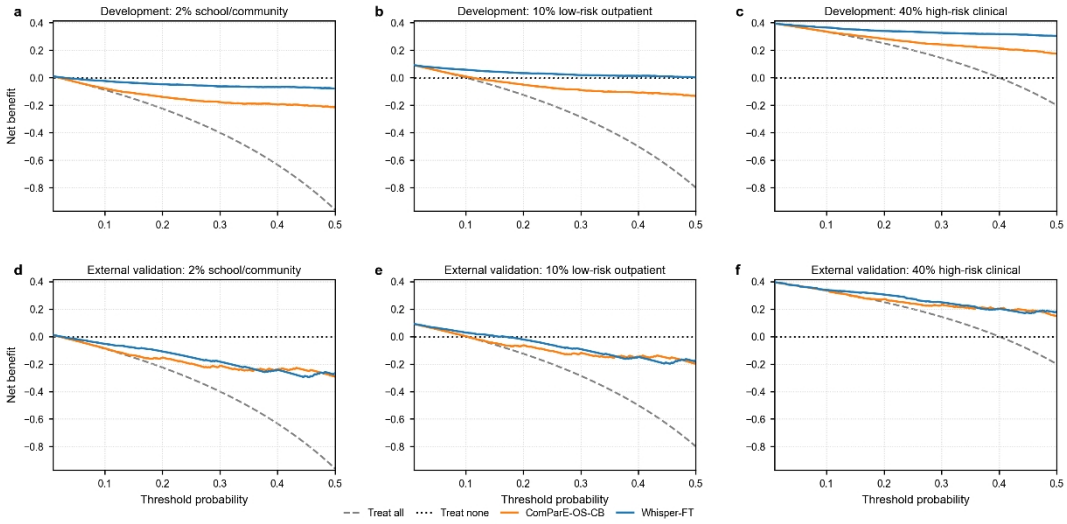
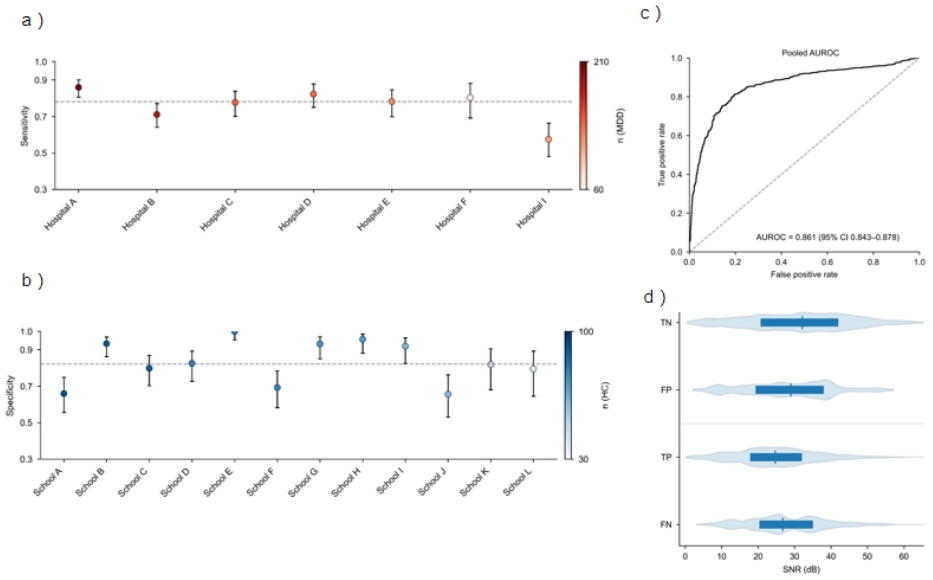
## Supplementary information

