Evaluation of an AI-Based Voice Biomarker Tool to Detect Signals Consistent With Moderate to Severe Depression

Alexa Mazur[1]
Harrison Costantino, MS[2]
Prentice Tom, MD[1]
Michael P. Wilson, MD, PhD[3]
Ronald G. Thompson, PhD[3]

[1]Kintsugi Mindful Wellness, Inc, San Francisco, California

[2]Department of Computer Science, University of California, Berkeley, California

[3]Departments of Psychiatry and Emergency Medicine, University of Arkansas for Medical Sciences, Little Rock, Arkansas

MORE ONLINE
www.annfammed.org

VISUAL ABSTRACT

Annals Early Access article

Annals Journal Club selection

Conflicts of interest: A.M., P.T., and H.C. are currently employed by Kintsugi Mindful Wellness, Inc, have equity in the company, and have played a role in the investigational device's development. Kintsugi Mindful Wellness received a grant from the National Science Foundation to conduct this research but had full control of the data and the decision to submit this manuscript to Annals of Family Medicine. M.P.W. and R.G.T. have no conflicts of interest to disclose.

CORRESPONDING AUTHOR

Alexa Mazur
Kintsugi Mindful Wellness, Inc
2210 Jackson Street 701
San Francisco, CA 94115
aam2213@columbia.edu

ABSTRACT

PURPOSE Mental health screening is recommended by the US Preventive Services Task Force for all patients in areas where treatment options are available. Still, it is estimated that only 4% of primary care patients are screened for depression. The goal of this study was to evaluate the efficacy of machine learning technology (Kintsugi Voice, v1, Kintsugi Mindful Wellness, Inc) to detect and analyze voice biomarkers consistent with moderate to severe depression, potentially allowing for greater compliance with this critical primary care public health need.

METHODS We performed a cross-sectional study from February 1, 2021 to July 31, 2022 to examine ≥25 seconds of free-form speech content from English-speaking samples captured from 14,898 unique adults in the United States and Canada. Participants were recruited via social media, provided informed consent, and their voice biomarker results were compared with a self-reported Patient Health Questionnaire-9 (PHQ-9) at a cut-off score of 10 (moderate to severe depression).

RESULTS From as few as 25 seconds of free-form speech, machine learning technology was able to detect vocal characteristics consistent with an increased PHQ-9 ≥10, with a sensitivity of 71.3 (95% CI, 69.0-73.5) and a specificity of 73.5 (95% CI, 71.5-75.5).

CONCLUSIONS Machine learning has potential utility in helping clinicians screen patients for moderate to severe depression. Further research is needed to measure the effectiveness of machine learning vocal detection and analysis technology in clinical deployment.

Ann Fam Med 2024;23:online. https://doi.org/10.1370/afm.240091

INTRODUCTION

Depression is a leading cause of disability, affecting an estimated 18 million Americans each year, with a lifetime prevalence of major depression approaching 30%.[1-3] In 2016, the US Preventive Services Task Force recommended universal depression screening for adult patients when adequate follow-up is available.[4,5] Still, depression screening rarely occurs in the outpatient setting, with some estimates placing screening rates at <4% of primary care encounters.[6-9] Even when identified to undergo screening, patients with depression are included <50% of the time.[6,10] Thus, there is a substantial opportunity and need to improve primary care screening for depression. Machine learning (ML) can help fill this care gap by augmenting clinical workflows without additional clerical burden, to increase the frequency of depression screening and accelerate patient triage.[11-14]

Individuals with an active depressive episode have distinct speech patterns such as more frequent stuttering and hesitations, longer and more frequent pauses, and slower speech cadence.[15-22] Vocal signatures associated with a clinical diagnosis are defined as voice biomarkers.[23] Using ML technology to evaluate these voice signatures represents a novel, noninvasive, quantitative, reproducible, and near seamless assessment that can be added to virtual encounters. We sought to assess whether ML can effectively detect vocal characteristics consistent with a moderate to severe acute depressive episode.

