Supplemental material for

Mazur A, Costantino H, Tom P, Wilson MP, Thompson RG. Evaluation of an AI-based voice biomarker tool to detect signals consistent with moderate to severe depression. *Ann Fam Med.* 2025;(23)1:60-65.

Supplemental Table 1. Subpopulation Participant Demographics

Demographic	False Positive	False Negative	Further Evaluation Recommended
Age (years)			
Average (SD)	35.5 (13.1)	37.8 (14.3)	36.0 (14.1)
Median	32.0	34.0	32.0
Mode	27.0	31.0	25.0
Range	18 - 80	18 - 82	18 - 85
Gender (%)			
Female	77.6	66.2	70.0
Male	19.8	31.7	27.3
Not Specified	1.2	0.4	1.0
Other	1.4	1.7	1.7
Race and Ethnicity (%)			
Asian or Pacific Islander	15.5	15.1	16.7
Black or African American	9.2	8.0	9.9
Hispanic or Latino	8.4	5.8	7.5
Native American / American Indian	0.8	1.5	1.3
Not Specified	2.2	1.5	1.8
Other or Mixed Race	5.9	6.5	6.5
White	58.0	61.6	56.2
Audio Duration (seconds)			
Average (SD)	56.3 (9.0)	53.8 (10.4)	54.0 (10.9)
Median	58.4	57.5	57.4
Mode	59.3	59.4	57.4
Range	25.5 - 74.9	25.1 - 74.5	25.2 - 74.8
PHQ -9 (Score)			
Average (SD)	6.1 (2.4)	14.2 (3.7)	9.7 (6.1)
Median	7.0	13.0	9.0
Mode	8.0	11.0	11.0

Supplemental Appendix 1. Data Processing, Model Architecture, and Training

The yielded raw audio data was used to create a unique feature set using two feature extraction tools: (1) Torchaudio to produce a 40-band Mel spectrogram and 20 mel-frequency cepstral coefficients (MFCCs), and (2) Meta AI's data2vec to extract 1024-dimensional representations of speech.[1,2] In addition to the raw audio, the three feature vectors and their deltas were produced in a method that emphasized precisely selected feature sets over more algorithmic layers. Functionals, including mean, standard deviation, and zero crossing rate were extracted for each vector before 512-dimensional speaker embeddings were extracted using Pyannote.[3] The vectors and embeddings were concatenated, which yielded a large final feature space. To address overfitting, Scikit-learn was used to scale the data to lie in the range of (-1,1) and the kernel principal component analysis was used to reduce the dimensionality.[4] Further overfitting reduction was performed using the rbf kernel, where 1024 components were retained. Principal Component Analysis (PCA) was used to assess the relationship between the features and PHQ-9 labels by projecting the high-dimensional data into 3-D space (**Supplemental Figure 1)**. In summary, these technical processes are the mechanisms by which the Kintsugi instrument utilizes a set of voice markers, as trained by Kintsugi, to examine key indicators of depression.