### Figure S1. Decision-curve analyses for the development and external cohorts.



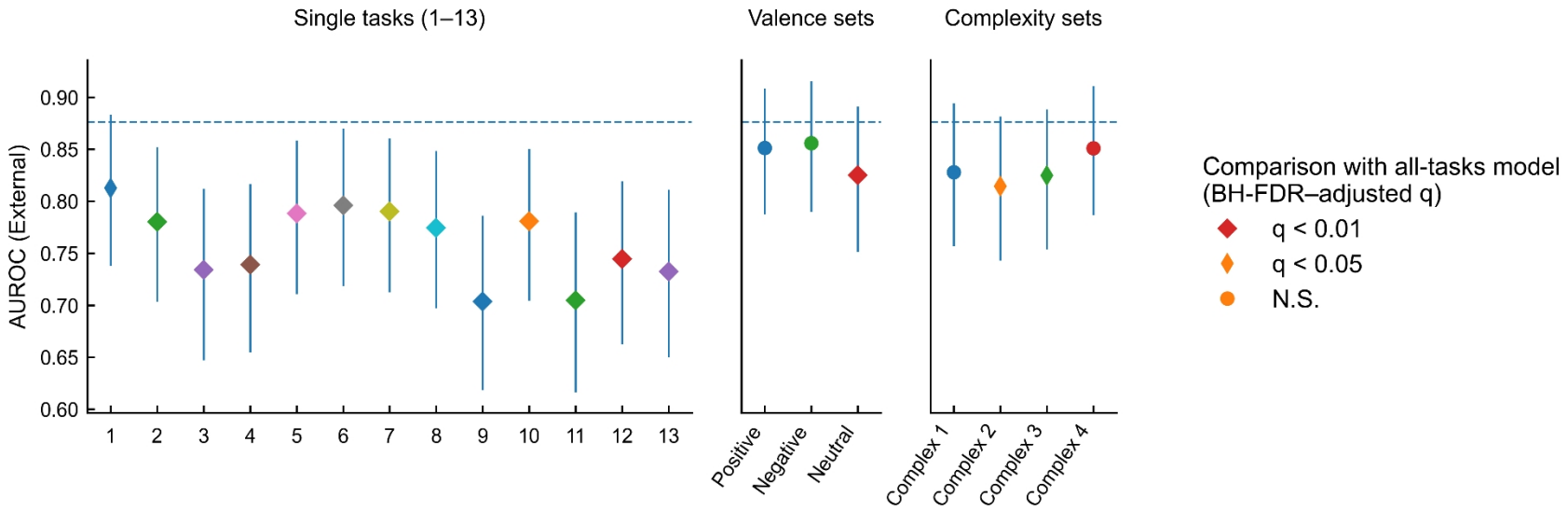

Note: Prevalence-adjusted decision curves show net benefit for Whisper-FT and ComParE-OS-CB under assumed MDD prevalences of 2%,10%,and40%. Panels a–c show the development cohort, and panels d–f show the external validation cohort. Dashed and dotted lines indicate treat-all and treat-none strategies.

Figure S2. Robustness to site and recording quality in the out-of-site evaluation.




Note: (A–B) Site-level sensitivity (hospitals) and specificity (schools) in the out-of-site evaluation; points are estimates with 95% bootstrap CIs; colour denotes site sample size, 2 sites with <20 participants were excluded; dashed lines indicate pooled operating levels. (C) Pooled ROC (AUROC with 95% bootstrap CI). (D) SNR (dB) distributions by classification outcome; violins show distributions and boxplots summarize median and interquartile range. Within-site SNR differences between errors and correct predictions were not significant, In site-adjusted logistic regression, SNR was not associated with misclassification (p=0.89).

**Figure S3. Task-level and task-set performance of Whisper-FT in external validation cohort.**




Note: Points indicate AUROC estimates and vertical bars show 95% confidence intervals for each single task (1–13), valence set (positive/negative/neutral), and complexity set (levels 1–4) evaluated on the external validation cohort. Each reduced task model was compared with the all-task Whisper-GE-FT reference model using a paired DeLong test. Benjamini-Hochberg FDR correction was applied within each predefined task-subset family: single tasks(m=13), valence sets(m=3), and complexity sets(m=4). Full statistics are provided in Table S13.