METHODS

Dataset

All data were obtained with approval from the Solutions IRB (https://www.solutionsirb.com) institutional review board. We performed this study in accordance

with relevant regulations and guidance including the Declaration of Helsinki. The study population included adults aged ≥18 years in the United States and Canada recruited via social media (ie, Reddit, Craigslist, Facebook, and Instagram). Because young and/or female individuals were more likely to self-enroll in this study, additional advertisements on social media using images of men and older individuals were directed at male and senior populations to strive for a more evenly distributed study sample. From February 1, 2021 to July 31, 2022, 14,898 participants provided informed consent, completed the Patient Health Questionnaire-9 (PHQ-9), and recorded a voice response to the prompt, "How was your day?" for at least 25 seconds in English using their personal electronic device's microphone from their remote location (Figure 1). Responses were captured using a secure online survey platform. Participants self-reported demographic information, which included age, gender, race and ethnicity, and country of residence, to assess sample representativeness and eligibility. We collected e-mail addresses to distribute compensation of $5 (USD or CAD) on study completion for eligible participants. Other identifying information, such as name or phone number, was not collected, for protection of participant privacy.

Study Measurement

Participants were evaluated via completion of the PHQ-9 questionnaire. Scores for the 9 items range from 0 ("Not at all") to 3 ("Nearly every day"), and total scores range from 0 to 27. A PHQ-9 score of ≥10 was the threshold for a moderate to severe acute depressive episode because it maximizes the sensitivity and specificity of the PHQ-9 instrument.[8,9]

Data Processing

Surveys were individually reviewed by study staff for completion, uniqueness, and authenticity. Incomplete, duplicate, or fraudulent surveys (eg, outside the United States or Canada) were excluded from analysis. The eligible audio recordings were captured as .wav files, and linear pulse code modulation, sampling rate, and voice activity were standardized to limit variations in quality introduced by differences in participants' personal electronic device microphones. Preserving consistent audio quality was accomplished by converting files to 16-kHz linear pulse code modulation, which is the standard for speech processing and minimizes file degradation.[24,25] Full details regarding data processing and model architecture and training (Kintsugi Voice, v1, Kintsugi Mindfull Wellness) are described in Supplemental Appendixes 1 and 2.

Model Evaluation

Predictions were normalized, scaled from 0 to 1. Values closer to 1 represented a greater confidence score for the model's belief that the participant had vocal characteristics consistent with a moderate to severe acute depressive episode. We selected the following 3 predicted model outputs: (1) Signs of Depression Detected for individuals with sufficient vocal characteristics consistent with an active depressive episode; (2) Signs of Depression Not Detected for individuals who had insufficient vocal characteristics consistent with an active depressive episode; and (3) Further Evaluation Recommended, which captured individuals for whom the model did not have sufficient confidence to yield output and would defer to clinician judgment for a formal screening determination in practice.

Quantitatively, Signs of Depression Detected corresponded to model output values >0.5631 or = 1 and anticipated a PHQ-9 score of ≥10. Signs of Depression Not Detected corresponded to model output values = 0 and <0.4449. Values from 0.4449 to 0.5631 were labeled Further Evaluation Recommended. We set threshold values were set to minimize the presence of false-positive and false-negative samples. We evaluated overall model performance by comparing model outputs to self-reported PHQ-9 scores.

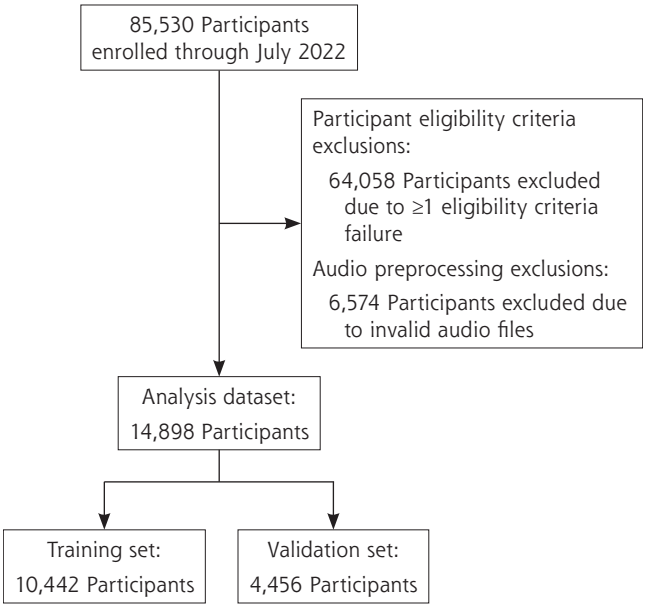
Statistical Analysis

We assessed sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), along with Wald 95% CIs.[26,27] We calculated subpopulation performance for subpopulations with a sufficient representation.