Supplemental Appendix 2. Model Architecture and Training

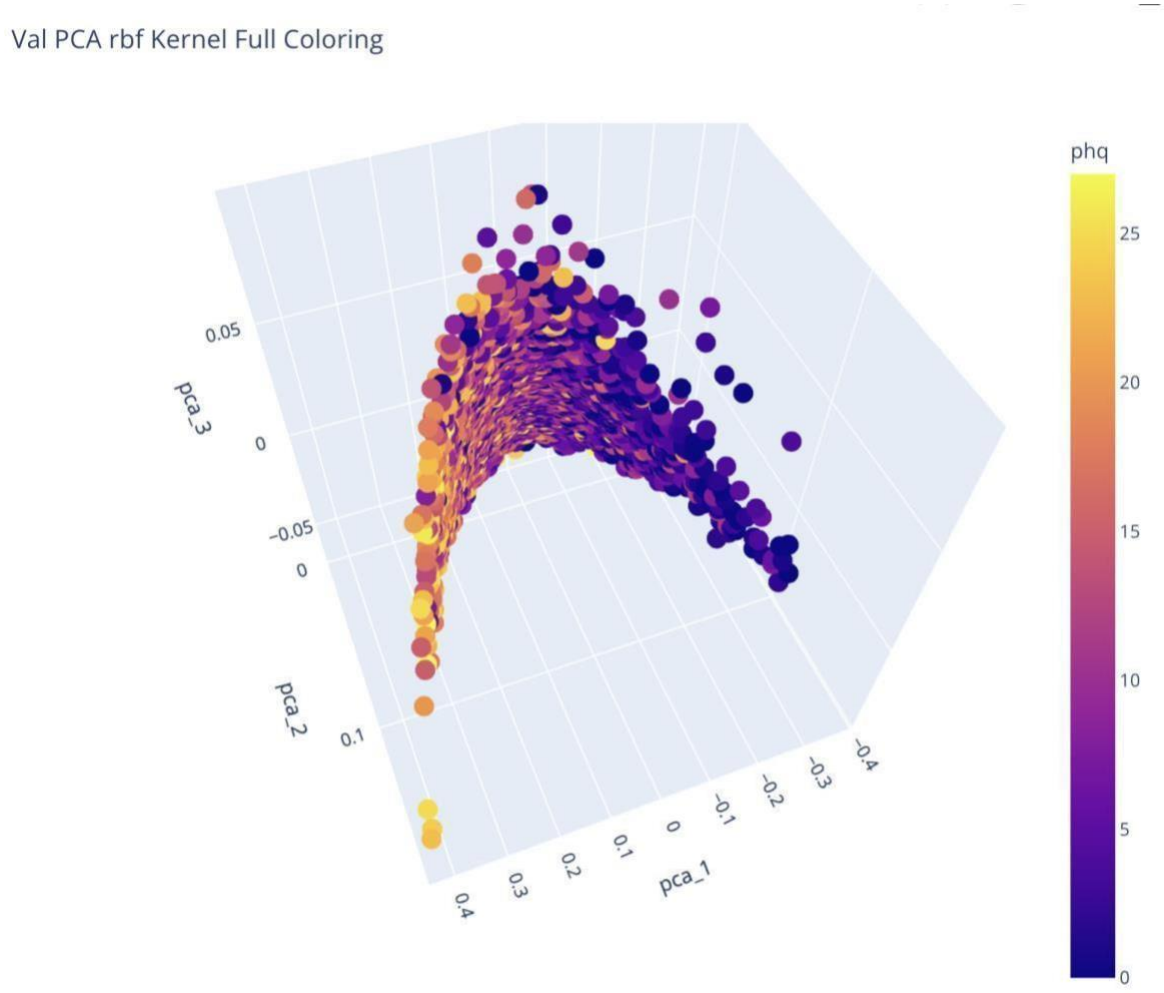
A feed-forward deep neural network composed of four algorithmic layers was trained in a supervised manner using PHQ-9 scores as labels to detect the presence of vocal characteristics consistent with a moderate to severe acute depressive episode in a single binary output. Training and validation datasets were split in a 70:30 ratio to ensure sufficient data were available in the validation set (**Figure 1 in main text**). The sets were balanced to include equal

proportions of demographic features and PHQ-9 scores without sample overlap to maintain validation fidelity. Training was performed in batches of 32 audio files introduced in a minibatch gradient descent to conserve memory and increase output speed, with a learning rate of 7.5e - 4 and using the AdamW optimizer.[5] Weights and biases were initialized randomly, and fine-tuning was performed between batches. Fifty epochs through all batches of training data were performed. The training objective was binary cross entropy weighted so that the depressed and nondepressed classes carried the same weight.

References

1. Yang, YY, Hira M, Ni Z, et al. TorchAudio: building blocks for audio and speech processing. arXiv:2110.15018. doi:10.48550/arXiv.2110.15018
2. Baevski A, Hsu WN, Xu Q, Babu A, Gu J, Auli M. data2vec: a general framework for self-supervised learning in speech, vision and language. arXiv:2202.03555. doi:10.48550/arXiv.2202.03555
3. Bredin H, Yin R, Coria JM, et al. pyannote.audio: neural building blocks for speaker diarization. arXiv:1911.01255. doi:10.48550/arXiv.1911.01255
4. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
5. Li M, Zhang T, Chen Y, Smola A. Efficient mini-batch training for stochastic optimization. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014: pp 661-670. doi: 10.1145/2623330.2623612

Supplemental Figure 1. The Model's Intermediate Feature Representation



PHQ-9 = Patient Health Questionnaire-9.

Note: The model's intermediate feature representation—the way the model encodes and numerically represents audio input projected into a 3-dimensional space—along with the corresponding PHQ-9 labels was established as a descriptive representation for the correlation between speaker audio and the magnitude of signs of depression. The correlation shows that the model has learned to extract useful information for predicting signs of depression from speech samples.