**Figure S4. Participant flow diagram.**

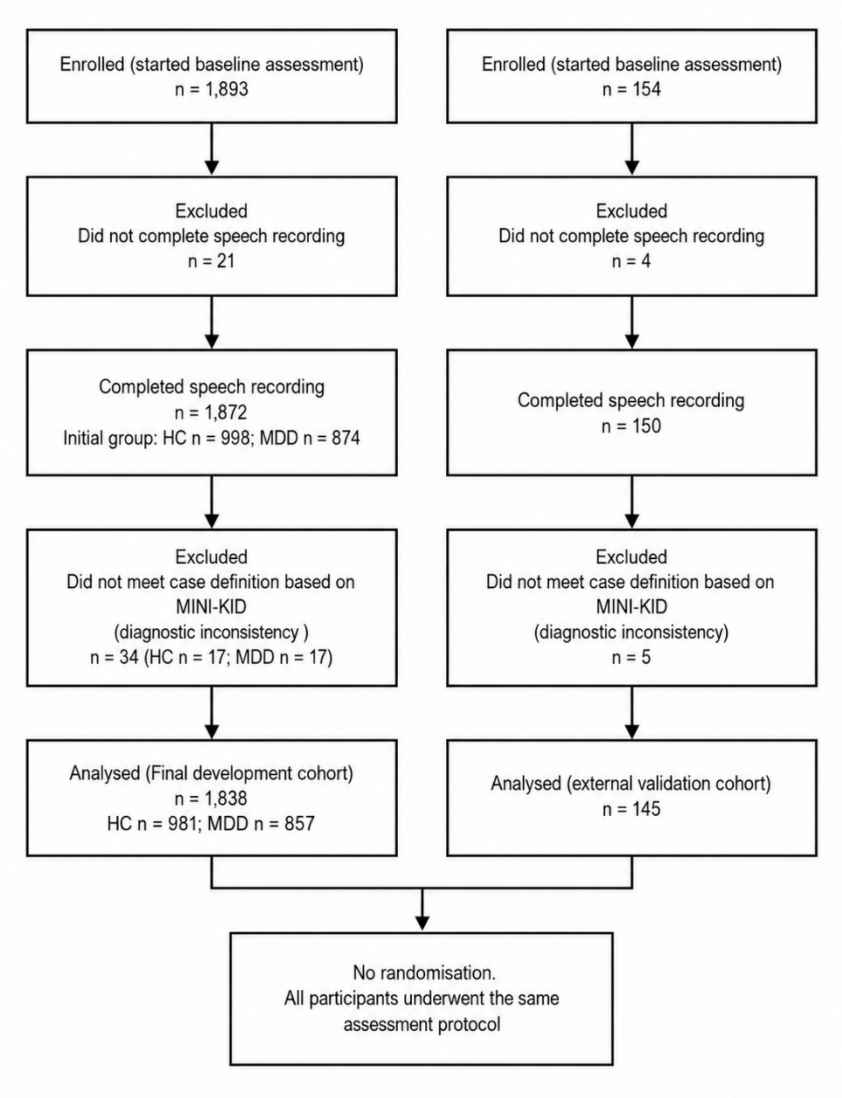


**Table S1. Baseline characteristics of the training and external validation cohorts.**

| Variable | Dev (n=1838) | Val(n=145) | ASMD |
|---|---|---|---|
| Age, Median [IQR] | 15.00 [14.00-16.00] | 15.00 [14.00-16.00] (n=145) | 0.16 |
| PHQ-8, Median [IQR] | 10.00 [3.00-17.00] | 8.00 [3.00-15.00] (n=145) | 0.12 |
| GAD-7, Median [IQR] | 6.00 [2.00-12.00] (n=1838) | 5.00 [1.00–11.00] (n=145) | 0.17 |
| Males(n%) | 682 (37.1%) | 98 (67.6%) | 0.61 |
| MDD proportion | 53.4% | 49.7% | 0.07 |

Abbreviations: IQR, interquartile range; ASMD, absolute standardized mean difference; PHQ-8, 8-item Patient Health Questionnaire; GAD-7, 7-item Generalized Anxiety Disorder scale; MDD, major depressive disorder; Dev, Development cohort; Val,external validation cohort; n, number of participants. ASMD > 0.2 was considered potentially meaningful.

**Table S2. Demographic, clinical , and recording-quality characteristics by diagnosis across cohorts.**

| Variable | Dev-MDD (n=981) | Dev-HC (n=857) | ASMD (in Train) | Val-MDD (n=72) | Val-HC (n=73) | ASMD (in Val) |
|---|---|---|---|---|---|---|
| Age, Median [IQR] | 15.00 [14.00-16.00] | 14.00 [13.00-16.00] | 0.3 | 15.00 [14.00-16.00] | 15.00 [14.00–16.00] | 0.23 |
| PHQ-8, Median [IQR] | 17.00 [13.00-20.00] | 2.00 [0.00–5.00] | 2.99 | 15.00 [12.00-18.00] | 4.00 [0.00–6.00] | 2.71 |
| GAD-7, Median [IQR] | 12.00 [7.00-16.00] | 2.00 [0.00–5.00] | 1.73 | 11.00 [6.00-15.00] | 1.00 [0.00–4.00] | 1.55 |
| Recording SNR Median [IQR] | 25.32 [18.13–33.15] | 31.73 [20.54–41.04] | 0.40 | 25.73 [21.19–33.97] | 30.45 [19.95–36.82] | 0.17 |
| Males(n%) | 284 (29.0%) | 398 (46.4%) | 0.36 | 57 (79.2%) | 41 (56.2%) | 0.49 |

Abbreviations: IQR, interquartile range; ASMD, absolute standardized mean difference; PHQ-9, 9-item Patient Health Questionnaire; PHQ-8, 8-item Patient Health Questionnaire; GAD-7, 7-item Generalized Anxiety Disorder scale; SNR: Signal-to-Noise Ratio; MDD, major depressive disorder; Dev, Development cohort; Val,external validation cohort; n, number of participants. ASMD > 0.2 was considered potentially meaningful.

**Table S3. Distribution of comorbidities in MDD**

| Comorbidity | Dev-MDD (n=981) | Val-MDD (n=72) |
|---|---|---|
| Anxiety Disorder | 961 (98.0%) | 66 (91.7%) |
| OCD | 152 (15.5%) | 3 (4.2%) |
| ODD | 112 (11.4%) | 8 (11.1%) |
| Other | 164 (16.7%) | 15 (20.8%) |

Abbreviations: Dev, development set; Val,external validation set; n, number of participants. OCD,obsessive-compulsive disorder;ODD,oppositional defiant disorder.

Table S4. DeLong comparison of AUCs between Whisper-FT and ComParE-OS-CB.

| Cohort | AUROC Whisper-FT | AUROC ComParE-OS-CB | ΔAUROC | Z | p |
|---|---|---|---|---|---|
| Development | 0.955 | 0.844 | 0.111 | 12.97 | <0.001 |
| External validation | 0.876 | 0.825 | 0.052 | 2.05 | 0.041 |

**Table S5. High-effect acoustic features (development vs external).**

| Family | Development set | External validation set |
|---|---|---|
| Spectral | pcm_fftMag_spectralSlope_sma_leftctime<br>audSpec_Rfilt_sma_de[4]_quartile1<br>audSpec_Rfilt_sma_de[5]_quartile1<br>audSpec_Rfilt_sma_de[19]_quartile1<br>audSpec_Rfilt_sma_de[20]_quartile1<br>audSpec_Rfilt_sma_de[21]_quartile1<br>audSpec_Rfilt_sma_de[24]_quartile1<br>audSpec_Rfilt_sma_de[25]_quartile1<br>pcm_fftMag_spectralRollOff25.0_sma_de_quartile1<br>pcm_fftMag_spectralRollOff50.0_sma_de_quartile1 | audSpec_Rfilt_sma_de[4]_quartile3<br>audSpec_Rfilt_sma_de[4]_iqr2-3<br>audSpec_Rfilt_sma_de[5]_quartile3<br>audSpec_Rfilt_sma_de[25]_iqr2-3 |
| Cepstral | mfcc_sma[2]_lpc0<br>mfcc_sma[3]_lpc0<br>mfcc_sma[4]_lpc0<br>mfcc_sma[6]_lpc0<br>mfcc_sma[8]_lpc0<br>mfcc_sma_de[1]_quartile1<br>mfcc_sma_de[2]_quartile1<br>mfcc_sma_de[3]_quartile1<br>mfcc_sma_de[3]_lpc0<br>mfcc_sma_de[4]_quartile1<br>mfcc_sma_de[4]_lpc0<br>mfcc_sma_de[5]_quartile1<br>mfcc_sma_de[6]_quartile1<br>mfcc_sma_de[6]_lpc0<br>mfcc_sma_de[8]_lpc0<br>mfcc_sma_de[9]_lpc0 | mfcc_sma_de[4]_percentile99.0<br>mfcc_sma_de[5]_quartile1<br>mfcc_sma_de[6]_quartile1<br>mfcc_sma[4]_meanRisingSlope<br>mfcc_sma[4]_stddevRisingSlope<br>mfcc_sma[5]_meanRisingSlope<br>mfcc_sma_de[3]_peakRangeAbs<br>mfcc_sma_de[4]_peakRangeAbs<br>mfcc_sma_de[4]_meanFallingSlope |