RESULTS

The number of participant files in the training and validation data sets were 10,442 and 4,456, respectively (Figure 1). The

Figure 1. Participant Exclusion and Audio Preprocessing Criteria Used to Create Training and Validation Sets to Train and Tune the Model and Evaluate its Performance



Note: Eligible participants for inclusion in the analysis data sets were adults aged ≥18 years living in the United States or Canada who provided a voice sample in English containing at least 25 seconds of speech content meeting audio quality parameters. The training set and validation set were split to evenly distribute samples on the basis of participant characteristics and audio length.

validation set had a speech content range of 25.0-74.9 seconds, (median = 57.9 seconds, mode = 58.5 seconds) and self-reported PHQ-9 score range of 0-27, (median = 9, mode = 0). Demographic characteristics are summarized in Table 1, model performance is shown in Table 2, and subpopulation model performance is shown in Table 3. The subpopulation participant demographics for misclassified samples are listed in Supplemental Table 1.

The model provided an output of Signs of Depression Detected and Signs of Depression Not Detected for 3,536 of the validation samples. The performance of these predictions was as follows: overall sensitivity of the model was 71.3 (95% CI, 69.0-73.5), specificity was 73.5 (95% CI, 71.5-75.5), PPV was 69.3 (95% CI, 67.1-71.5), and NPV was 75.3 (95% CI, 73.3-77.2) (Table 2). The output Further Evaluation Recommended was returned for 20% of the overall validation set, or 920 samples.

Within subpopulations, model sensitivity was greatest for Hispanic or Latine (80.3; 95% CI, 72.6-86.6) and Black or African American (72.4; 95% CI, 64.0-79.8) populations, and model specificity was greatest for Asian or Pacific Islander (77.5; 95% CI, 72.8-81.8) and Black or African American (75.9; 95% CI, 69.3-81.7) populations, which all had wider CIs relative to the full sample and White subpopulation. Sensitivity and specificity for women and men were notably different; sensitivity and specificity for women were 74 (95% CI, 71.4-76.5) and 68.9 (95% CI, 66.2-71.4), and those for men were 59.3 (95% CI, 54.0-64.4) and 83.9 (95% CI, 80.8-86.7). The population aged <60 years had a sensitivity (71.9; 95% CI, 69.5-74.2) and specificity (71.8; 95% CI, 69.6-73.9) with narrower CIs than the population aged ≥60 years (63.4; 95% CI, 54.3-71.9) and (86.8; 95% CI, 81.6-91.0).

Table 1. Participant Demographic Characteristics

Characteristic	Training	Validation
Age, y		
Average (SD)	37.3 (14.3)	37.3 (14.2)
Median	34.0	34.0
Mode	22.0	25.0
Range	18-93	18-86
Gender, %		
Female	69.5	69.4
Male	27.3	27.9
Not specified	2.4	2.1
Other	0.9	0.7
Race/ethnicity, %		
Asian or Pacific Islander	15.9	16.2
Black or African American	9.4	9.4
Hispanic or Latine	7.5	7.7
Native American or American Indian	1.2	1.3
Not specified	1.5	1.8
Other or mixed race	5.9	5.8
White	58.5	57.8
Audio duration, s		
Average (SD)	55.1 (10.1)	55.0 (10.1)
Median	57.9	57.9
Mode	58.6	58.5
Range	25.0-74.9	25.0-74.9
PHQ-9 score		
Average (SD)	9.8 (6.7)	9.7 (6.7)
Median	9.0	9.0
Mode	9.0	0
Range	0-27	0-27

PHQ-9 = Patient Health Questionnaire-9

Table 2. Model Performance

Metric	Value (95% CI)
Sensitivity	71.3 (69.0-73.5)
Specificity	73.5 (71.5-75.5)
PPV	69.3 (67.1-71.5)
NPV	75.3 (73.3-77.2)

NPV = negative predictive value; PPV = positive predictive value.

