Scientia Psychiatrica
Vol. 7, No. 4, 2026, pp. 213–225 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Participant-Wise Validation of Transcript-Only Models for Adolescent Depression and Anxiety Triage in MMPsy

Winata Putri[1], Mischa Chantal Adella[2], Agnes Mariska[3], Husin Sastranagara[4*]

[1]*Department of Pediatrics, CMHC Research Center, Palembang, Indonesia*
[2]*Department of Forensic Medicine and Medicolegal, Phlox Institute, Palembang, Indonesia*
[3]*Department of Internal Medicine, Jayawijaya Private Clinic, Jayapura, Indonesia*
[4]*Department of Neurology, CMHC Research Center, Palembang, Indonesia*

ARTICLE INFO

Article history
Received: August 22, 2026
Revised: September 20, 2026
Accepted: October 16, 2026
Available online: October 25, 2026
Published: October 30, 2026

Keywords
Adolescents
Anxiety
Calibration
Depression
Natural language processing
Prediction model

***Corresponding author**
Husin Sastranagara
E-mail:
husin.sastranagara@cattleyacenter.id
ORCID: 0009-0006-7680-5608

DOI
10.37275/scipsy.v7i4.223

ABSTRACT

Background. Digital interviews may offer scalable signals for identifying adolescents needing further assessment, but language models can appear accurate because of participant leakage, response length, or tuning-data evaluation.

Objective. To develop and internally validate transcript-only probabilistic models of Patient Health Questionnaire-9 (PHQ-9) and Generalized Anxiety Disorder-7 (GAD-7) screening positivity in the Multi-Modal Psychological Assessment (MMPsy) corpus and quantify incremental performance beyond response length.

Methods. This secondary prediction-modeling study used 1,275 participants with Mandarin transcripts and complete PHQ-9 and GAD-7 scores. Transcripts were represented by term frequency-inverse document frequency character 2–4-grams and modeled with elastic-net logistic regression. Five-fold outer cross-validation estimated performance in unseen participants; four-fold inner cross-validation selected penalization, fitted Platt calibration, and set fold-specific thresholds targeting 0.80 sensitivity. Comparators were response length and training-fold prevalence. Intervals used 2,000 participant bootstraps.

Results. Screening positivity occurred in 258 (20.2%) PHQ-9 and 136 (10.7%) GAD-7 participants. Transcript areas under the receiver operating characteristic curve were 0.791 (95% confidence interval 0.761–0.820) for PHQ-9 and 0.822 (0.784–0.857) for GAD-7; precision-recall areas were 0.453 and 0.356. Calibration intercepts were 0.02 and −0.02; slopes were 1.10 for both. At sensitivity-directed thresholds, sensitivity was 81.4% and 80.1%, positive predictive value 36.1% and 23.5%. Improvements over response length were small and uncertain. Decision-curve patterns favored transcript content more consistently for GAD-7 than PHQ-9.

Conclusion. Participant-wise transcript models identified screening positivity with useful internal ranking and high negative predictive values, but incremental value and clinical utility were outcome-dependent. External, prospective, subgroup, and workflow validation are required before operational use.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Depressive and anxiety symptoms often first appear in adolescence. Social demands, identity, schooling, sleep, family relationships, and neurodevelopment all interact during these years, and recent studies from China, the United Kingdom, South Africa, and Kenya report that symptom frequency and correlates differ by setting, instrument, social context, and observation period.[1–4] Anxiety and depression can disrupt education, relationships, physical health, and quality of life. They may also coexist with self-harm or other conditions that need timely assessment. Primary-care evidence shows that screening reach is uneven, and implementation studies find that an instrument helps only when it connects to a defined evaluation and referral pathway.[5,6] Screening therefore depends on accessible assessment, proportionate communication, safety procedures, and the capacity to manage positive and negative results alike.

Brief self-report measures sit at the center of these pathways. Recent adolescent and student studies support the reliability and construct validity of Patient Health Questionnaire-9 (PHQ-9) and Generalized

Anxiety Disorder-7 (GAD-7) scores, and they also show that item selection, thresholds, and omitted concerns vary with population and intended use.[7–11] The original PHQ-9 and GAD-7 studies set out the instruments' measurement properties.[12,13] Neither questionnaire establishes a psychiatric diagnosis by itself. A threshold of 10 is an operational definition of screening positivity, not proof of a depressive or anxiety disorder, and it cannot replace assessment of impairment, duration, differential diagnosis, or immediate safety. Digital mental-health systems increasingly collect free text, spoken responses, passive sensor data, and longitudinal context. The Multi-Modal Psychological Assessment (MMPsy) study paired audio with text to estimate mental-health outcomes in an automated interview.[14] Other adolescent studies have used interview audio and text, multi-wave psychosocial data, mobile and wearable sensors, or longitudinal childhood records.[15–18] Natural language processing work adds that psychiatric records, psychotherapy dialogue, self-disclosed anxiety on social media, clinical interviews, voice, and population-level language can all carry mental-health signal.[19–24] Context still matters. Language models stay sensitive to domain and platform, so validation has to be done in the setting of intended use.[25]