Note: High-effect acoustic features from the ComParE-2016 set identified by univariate effect-size screening in the development cohort and replicated in the external validation cohort. Features are grouped by acoustic family (spectral vs. cepstral), and entries list those with the largest effects in each dataset.

**Table S6. Discrimination performance of PHQ-8 (≥10) and the fusion model (speech+PHQ-8).**

| Model | Threshold | AUROC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | F1 (95% CI) | Accuracy (95% CI) |
|---|---|---|---|---|---|---|
| PHQ-8 (internal) | 10 | 0.92 (0.90–0.93) | 0.89 (0.87–0.91) | 0.94 (0.93–0.96) | 0.92 (0.90–0.93) | 0.91 (0.90–0.93) |
| PHQ-8 (external) | | 0.93 (0.89–0.97) | 0.89 (0.81–0.96) | 0.97 (0.93–1.00) | 0.93 (0.88–0.97) | 0.93 (0.89–0.97) |
| Fusion (internal) | 0.546 | 0.99 (0.98–0.99) | 0.95 (0.94–0.97) | 0.96 (0.95–0.97) | 0.96 (0.95–0.97) | 0.96 (0.95–0.97) |
| Fusion (external) | | 0.97 (0.94–0.99) | 0.92 (0.85–0.97) | 0.88 (0.79–0.95) | 0.90 (0.85–0.94) | 0.90 (0.85–0.94) |

Note: Thresholds correspond to the Youden-optimal cut-point in the development cohort ($\tau = 0.546$ for Fusion), and the same threshold was applied unchanged to the external validation cohort. Values are shown as point estimates with 95% confidence intervals.

**Table S7. Prevalence-adjusted predictive values of PHQ-8, Whisper-FT, and fusion models across prevalence scenarios.**

| Cohort | Assumed prevalence | Clinical context | Model | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) |
|---|---|---|---|---|---|---|---|
| Development | 2% | Community/school screening | PHQ-8 | 0.89 (0.87–0.91) | 0.94 (0.92–0.96) | 0.24 (0.19–0.30) | 1.00 (1.00–1.00) |
| | | | Whisper-FT | 0.91 (0.89–0.92) | 0.91 (0.89–0.92) | 0.16 (0.14–0.20) | 1.00 (1.00–1.00) |
| | | | Fusion | 0.95 (0.94–0.97) | 0.96 (0.95–0.97) | 0.33 (0.27–0.42) | 1.00 (1.00–1.00) |
| | 10% | Low-risk outpatient screening | PHQ-8 | 0.89 (0.87–0.91) | 0.94 (0.92–0.96) | 0.63 (0.57–0.70) | 0.99 (0.98–0.99) |
| | | | Whisper-FT | 0.91 (0.89–0.92) | 0.91 (0.89–0.92) | 0.52 (0.47–0.57) | 0.99 (0.99–0.99) |
| | | | Fusion | 0.95 (0.94–0.97) | 0.96 (0.95–0.97) | 0.73 (0.67–0.80) | 0.99 (0.99–1.00) |
| | 40% | High-risk psychiatric setting | PHQ-8 | 0.89 (0.87–0.91) | 0.94 (0.92–0.96) | 0.91 (0.89–0.93) | 0.93 (0.92–0.94) |
| | | | Whisper-FT | 0.91 (0.89–0.92) | 0.91 (0.89–0.92) | 0.86 (0.84–0.89) | 0.93 (0.92–0.95) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | Fusion | 0.95 (0.94–0.97) | 0.96 (0.95–0.97) | 0.94 (0.92–0.96) | 0.97 (0.96–0.98) |
| Validation | 2% | Community/school screening | PHQ-8 | 0.89 (0.81–0.96) | 0.97 (0.93–1.00) | 0.40 (0.21–1.00) | 1.00 (1.00–1.00) |
| | | | Whisper-FT | 0.83 (0.74–0.91) | 0.79 (0.70–0.88) | 0.08 (0.05–0.13) | 1.00 (0.99–1.00) |
| | | | Fusion | 0.92 (0.85–0.97) | 0.88 (0.80–0.94) | 0.13 (0.08–0.25) | 1.00 (1.00–1.00) |
| | 10% | Low-risk outpatient screening | PHQ-8 | 0.89 (0.81–0.96) | 0.97 (0.93–1.00) | 0.78 (0.59–1.00) | 0.99 (0.98–1.00) |
| | | | Whisper-FT | 0.83 (0.74–0.91) | 0.79 (0.70–0.88) | 0.31 (0.23–0.44) | 0.98 (0.96–0.99) |
| | | | Fusion | 0.92 (0.85–0.97) | 0.88 (0.80–0.94) | 0.45 (0.33–0.64) | 0.99 (0.98–1.00) |
| | 40% | High-risk psychiatric setting | PHQ-8 | 0.89 (0.81–0.96) | 0.97 (0.93–1.00) | 0.96 (0.89–1.00) | 0.93 (0.88–0.97) |
| | | | Whisper-FT | 0.83 (0.74–0.91) | 0.79 (0.70–0.88) | 0.73 (0.64–0.83) | 0.88 (0.82–0.93) |
| | | | Fusion | 0.92 (0.85–0.97) | 0.88 (0.80–0.94) | 0.83 (0.75– | 0.94 (0.90– |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | 0.92) | 0.98) |