Table 3. Subpopulation Performance

Metric	Value (95% CI)
Sensitivity	
All	71.3 (69.0-73.5)
Gender	
Female	74.0 (71.4-76.5)
Male	59.3 (54.0-64.4)
Age, y	
<60	71.9 (69.5-74.2)
≥60	63.4 (54.3-71.9)
Race/ethnicity	
Asian or Pacific Islander	67.4 (60.7-73.7)
Black or African American	72.4 (64.0-79.8)
Hispanic or Latine	80.3 (72.6-86.6)
White	70.7 (67.7-73.5)
Specificity	
All	73.5 (71.5-75.5)
Gender	
Female	68.9 (66.2-71.4)
Male	83.9 (80.8-86.7)
Age, y	
<60	71.8 (69.6-73.9)
≥60	86.8 (81.6-91.0)
Race/ethnicity	
Asian or Pacific Islander	77.5 (72.8-81.8)
Black or African American	75.9 (69.3-81.7)
Hispanic or Latine	68.6 (60.1-76.3)
White	72.8 (70.0-75.4)

DISCUSSION

The present study showed the preliminary effectiveness of ML to detect vocal characteristics consistent with a moderate to severe acute depressive episode from audio clips of ≥25 seconds of free-form speech content. The ML system showed an overall sensitivity of 71.3, specificity of 73.5, PPV of 69.3, and NPV of 75.3 relative to a PHQ-9 with a cutoff score of 10. Many mental health diagnostic and screening inventories have a performance ranging from 60.0-90.0 for both sensitivity and specificity.[8,28,29] Thus, the performance of the tool relative to the PHQ-9 suggests it might be effective for an ML device to assist in screening and identifying individuals with depression.[6,13] Machine learning voice-based approaches used to detect other conditions, such as bulbar dysfunction in amyotrophic lateral sclerosis show similar performance criteria.[30] An ML-based instrument for depression screening holds promise because it could increase the proportion of patients screened, without undue clinician clerical burden.

As with any medical device, it is important to consider false positives and false negatives. Minimizing false positives versus false negatives is a natural trade-off and can be adjusted via the model threshold depending on the demands and objectives of the clinical setting. We set our threshold to >0.5631 for this study. Given that the majority of persons meeting binary diagnostic criteria for a depressive episode have mild to moderate symptoms, and many do not need clinical interventions and are able to successfully manage a mild depressive episode without formal therapy or medication, increased sensitivity might be a worthwhile aim for future exploration and warrants clinician feedback via formal study.[4,5,10,12,14,17,31-33] False-negative detection might result in patients experiencing an active depressive episode missing formal screening and gaining access to subsequent behavioral health treatment; however, because the tool is intended as an adjuvant screening tool, and the proportion of false negatives is in line with the performance of other screening tools, this risk should be minimal within the context of current practices.[6-9] Additional acceptability studies tailored to specific environments will be required to quantify and qualify the implication of false-positive and false-negative screenings using ML technology to augment clinical workflow for depression screening.

Among the false negatives identified, there was a greater proportion of men (31.7%) relative to the population in the overall validation set (27.9%). There was an observable difference in the sensitivity measure for men (59.3; 95% CI, 54.0-64.4) relative to the overall population (71.3; 95% CI, 69.0-73.5). Whereas there is less precision regarding the estimate for men relative to the full population, owing to a smaller population of men in our sample, the sensitivity measure is still within the bounds of other precedent depression inventories.[8,29] The study team made efforts to recruit additional men and seniors to participate in the study; however, there is documented resistance to participating in depression-related research among these groups, and depression in the general population is recorded to be greatest among women and individuals aged <25 years, which could have influenced participants' motivations to volunteer.[34,35] The lower representation of men (27.3%) in training data relative to women (69.5%) might have resulted in decreased exposure to the population's characteristics for model learning, contributing to lower observed performance on validation data. The lower sensitivity suggests that the model might need to be better trained at identifying signs of depression in men, given that research has shown that artificial intelligence algorithms might falsely correlate a more masculine voice with decreased likelihood of depression, owing to the fact that depression is less prevalent among men.[36]

