental Linguistics; Saint-Malo, France: June 19-23, 2010.
77. Kliper R, Portuguese S, Weinshall D. Prosodic analysis of speech and the underlying mental state. In Serino S, Matic A, Giakoumis D, Lopez G, Cipresso, P (Eds.), *Pervasive Computing Paradigms for Mental Health*. NY, NY: Springer International Publishing; 2016:52-62.
78. Perlini C, Marini A, Garzitto M, et al. Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder. *Acta Psychiatr Scand*. 2012;126(5):363-376.
79. Tahir Y, Yang Z, Chakraborty D, et al. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. *PLoS One*. 2019;14(4):e0214314.
80. Rapcan V, D'Arcy S, Yeap S, Afzal N, Thakore J, Reilly RB. Acoustic and temporal analysis of speech: a potential biomarker for schizophrenia. *Med Eng Phys*. 2010;32(9):1074-1079.

81. Guidi A, Schoentgen J, Bertschy G, Gentili C, Scilingo EP, Vanello N. Features of vocal frequency contour and speech rhythm in bipolar disorder. *Biomed Signal Process Control*. 2017;37:23-31.
82. Guidi A, Scilingo EP, Gentili C, Bertschy G, Landini L, Vanello N. Analysis of running speech for the characterization of mood state in bipolar patients. 2015 AEIT International Annual Conference (AEIT); Naples, Italy: 2015.
83. Zhang J, Pan Z, Gui C, et al. Analysis on speech signal features of manic patients. *J Psychiatr Res*. 2018;98:59-63.
84. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6(4):278-296.
85. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry*. 1978;133:429-435.
86. Weeks JW, Lee C-Y, Reilly AR, et al. "The sound of fear": assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder. *J Anxiety Disord*. 2012;26(8):811-822.
87. Galili L, Amir O, Gilboa-Schechtman E. Acoustic properties of dominance and request utterances in social anxiety. *J Soc Clin Psychol*. 2013;32(6):651-673.
88. Weeks JW, Srivastav A, Howell AN, Menatti AR. "Speaking more than words": classifying men with social anxiety disorder via vocal acoustic analyses of diagnostic interviews. *J Psychopathol Behav Assess*. 2016;38(1):30-41.
89. Özseven T, Düğenci M, Doruk A, Kahraman Hİ. Voice traces of anxiety: acoustic parameters affected by anxiety disorder. *Arch Acoust*. 2018;43(4):625-636.
90. Silber-Varod V, Kreiner H, Lovett R, Levi-Belz Y, Amir N. Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in social anxiety disorder individuals. Proceedings of Speech Prosody. Boston, USA: 2016:1211-1215.
91. Cassol M, Reppold CT, Ferrão Y, Gurgel LG, Almada CP. Análise de características vocais e de aspectos psicológicos em indivíduos com transtorno obsessivo-compulsivo [Analysis of vocal characteristics and psychological aspects in individuals with obsessive-compulsive disorder]. *Rev Soc Bras Fonoaudiol*. 2010;15(4):491-496.
92. Cielo CA, Didoné DD, Torres EMO, de Moraes JP. Laryngopharyngeal reflux and bulimia nervosa: laryngeal and voice disorders. *Rev CEFAC*. 2011;13(2):352-361.
93. Rajiah K, Mathew EM, Veettil SK, Kumar S. Bulimia nervosa and its relation to voice changes in young adults: a simple review of epidemiology, complications, diagnostic criteria and management. *J Res Med Sci*. 2012;17(7):689-693.
94. Balata P, Colares V, Petribu K, Leal M de C. Bulimia nervosa as a risk factor for voice disorders–literature review. *Braz J Otorhinolaryngol*. 2008;74(3):447-451.
95. Rothstein SG, Rothstein JM. Bulimia: the otolaryngology head and neck perspective. *Ear Nose Throat J*. 1992;71(2):78-80.
96. Rothstein SG. Reflux and vocal disorders in singers with bulimia. *J Voice*. 1998;12(1):89-90.
97. Maciejewska B, Rajewska-Rager A, Maciejewska-Szaniec Z, Michalak M, Rajewski A, Wiskirska-Woźnica B. The assessment of the impact of anorexia nervosa on the vocal apparatus in adolescent girls—a preliminary report. *Int J Pediatr Otorhinolaryngol*. 2016;85:141-147.
98. Garcia-Santana C, Capilla P, Blanco A. Alterations in tone of voice in patients with restrictive anorexia nervosa: a pilot study. *Clin Salud*. 2016;27(2):71-87.
99. Kächele M, Schels M, Schwenker F. The influence of annotation, corpus design, and evaluation on the outcome of automatic classification of human emotions. *Front ICT*. 2016;3:17.
100. Yang T-H, Wu C-H, Huang K-Y, Su M-H. Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. *J Amb Intel Hum Comput*. 2017;8(6):895-906.
101. Wang J, Sui X, Zhu T, Flint J. Identifying comorbidities from depressed people via voice analysis. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Kansas City, USA: Nov. 13-16, 2017:986-991.