**Table S8. Subgroup performance of Whisper-FT and ComParE-OS-CB models .**

| Stratifier | Subgroup | N | Whisper Sens (95% CI) | ComParE Sens (95% CI) | Whisper FPR (95% CI) | ComParE FPR (95% CI) |
|---|---|---|---|---|---|---|
| Age | 13–14years | 859 | 0.896 (0.867–0.926) | 0.758 (0.716–0.800) | 0.079 (0.055–0.106) | 0.156 (0.126–0.189) |
| | 15–18years | 979 | 0.910 (0.885–0.932) | 0.734 (0.700–0.769) | 0.112 (0.082–0.144) | 0.203 (0.164–0.243) |
| Sex | Male | 682 | 0.859 (0.817–0.898) | 0.687 (0.634–0.739) | 0.068 (0.045–0.093) | 0.138 (0.106–0.171) |
| | Female | 1156 | 0.923 (0.902–0.941) | 0.768 (0.737–0.799) | 0.118 (0.089–0.148) | 0.214 (0.176–0.251) |
| Symptom severity | Low dep/low anx | 878 | 0.833 (0.756–0.900) | 0.689 (0.589–0.778) | 0.090 (0.071–0.110) | 0.178 (0.151–0.204) |
| | High dep/low anx | 316 | 0.915 (0.880–0.947) | 0.756 (0.707–0.806) | 0.152 (0.030–0.273) | 0.152 (0.030–0.273) |
| | High dep/high anx | 607 | 0.915 (0.892–0.937) | 0.751 (0.715–0.785) | 0.118 (0.000–0.294) | 0.235 (0.059–0.471) |

Sensitivity and false positive rate (FPR) for Whisper-FT and ComParE-OS-CB across age, sex and symptom-severity subgroups at fixed Youden-optimal thresholds. Values are point estimates with 95% bootstrap confidence intervals (5,000 resamples). Symptom severity strata were defined using PHQ-8 and GAD-7: low dep/low anx (PHQ-8 < 10 and GAD-7 < 10), high dep/low anx (PHQ-8 ⩾ 10 and GAD-7 < 10), high dep/high anx (PHQ-8 ⩾ 10 and GAD-7 ⩾ 10).

**Table S9. Subgroup disparities for Whisper-FT versus ComParE-OS-CB.**

| Stratifier | Contrast | Whispe ΔSens (95% CI) | ComParE ΔSens (95% CI) | ΔΔSens (Whisper−ComParE) (95% CI) | Whispe ΔFPR (95% CI) | ComParE ΔFPR (95% CI) | ΔΔFPR (Whisper−ComParE) (95% CI) |
|---|---|---|---|---|---|---|---|
| Age | Older-younger | 0.013 (-0.026–0.051) | -0.024 (-0.079–0.031) | 0.037 (-0.017–0.090) | 0.032 (-0.006–0.071) | 0.047 (-0.004–0.098) | -0.015 (-0.070–0.038) |
| Sex | Female- Male | 0.063* (0.020–0.109) | 0.081 (0.021–0.143) | -0.018 (-0.081–0.041) | 0.050 (0.011–0.089) | 0.075 (0.023–0.126) | -0.026 (-0.079–0.029) |
| Symptom severity | High dep/low anx − Low dep/low anx | 0.082 (0.002–0.170) | 0.067 (-0.035–0.178) | 0.015 (-0.085–0.109) | 0.061 (-0.052–0.195) | -0.026 (-0.141–0.103) | 0.088 (-0.107–0.272) |
| | High dep/high anx −Low dep/low anx | 0.082 (0.007–0.165) | 0.062 (-0.038–0.168) | 0.020 (-0.076–0.110) | 0.028 (-0.099–0.203) | 0.058 (-0.130–0.280) | -0.030 (-0.208–0.098) |
| | High dep/high anx −High dep/low anx | 0.000 (-0.038–0.040) | -0.005 (-0.065–0.058) | 0.005 (-0.056–0.062) | -0.034 (-0.214–0.173) | 0.084 (-0.127–0.321) | -0.118 (-0.360–0.121) |

Note: Subgroup disparities in sensitivity (ΔSens) and false positive rate (ΔFPR) for Whisper-FT and ComParE-OS-CB at fixed decision thresholds. Each contrast is defined as subgroup minus reference: older − younger (15–18 vs 13–14 years),ΔΔSens and ΔΔFPR denote the between-model differences in disparities (Whisper-FT disparity − ComParE-OS-CB disparity). Values are shown as point estimates with 95% bootstrap confidence intervals. Two-sided p values for Δ and ΔΔ were obtained by non-parametric bootstrap and adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) procedure. * indicates $q < 0.05$.