Segmenting by age, the <60 years population comprised a larger proportion of the data set and had narrower CIs than the ≥60 years group. We used the comprehensive sample age because we believe the results were clinically significant across the entire age range to warrant consideration of the voice biomarkers across the age spectrum. Many biomarkers (electrocardiogram morphology, blood pressure, lipid profile) are age dependent, and we suspect that this might be true of voice biomarkers. The etiologies of these age-related changes in voice biomarkers are difficult to speculate on and could include both age-related differences in voice as well as age-dependent neuromotor manifestations of depression. Further study of this phenomenon is warranted and could result in even more accurate screening via use of age-specific voice analytic biomarker tools. Similarly, further study and honing of the ML device to other patient characteristics that might allow for increased accuracy and value to clinicians is also warranted.

We note several strengths and limitations that need to be addressed in future ML model depression screening studies. First, as a substantial strength, the data set was diverse in socioeconomic population characteristics and consisted of a diverse regional representation across the United States and Canada, which captures a breadth of speech patterns and accents and is comparable in distribution to the racial makeup of the United States and Canada according to aggregate Census data.[37,38] Because we did not collect information on comorbid conditions, future studies should expand the overall data set and capture the relevant medical history of conditions affecting vocal production to help further understand any effects on voice biomarkers.

To correct imbalances in the representativeness of the study sample, we used targeted ads with images of men and seniors during recruitment. Persistent sample bias might be due to recruiting via social media or because depressed individuals might be more likely to participate in depression-based research. The average PHQ-9 score of the sample was 9.8, and just over 45% of participants scored >10, which is increased relative to the 8.6% prevalence of major depressive episodes in the United States.[39] Although sensitivity and specificity are generally stable predictors of test performance, PPV is increased and NPV is decreased based on

the prevalence of disease in the sample and might have been affected in our results.[40] Future study designs should use purposeful rather than convenient sampling frames to achieve representativeness. The increased prevalence in our training sample might allow the model to gain exposure to a broad spectrum of depression cases, however, which is important for generalizability given the nonuniform clinical presentation of depression.

Finally, the ML device was trained using the PHQ-9, which has reliably shown both a sensitivity and specificity of 88% in screening for depression; however, like the PHQ-9, the ML device is not intended as a substitute for a formal clinical interview and qualified clinician assessment, which remains the reference standard for confirming the presence of a depressive episode or clinical depression, nor is it meant to substitute for a comprehensive psychiatric evaluation for those that might be experiencing a mood disorder such as a major depressive disorder.[8,9] The ML device is not intended as a standalone tool for screening or diagnosing depression, and we are presenting these data to show how the ML device might be used by qualified clinicians, particularly primary care physicians such as family medicine doctors, as an adjuvant tool to help in their monitoring and screening of their patients for depression.

The present study represents one of the first attempts to train and validate ML technology to evaluate clips of free-form speech to detect signs of a depressive episode. Findings from this study suggest that harnessing ML technology to evaluate speech for the detection of signs of a depressive episode is effective compared with the PHQ-9 at a cutoff score of 10. This study supports that the use of ML technology as a clinical decision-support tool might be a step toward universal depression screening, a primary care objective recommended by the US Preventive Services Task Force.[4,5] Although this ML device technology is a breakthrough, and we believe it is important to communicate the performance of this technology at this juncture, we emphasize that this study represents an initial study validating how an ML device that analyzes a purely physiologic biometric (voice biomarkers), not dependent on patient or clinician interpretation and thus not subject to the inherent biases of natural language processing devices that interpret speech content, can be used to help validate and direct clinician action. We recognize that future studies are needed and that there is an expectation that this technology will continue to evolve and improve. Future studies will be directed toward determining the acceptability of augmenting primary care workflows with ML technology as a clinical decision-support tool and assessing the effect of other conditions that might influence depression voice biomarker analysis.

Read or post commentaries in response to this article.

Key words: machine learning; artificial intelligence; depression; voice biomarkers

Submitted February 20, 2024; submitted, revised, September 18, 2024; accepted September 19, 2024.

Funding support: A.M., H.C., and P.T. were employed by Kintsugi Health (dba: Kintsugi Mindful Wellness, Inc) during the period this study was conducted. This work was supported by the National Science Foundation (grants #2036213 and #1938831).