102. Bernardini F, Lunden A, Covington M, et al. Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in Italy and the United States. *Psychiatry Res*. 2016;239:253-258.
103. Arseniev-Koehler A, Mozgai S, Scherer S. What type of happiness are you looking for? A closer look at detecting mental health from language. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; 2018:1-12.
104. Crawford AA, Lewis S, Nutt D, et al. Adverse effects from antidepressant treatment: randomised controlled trial of 601 depressed individuals. *Psychopharmacology*. 2014;231(15):2921-2931.
105. De Hert M, Detraux J, van Winkel R, Yu W, Correll CU. Metabolic and cardiovascular adverse effects associated with antipsychotic drugs. *Nat Rev Endocrinol*. 2011;8(2):114-126.
106. Pan W, Flint J, Shenhav L, et al. Re-examining the robustness of voice features in predicting depression: compared with baseline of confounders. *PLoS One*. 2019;14(6):e0218172.
107. Abu-Mostafa YS, Magdon-Ismail M, Lin HT. *Learning from Data*. Vol 4. New York, NY: AML Book; 2012.
108. Ferreira CP, Gama ACC, Santos MAR, Maia MO. Laryngeal and vocal analysis in bulimic patients. *Braz J Otorhinolaryngol*. 2010;76(4):469-477.
109. Ziegler W. Task-related factors in oral motor control: speech and oral diadochokinesis in dysarthria and apraxia of speech. *Brain Lang*. 2002;80(3):556-575.
110. Kiss G, Vicsi K. Comparison of read and spontaneous speech in case of automatic detection of depression. 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom); Debrecen, Hungary: 2017:213-218.
111. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of voice acoustics as predictors of clinical depression scores. *J Voice*. 2017;31(2):256.e1-256.e6.
112. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguist*. 2007;20(1):50-64.
113. Karam ZN, Baveja SS, Mcinnis M, Provost EM. Mood monitoring of bipolar disorder using speech analysis. US Patent June 2017. https://patentimages.storage.googleapis.com/c8/59/21/9dddce335fd4fd/US9685174.pdf. Accessed July 30, 2019.
114. Kane J, Aylett M, Yanushevskaya I, Gobl C. Phonetic feature extraction for context-sensitive glottal source processing. *Speech Commu*. 2014;59:10-21.
115. van den Broek EL, van der Sluis F, Dijkstra T. Telling the story and re-living the past: how speech analysis can reveal emotions in post-traumatic stress disorder (PTSD) patients. In: Westerink J, Krans M, Ouwerkerk M, eds. *Sensing Emotions: The Impact of Context on Experience Measurements*. Dordrecht, The Netherlands: Springer; 2011:153-180.
116. Scherer S, Stratou G, Gratch J, Morency L-P. Investigating voice quality as a speaker-independent indicator of depression and PTSD. *Interspeech*. Lyon, France: Aug. 25-29, 2013;847-851.
117. Bylsma LM, Morris BH, Rottenberg J. A meta-analysis of emotional reactivity in major depressive disorder. *Clin Psychol Rev*. 2008;28(4):676-691.
118. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. Detecting depression: a comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; Vancouver, Canada: 2013:7547-7551.
119. DeVault D, Artstein R, Benn G, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14). Paris, France: 2014:1061-1068.

120. Hartholt A, Traum D, Marsella SC, et al. All together now. *Intelligent Virtual Agents*. Berlin, Germany: Springer; 2013:368-381.
121. Burton C, Tatar AS, McKinstry B, et al. Pilot randomised controlled trial of Help4Mood, an embodied virtual agent-based system to support treatment of depression. *J Telemed Telecare*. 2016;22(6):348-355.
122. Cummins N, Epps J, Sethu V, Krajewski J. Variability compensation in small data: oversampled extraction of i-vectors for the classification of depressed speech. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Florence, Italy: 2014:970-974.
123. Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia (MM '10). Indianapolis, USA: 2010:1459-1462.
124. Nemes V, Nikolic D, Barney A, Garrard P. A feasibility study of speech recording using a contact microphone in patients with possible or probable Alzheimer's disease to detect and quantify repetitions in a natural setting. *Alzheimers Dement*. 2012;8(4):P490-P491.
125. McClure P, Zheng CY, Kaczmarzyk J. Distributed weight consolidation: a brain segmentation case study. Adv Neural Inf Process Syst 2018. Montreal, Canada: 2018.
126. Smilkov D, Thorat N, Assogba Y, et al. TensorFlow.js: machine learning for the web and beyond. *arXiv [csLG]*. January 2019.
127. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci*. 2017;12(6):1100-1122.
128. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proc Natl Acad Sci U S A*. 2018;115(11):2600-2606.
129. Schuller B, Steidl S, Batliner A, et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France; 2013.
130. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol*. 2010;29(1):24-54.
131. Tacchetti M. User Guide for ELAN Linguistic Annotator; 2017. http://www.mpi.nl/corpus/manuals/manual-elan_ug.pdf. Accessed on December 25, 2019.
132. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
133. Lipton ZC. The mythos of model interpretability. arXiv [csLG]. June 2016.
134. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv [statML]. February 2017.
135. Molnar C. Interpretable machine learning. Lulu.com; 2019. https://christophm.github.io/interpretable-ml-book/. Accessed December 25, 2019.
136. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: a unified framework for machine learning interpretability. arXiv [csLG]. September 2019.
137. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. Vol.30. Red Hook, NY: Curran Associates; 2017:4765-4774.
138. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215.
139. Boettiger C. An introduction to docker for reproducible research. *Oper Syst Rev*. 2015;49(1):71-79.
140. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. *PLoS One*. 2017;12(5):e0177459.
141. Oakden-Rayner L. AI Competitions Don't Produce Useful Models. https://lukeoakdenrayner.wordpress.com/2019/09/19/ai-competitions-dont-produce-useful-models/. Published September 19, 2019. Accessed December 25, 2019.
142. Mount J. A Deeper Theory of Testing. Win-Vector Blog. http://www.win-vector.com/blog/2015/09/a-deeper-theory-of-testing/. Published September 26, 2015. Accessed December 25, 2019.
143. Blum A, Hardt M. The ladder: a reliable leaderboard for machine learning competitions. arXiv [csLG]. February 2015. http://arxiv.org/abs/1502.04585.
144. Alghowinem S, Goecke R, Epps R, Wagner M, Cohn J. Cross-cultural depression recognition from vocal biomarkers. *Interspeech*. San Francisco, USA: Sept. 8-12, 2016;1943-1947.
145. Mitra V, Shriberg E, Vergyri D, Knoth B, Salomon RM. Cross-corpus depression prediction from speech. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Brisbane, USA: 2015:4769-4773.
146. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Commun*. 2015;75:27-49.
147. Stasak B, Epps J. Differential performance of automatic speech-based depression classification across smartphones. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW); San Antonio, USA: 2017:171–175.
148. Mitra V, Tsiartas A, Shriberg E. Noise and reverberation effects on depression detection from speech. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Shanghai, China: 2016:5795-5799.
149. Karam ZN, Provost EM, Singh S, et al. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. *Proc IEEE Int Conf Acoust Speech Signal Process*. Florence, Italy: 2014; 2014:4858-4862.
150. Muaremi A, Gravenhorst F, Grünerbl A, Arnrich B, Tröster G. Assessing bipolar episodes using speech cues derived from phone calls. In Serino S, Matic A, Giakoumis D, Lopez G, Cipresso P (eds.). *Pervasive Computing Paradigms for Mental Health*. NY, NY: Springer International Publishing; 2014:103-114.
151. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. *IEEE Trans Affect Comput*. 2013;4(2):142-150.
152. He L, Jiang D, Sahli H. Multimodal depression recognition with dynamic visual and audio cues. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII); Xian, China: 2015:260-266.
153. Acharya UR, Rajendra Acharya U, Oh SL, et al. Automated EEG-based screening of depression using deep convolutional neural network. *Comput Methods Programs Biomed*. 2018;161:103-113.
154. Friston KJ, Redish AD, Gordon JA. Computational nosology and precision psychiatry. *Comput Psychiatr*. 2017;1:2-23.
155. Allsopp K, Read J, Corcoran R, Kinderman P. Heterogeneity in psychiatric diagnostic classification. *Psychiatry Res*. 2019;279:15-22.
156. Custers B. Click here to consent forever: expiry dates for informed consent. *Big Data Soc*. 2016;3(1):2053951715624935.
157. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv [csAI]. June 2016.
158. Baker MG, Kale R, Menken M. The wall between neurology and psychiatry. *BMJ*. 2002;324(7352):1468-1469.
159. Ibáñez A, García AM, Esteves S, et al. Social neuroscience: undoing the schism between neurology and psychiatry. *Soc Neurosci*. 2018;13(1):1-39.
160. Iverach L, O'Brian S, Jones M, et al. Prevalence of anxiety disorders among adults seeking speech therapy for stuttering. *J Anxiety Disord*. 2009;23(7):928-934.

161. Paolini AG. Trait anxiety affects the development of tinnitus following acoustic trauma. *Neuropsychopharmacology*. 2012;37(2):350-363.
162. Gomaa MAM, Elmagd MHA, Elbadry MM, Kader RMA. Depression, anxiety and stress scale in patients with tinnitus and hearing loss. *Eur Arch Otorhinolaryngol*. 2014;271(8):2177-2184.
163. Marmor S, Horvath KJ, Lim KO, Misono S. Voice problems and depression among adults in the United States. *Laryngoscope*. 2016;126(8):1859-1864.
164. Martinez CC, Cassol M. Measurement of voice quality, anxiety and depression symptoms after speech therapy. *J Voice*. 2015;29(4):446-449.
165. Lannin DG, Guyll M, Vogel DL, Madon S. Reducing the stigma associated with seeking psychotherapy through self-affirmation. *J Couns Psychol*. 2013;60(4):508-519.

How to cite this article: Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*. 2020;5:96–116. https://doi.org/10.1002/lio2.354