**Table S10. Associations between calibrated Whisper-FT predicted risk and symptom severity within adolescents with MDD.**

| Cohort | Severity measure | n | Spearman ρ (95% CI) | p |
|---|---|---|---|---|
| Development | PHQ-8 | 981 | 0.139 (0.077, 0.197) | <0.01 |
| External validation | | 72 | 0.084 (-0.155, 0.318) | 0.483 |
| Development | GAD-7 | 981 | -0.006 (-0.066, 0.055) | 0.854 |
| External validation | | 72 | 0.015 (-0.228, 0.252) | 0.898 |

Note: Spearman rank correlations (ρ) were computed between the Platt-scaled Whisper-FT predicted probability (y_platt) and continuous symptom-severity measures (PHQ-8 total score and GAD-7 total score) within MDD cases only. Values are point estimates with 95% bootstrap confidence intervals. Two-sided p values were obtained from Spearman correlation tests.

**Table S11. Overall performance of Whisper-FT and Whisper-GE-FT at the Youden-optimal threshold.**

| Model | AUROC (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | F1 (95% CI) | Accuracy (95% CI) |
|---|---|---|---|---|---|---|---|
| Whisper-FT (internal) | 0.95 (0.95–0.96) | 0.90 (0.89–0.92) | 0.91 (0.89–0.92) | 0.92 (0.90–0.93) | 0.89 (0.87–0.91) | 0.91 (0.90–0.92) | 0.90 (0.89–0.92) |
| Whisper-GE-FT (internal) | 0.96 (0.95–0.97) | 0.91 (0.89–0.93) | 0.92 (0.90–0.94) | 0.93 (0.91–0.94) | 0.90 (0.88–0.92) | 0.92 (0.91–0.93) | 0.91 (0.90–0.93) |
| Whisper-FT (external) | 0.88 (0.82–0.93) | 0.83 (0.75–0.92) | 0.79 (0.70–0.88) | 0.80 (0.73–0.88) | 0.83 (0.76–0.91) | 0.82 (0.75–0.88) | 0.81 (0.75–0.88) |
| Whisper-GE-FT (external) | 0.87 (0.81–0.93) | 0.81 (0.71–0.89) | 0.81 (0.71–0.90) | 0.81 (0.71–0.89) | 0.81 (0.72–0.89) | 0.81 (0.73–0.87) | 0.81 (0.74–0.87) |

Note: Thresholds correspond to the Youden-optimal cut-point in the development cohort based on Platt-calibrated probabilities, and the same thresholds were applied unchanged to the external validation cohort. Values are shown as point estimates with 95% confidence intervals.

**Table S12. Sex-specific performance and sex gaps (M–F) for Whisper-FT and Whisper-GE-FT.**

| Metric | Whisper-FT F (95% CI) | Whisper-FT M (95% CI) | Whisper-GE-FT F (95% CI) | Whisper-GE-FT M (95% CI) | Whisper-FT gap M–F (95% CI) | Whisper-GE-FT gap M–F (95% CI) |
|---|---|---|---|---|---|---|
| AUC | 0.953 (0.941–0.964) | 0.954 (0.939–0.968) | 0.958 (0.947–0.968) | 0.959 (0.945–0.973) | 0.001 (-0.018–0.020) | 0.002 (-0.016–0.019) |
| Sens | 0.923 (0.902–0.941) | 0.863 (0.820–0.901) | 0.921 (0.901–0.941) | 0.884 (0.849–0.919) | -0.060 (-0.105–-0.017) | -0.037 (-0.080–0.004) |
| Spec | 0.882 (0.852–0.911) | 0.932 (0.907–0.955) | 0.906 (0.880–0.932) | 0.935 (0.910–0.957) | 0.050 (0.011–0.089) | 0.028 (-0.008–0.064) |
| FPR | 0.118 (0.089–0.148) | 0.068 (0.045–0.093) | 0.094 (0.068–0.120) | 0.065 (0.043–0.090) | -0.050 (-0.089–-0.011) | -0.028 (-0.064–0.008) |
| PPV | 0.923 (0.904–0.940) | 0.901 (0.868–0.933) | 0.937 (0.921–0.954) | 0.906 (0.874–0.937) | -0.022 (-0.059–0.015) | -0.031 (-0.067–0.004) |
| minSens | 0.863 (0.820–0.901) | | 0.884 (0.849–0.915) | | | |

Note: Performance metrics for Whisper-FT and the gender-embedding variant Whisper-GE-FT are reported separately for female and male adolescents at the Youden-optimal thresholds, together with male–female gaps (M–F) and 95% bootstrap confidence intervals (CIs). Negative gaps indicate better performance in females, whereas positive gaps indicate better performance in males. Δgap quantifies the change in gap magnitude when moving from Whisper-FT to Whisper-GE-FT.