Data statement: This manuscript complies with the policies and practices outlined for its respective program by the National Science Foundation SBIR funding, under which some data components are considered proprietary and may not be shared. The authors will respond to reasonable requests and comply as appropriate.

Acknowledgments: The authors extend their sincerest gratitude to Victoria Graham, Chase Walker, and Brandn Green for their invaluable contributions to this work.

Supplemental materials

References

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1545-1602. doi: 10.1016/S0140-6736(16)31678-6
2. Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*. 2005;62(6):617-627. doi: 10.1001/archpsyc.62.6.617
3. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen HU. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *Int J Methods Psychiatr Res*. 2012;21(3):169-184. doi: 10.1002/mpr.1359
4. US Preventive Services Task Force. Screening for depression in adults. Published Jan 26, 2016. Accessed Nov 2, 2024. https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/depression-in-adults-screening
5. Park LT, Zarate CA Jr. Depression in the primary care setting. *N Engl J Med*. 2019;380(6):559-568. doi: 10.1056/NEJMcp1712493
6. Bhattacharjee S, Goldstone L, Vadiei N, Lee JK, Burke WJ. Depression screening patterns, predictors, and trends among adults without a depression diagnosis in ambulatory settings in the United States. *Psychiatr Serv*. 2018;69(10):1098-1100. doi: 10.1176/appi.ps.201700439
7. Jackson JL, Kuriyama A, Bernstein J, Demchuk C. Depression in primary care, 2010-2018. *Am J Med*. 2022;135(12):1505-1508. doi: 10.1016/j.amjmed.2022.06.022
8. Levis B, Benedetti A, Thombs BD; DEPRESSion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. 2019;365:l1476. doi: 10.1136/bmj.l1476
9. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA*. 1999;282(18):1737-1744. doi: 10.1001/jama.282.18.1737
10. Pignone MP, Gaynes BN, Rushton JL, et al. Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2002;136(10):765-776. doi: 10.7326/0003-4819-136-10-200205210-00013
11. Coulehan JL, Schulberg HC, Block MR, Madonia MJ, Rodriguez E. Treating depressed primary care patients improves their physical, mental, and social functioning. *Arch Intern Med*. 1997;157(10):1113-1120.
12. Frank RG, Huskamp HA, Pincus HA. Aligning incentives in the treatment of depression in primary care with evidence-based practice. *Psychiatr Serv*. 2003;54(5):682-687. doi: 10.1176/appi.ps.54.5.682
13. Kroenke K, Jackson JL, Chamberlin J. Depressive and anxiety disorders in patients presenting with physical complaints: clinical predictors and outcome. *Am J Med*. 1997;103(5):339-347. doi: 10.1016/s0002-9343(97)00241-6
14. Pfoh ER, Janmey I, Anand A, Martinez KA, Katzan I, Rothberg MB. The impact of systematic depression screening in primary care on depression identification and treatment in a large health care system: a cohort study. *J Gen Intern Med*. 2020;35(11):3141-3147. doi: 10.1007/s11606-020-05856-5