**Table S13. DeLong result for task-subset comparisons.**

| Family | Task subset | AUROC | ΔAUROC vs all-task | Z | p | BH-FDR q |
|---|---|---|---|---|---|---|
| Single task | Task-1 | 0.813 | -0.064 | -2.26 | 0.024 | 0.024 |
| | Task-2 | 0.780 | -0.096 | -3.07 | 0.002 | 0.003 |
| | Task-3 | 0.734 | -0.142 | -4.14 | <0.001 | <0.001 |
| | Task-4 | 0.739 | -0.137 | -4.31 | <0.001 | <0.001 |
| | Task-5 | 0.788 | -0.088 | -2.66 | 0.008 | 0.008 |
| | Task-6 | 0.796 | -0.080 | -2.74 | 0.006 | 0.007 |
| | Task-7 | 0.790 | -0.086 | -2.88 | 0.004 | 0.005 |
| | Task-8 | 0.774 | -0.102 | -3.54 | <0.001 | <0.001 |
| | Task-9 | 0.704 | -0.173 | -5.38 | <0.001 | <0.001 |
| | Task-10 | 0.781 | -0.096 | -3.03 | 0.002 | 0.004 |
| | Task-11 | 0.705 | -0.172 | -5.39 | <0.001 | <0.001 |
| | Task-12 | 0.744 | -0.132 | -3.89 | <0.001 | <0.001 |
| | Task-13 | 0.732 | -0.144 | -4.64 | <0.001 | <0.001 |
| Valence set | Positive | 0.851 | -0.025 | -1.86 | 0.064 | 0.095 |
| | Neutral | 0.825 | -0.051 | -3.03 | 0.002 | 0.007 |
| | Negative | 0.856 | -0.020 | -1.12 | 0.264 | 0.264 |
| Complexity set | Complexity level 1 | 0.828 | -0.048 | -1.99 | 0.047 | 0.063 |
| | Complexity level 2 | 0.814 | -0.062 | -2.64 | 0.008 | 0.033 |
| | Complexity level 3 | 0.825 | -0.052 | -2.26 | 0.024 | 0.048 |
| | Complexity level 4 | 0.851 | -0.025 | -1.74 | 0.082 | 0.082 |

**Table S14. Summary of sample sizes by sites (anonymized).**

A. Development cohort

| Setting | Site | Group | n |
|---|---|---|---|
| Hospital | Hospital A | MDD | 206 |
| Hospital | Hospital B | MDD | 183 |
| Hospital | Hospital C | MDD | 139 |
| Hospital | Hospital D | MDD | 135 |
| Hospital | Hospital E | MDD | 119 |
| Hospital | Hospital F | MDD | 66 |
| Hospital | Hospital G | MDD | 18 |

| | | | |
|---|---|---|---|
| Hospital | Hospital H | MDD | 7 |
| Hospital | Hospital I | MDD | 108 |
| School | School A | HC | 91 |
| School | School B | HC | 90 |
| School | School C | HC | 89 |
| School | School D | HC | 80 |
| School | School E | HC | 80 |
| School | School F | HC | 78 |
| School | School G | HC | 73 |
| School | School H | HC | 70 |
| School | School I | HC | 62 |
| School | School J | HC | 61 |
| School | School K | HC | 44 |
| School | School L | HC | 39 |

B. External validation cohort

| Site | MDD | HC | Total |
|---|---|---|---|
| Hospital A | 9 | 10 | 19 |
| Hospital B | 2 | 10 | 12 |
| Hospital C | 9 | 10 | 19 |
| Hospital E | 31 | 25 | 56 |
| Hospital F | 1 | 0 | 1 |
| Hospital G | 10 | 10 | 20 |
| Hospital I | 10 | 8 | 18 |

Note: To protect privacy and comply with data-use agreements, sites are anonymized (Hospital A–I; School A–L) and specific institutional names re not disclosed.

**Table S15. Recording settings in the external validation cohort.**

| Recording setting | MDD (n=72) | HC (n=73) |
|---|---|---|
| Home | 43 (59.7%) | 57 (78.1%) |
| Clinic | 23 (31.9%) | 11 (15.1%) |
| Ward | 5 (6.9%) | 0 (0.0%) |
| Other | 1 (1.4%) | 5 (6.8%) |

**Table S16. Speech task prompts, valence, and complexity levels**

| Item | Task type | Prompt | Valence | Complexity level |
|---|---|---|---|---|
| Item 0 | Environment noise (10 s) | To assess background noise in your current recording environment, please record 10 seconds of ambient sound. Do not speak during the recording and remain quiet. | N/A | N/A |
| Item 1 | Counting | Read the numbers from 1 to 100 quickly and continuously in order. | Neutral | Level 1 |
| Item 2 | Syllable repetition | Repeat the syllable “la” quickly and continuously. | Neutral | Level 1 |
| Item 3 | Image description | Please carefully observe the expressions of the people in the picture, and describe their mood and possible emotions in two or three sentences. (Happiness) | Positive | Level 2 |
| Item 4 | Image description | Please carefully observe the expressions of the people in the picture, and describe their mood and possible emotions in two or three sentences. (Neutral) | Neutral | Level 2 |
| Item 5 | Image description | Please carefully observe the expressions of the people in the picture, and describe their mood and possible emotions in two or three sentences. (Sadness) | Negative | Level 2 |
| Item 6 | Sentence reading | Read the sentence aloud: This book is a comprehensive introduction and survey text. It covers the latest developments across all major topics and areas, with more than 400 references. It is suitable for students, researchers, and practitioners. Instructors can conveniently use it in class to teach courses such as natural language processing, social media analytics, text mining, and data mining. Lecture slides are also available online. | Neutral | Level 3 |
| Item 7 | Sentence reading | Read the sentence aloud: Director Frank Darabont could be remembered in history for this enduring work alone. Its powerful message, conveyed through a restrained pace, inspires our growth and will continue to be passed on. Within the film’s confined space, a grand vision of freedom and hope in adversity unfolds; it may not be the most dazzling film, but when you feel lost in life, it is powerful enough to awaken clarity and courage. | Neutral | Level 3 |
| Item 8 | Sentence reading | Read the sentence aloud: The press conference on the Xiaoxian traffic accident has concluded, confirming 23 deaths at the scene. Two truck drivers died on site, and 21 passengers on the coach died on site. We later learned that among the three injured people, one died despite rescue efforts. The total number of deaths therefore rose to 24. | Negative | Level 3 |