15. Moses PJ. *The Voice of Neurosis.* Grune and Stratton; 1954.
16. Darby JK, Hollien H. Vocal and speech patterns of depressive patients. *Folia Phoniatr (Basel).* 1977;29(4):279-291. doi: 10.1159/000264098
17. Darby JK, Simmons N, Berger PA. Speech and voice parameters of depression: a pilot study. *J Commun Disord.* 1984;17(2):75-85. doi: 10.1016/0021-9924(84)90013-3
18. Nilsonne A. Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr Scand.* 1987;76(3):235-245. doi: 10.1111/j.1600-0447.1987.tb02891.x
19. Scherer KR, Zei B. Vocal indicators of affective disorders. *Psychother Psychosom.* 1988;49(3-4):179-186. doi: 10.1159/000288082
20. Stassen HH. Modelling affect in terms of speech parameters. *Psychopathology.* 1988;21(2-3):83-88. doi: 10.1159/000284547
21. Stassen HH, Bomben G, Günther E. Speech characteristics in depression. *Psychopathology.* 1991;24(2):88-105. doi: 10.1159/000284700
22. Kuny S, Stassen HH. Speaking behavior and voice sound characteristics in depressive patients during recovery. *J Psychiatr Res.* 1993;27(3):289-307. doi: 10.1016/0022-3956(93)90040-9
23. Shin D, Cho WI, Park CHK, et al. Detection of minor and major depression through voice as a biomarker using machine learning. *J Clin Med.* 2021;10(14):3046. doi: 10.3390/jcm10143046
24. Library of Congress. Linear pulse code modulated audio (LPCM). Published Apr 19, 2022. Last updated Mar 26, 2024. Accessed Jul 11, 2023. https://www.loc.gov/preservation/digital/formats/fdd/fdd000011.shtml
25. Abhang PA, Gawali BW, Mehrotra SC. Technological basics of EEG recording and operation of apparatus. In: Abhang PA, Gawali BW, Mehrotra SC, eds. *Introduction to EEG- and Speech-Based Emotion Recognition.* Academic Press; 2016:19-50.
26. Patino CM, Ferreira JC. Confidence intervals: a useful statistical tool to estimate effect sizes in the real world. *J Bras Pneumol.* 2015;41(6):565-566. doi: 10.1590/S1806-37562015000000314
27. Bonett DG, Price RM. Adjusted Wald confidence interval for a difference of binomial proportions based on paired data. *J Educ Behav Stat.* 2012;37(4):479-488. www.jstor.org/stable/23256833
28. Costa MV, Diniz MF, Nascimento KK, et al. Accuracy of three depression screening scales to diagnose major depressive episodes in older adults without neurocognitive disorders. *Br J Psychiatry*. 2016;38(2):154-156. doi: 10.1590/1516-4446-2015-1818
29. Arroll B, Goodyear-Smith F, Crengle S, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med*. 2010;8(4):348-353. doi: 10.1370/afm.1139
30. Tena A, Clarià F, Solsona F, Povedano M. Voiceprint and machine learning models for early detection of bulbar dysfunction in ALS. *Comput Methods Programs Biomed*. 2023;229:107309. doi: 10.1016/j.cmpb.2022.107309
31. Dorow M, Löbner M, Pabst A, Stein J, Riedel-Heller SG. Preferences for depression treatment including internet-based interventions: results from a large sample of primary care patients. *Front Psychiatry*. 2018;9:181.doi: 10.3389/fpsyt.2018.00181
32. Zuroff DC, Blatt SJ. Vicissitudes of life after the short-term treatment of depression: roles of stress, social support, and personality. *J Soc Clin Psychol*. 2002;21(5):473-496. doi: 10.1521/jscp.21.5.473.22622
33. Nasser EH, Overholser JC. Recovery from major depression: the role of support from family, friends, and spiritual beliefs. *Acta Psychiatr Scand*. 2005;111(2):125-132. doi: 10.1111/j.1600-0447.2004.00423.x
34. Woodall A, Morgan C, Sloan C, Howard L. Barriers to participation in mental health research: are there specific gender, ethnicity and age related barriers? *BMC Psychiatry*. 2010;10:103. doi: 10.1186/1471-244X-10-103
35. National Institute of Mental Health. Major depression. Published 2022. Last updated Jul 2023. Accessed Jul 11, 2023. https://www.nimh.nih.gov/health/statistics/major-depression
36. Bailey A, Plumbley MD. Gender bias in depression detection using audio features. arXiv:2010:15120. doi: **10.48550/arXiv.2010.15120**
37. United States Census Bureau. Census QuickFacts. Accessed Aug 28, 2024. www.census.gov/quickfacts/fact/table/US/PST045222
38. Statistics Canada. Census profile, 2021 census of population. Profile table. Published Nov 15, 2023. Date modified Aug 2, 2024. Accessed Nov 2, 2024. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?LANG=E&GENDERlist=1,2,3&STATISTIClist=4&DGUIDlist=2021A0000\11124&HEADERlist=2,31&SearchText=Canada
39. Substance Abuse and Mental Health Services Administration. 2023 National Survey on Drug Use and Health (NSDUH) releases. Accessed Nov 2, 2024. https://www.samhsa.gov/data/release/2023-national-survey-drug-use-and-health-nsduh-releases#detailed-tables
40. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina (Kaunas).* 2021;57(5):503. doi: 10.3390/medicina57050503