| | | | | |
|---|---|---|---|---|
| Item 9 | Free speech (description) | Please describe what the weather is like outside. | Neutral | Level 4 |
| Item 10 | Free speech (positive recall) | Please talk about a recent experience that made you happy. | Positive | Level 4 |
| Item 11 | Free speech (preference) | Please describe the person you like the most and explain why you like them. | Positive | Level 4 |
| Item 12 | Free speech (sad recall) | Please talk about a recent experience that made you sad. | Negative | Level 4 |
| Item 13 | Free speech (dislike) | Please describe the person you dislike the most and explain why you dislike them. | Negative | Level 4 |

# CONSORT 2025

| Section | No. | Checklist item | Suggested page(s) | Notes |
|---|---|---|---|---|
| **Title and abstract** | 1a | Identification as a randomised trial | N/A | Non-randomised prospective observational study |
| Title and abstract | 1b | Structured summary of the design, methods, results, and conclusions | 2 | Title p1 Abstract p2 |
| **Open science** | 2 | Trial registration | 2, 16 | Reported in the abstract and in Methods > Study design and oversight. |
| Open science | 3 | Protocol and statistical analysis plan access | 16 | Registry URL is reported on p16; full SAP accessible |
| Open science | 4 | Data sharing / code / materials access | 32 | Data availability and code availability statements on p32. |
| **Funding and conflicts of interest** | 5a | Funding source and role of funders | 31 | Acknowledgements |
| Funding and conflicts of interest | 5b | Conflicts of interest | 31 | Competing interests statement. |
| **Introduction** | 6 | Scientific background and rationale | 3-4 | Introduction. |
| Introduction | 7 | Specific objectives | 4 | Last paragraph of Introduction. |
| **Methods** | 8 | Patient and public involvement | NR | No explicit patient/public involvement |
| Methods | 9 | Trial design | 16-17 | Study design and oversight; the study is observational, not randomized. |
| Methods | 10 | Changes to protocol after commencement | 17 | No changes to eligibility criteria, outcomes or analysis plan after commencement. |
| Methods | 11 | Trial setting and locations | 16-18 | Methods: Study design and oversight / Study Design. |
| Methods | 12a | Eligibility criteria for participants | 17 | Methods: Eligibility criteria. |
| Methods | 12b | Eligibility criteria for sites / individuals delivering interventions | 17-18 | No intervention provider. |

| Section | No. | Checklist item | Suggested page(s) | Notes |
|---|---|---|---|---|
| Methods | 13 | Intervention and comparator details | N/A | |
| Methods | 14 | Primary and secondary outcomes | 20-21 | Methods: Evaluation Metrics and Statistical Analysis. |
| Methods | 15 | Harms | N/A | No important harms or unintended effects were anticipated or observed. As this was an observational study involving speech recordings and questionnaire assessments only, participant risk was considered minimal. |
| Methods | 16a | Sample size determination | 18 | Methods: Study Design. |
| Methods | 16b | Interim analyses and stopping guidelines | N/A | Not applicable for this observational study. |
| **Randomisation** | 17a | Sequence generation | N/A | Not applicable. |
| Randomisation | 17b | Type of randomisation / restriction | N/A | Not applicable. |
| Randomisation | 18 | Allocation concealment mechanism | N/A | Not applicable. |
| Randomisation | 19 | Implementation | N/A | Not applicable. |
| Randomisation | 20a | Blinding after assignment | N/A | Not applicable. |
| Randomisation | 20b | How blinding was achieved | N/A | Not applicable. |
| **Statistical methods** | 21a | Statistical methods for primary/secondary outcomes | 20-21 | Methods: Evaluation Metrics and Statistical Analysis. |
| Statistical methods | 21b | Who was included in each analysis, and in which group | 18-20 | Study Design and Model Development. |
| Statistical methods | 21c | How missing data were handled | 18 | Incomplete speech batteries were excluded; no imputation. |
| Statistical methods | 21d | Additional analyses (subgroup / sensitivity) | 10-11, 20-21 | Result and Methods |
| **Results** | 22a | Participant flow | 18 , Fig. S4 | |
| Results | 22b | Losses and exclusions after randomisation | N/A | Not applicable. |

| Section | No. | Checklist item | Suggested page(s) | Notes |
|---|---|---|---|---|
| Results | 23a | Recruitment and follow-up dates | 18 | Recruitment dates reported; longitudinal follow-up is not applicable in this cross-sectional analysis. |
| Results | 23b | Why the trial ended or was stopped | N/A | Not applicable. |
| Results | 24a | Intervention/comparator as actually administered | N/A | |
| Results | 24b | Concomitant care | N/A | Not applicable. No trial intervention or comparator treatment was administered. |
| Results | 25 | Baseline demographic and clinical characteristics | 5; Tables S1-S2 | |
| Results | 26 | Numbers analysed, outcomes and estimation | 5-10 | Main results text and Tables 1-2. |
| Results | 27 | Harms / unintended events | N/A | No important harms or unintended effects were identified. |
| Results | 28 | Ancillary analyses | 10-11; 20-21 | Robustness, subgroup, fairness and task-set analyses. |
| **Discussion** | 29 | Interpretation | 12-15 | Discussion. |
| Discussion | 30 | Limitations | 14-15 | Discussion |