MMPsy is a useful test bed because participant-level Mandarin transcripts and questionnaire scores are public. It also poses a hard methodological problem. Each participant contributes several interview segments. If segments rather than participants are split, related language from one person can land in both development and evaluation data. The interview path is adaptive, too. Response length, segment count, punctuation, and prompt-dependent structure may therefore track questionnaire scores, so a complex model could look predictive while mostly capturing how much a participant said or how the interview branched. Several aspects of prediction-model evaluation matter here. Every data-dependent step must stay inside the training data. Vocabulary construction, inverse-document-frequency weights, penalty selection, calibration, and threshold selection can each pass information from an evaluation fold if done globally. Discrimination alone is not enough. The area under the receiver operating characteristic curve (AUROC) describes ranking across thresholds, not whether predicted probabilities are numerically reliable. Precision-recall performance adds useful detail when screening positivity is uncommon, and calibration and sensitivity-directed threshold metrics show whether probabilities and decisions can be used. A response-length comparator checks whether lexical content adds anything beyond verbosity and interview duration. Decision-curve analysis compares consequences across threshold probabilities, yet it depends on an intended use and cannot replace a prospective impact study.

Current reporting and appraisal guidance supports this multidimensional approach. TRIPOD+AI asks for transparent specification of the population, outcome, predictors, modeling procedure, validation design, and performance measures.[26] PROBAST+AI points to risks from participant selection, outcome definition, predictor handling, sample size, overfitting, and incomplete validation.[27] These principles matter more for language-based mental-health models, because linguistic, cultural, temporal, and platform shifts can exceed the random variation that conventional cross-validation captures. This framing separates three questions that digital screening research often merges. First: can an algorithm rank screening-positive participants above screening-negative ones within a single dataset? Second: does transcript content improve decisions beyond simpler information already available from the interview process? Third: does using the resulting probabilities improve care in a new service? The present study addresses the first two directly under internal validation. It uses decision analysis to inform the third but not to answer it. Keeping the questions apart stops internal discrimination from being mistaken for transportability or clinical effectiveness.

The evidence gap is not whether multimodal models can fit MMPsy. It is whether transcript content alone gives reproducible, calibrated, participant-level information beyond simple response quantity when every modeling choice is isolated from the evaluation participants. The novelty of this study lies in combining participant-wise nested cross-validation, fold-specific vocabulary construction, inner-fold probability calibration, sensitivity-directed threshold selection, an explicit response-length comparator, paired bootstrap uncertainty, and decision-curve evaluation for two separate screening outcomes. The primary aim was to develop and internally validate transcript-only probabilistic models for PHQ-9 and GAD-7 screening positivity in MMPsy. Secondary objectives were to quantify incremental performance beyond response length, to evaluate calibration and threshold-dependent triage characteristics, and to define the evidence required before cultural adaptation or clinical implementation.

2. Methods

2.1. Study design and setting

This was a secondary prediction-model development study using publicly available participant-level data. The analysis was designed to estimate internal generalization performance for new participants sampled from the MMPsy distribution, rather than to estimate population prevalence, diagnose psychiatric disorders, or infer a causal effect of language on mental health. Two separate binary probabilistic models were developed: one for PHQ-9 screening positivity and one for GAD-7 screening positivity. The manuscript follows TRIPOD+AI principles for prediction-model studies.[26] Methodological decisions were also considered against PROBAST+AI domains.[27]

The analysis unit was the participant. All segments belonging to a participant remained together throughout resampling. Model development used nested cross-validation: an outer loop estimated performance in participants excluded from every

model-building operation, while an inner loop selected penalization, generated training-only predictions for calibration, and selected the classification threshold. This structure was chosen because a single train-test split would use data inefficiently and could produce an unstable estimate, whereas ordinary cross-validation without nesting would allow performance estimation to depend on tuning decisions made with the same held-out observations.

2.2. Participants

Data source, provenance, and participant inclusion. The MMPsy dataset was obtained from the official distribution linked by the source repository. The analyzed wav_select.json file had SHA-256 fingerprint 4f19f1d0b0518ebaedbc0e2a067214b5d3c33d68d670 6b7d19d4d3f38df24db7. This fingerprint was recorded before analysis to distinguish the locked analytic source from later repository revisions or formatting changes. The public file contains a participant identifier, PHQ-9 score and level, GAD-7 score and level, and a mapping of interview audio-segment filenames to Mandarin transcripts. The repository describes data collection among students in Guangdong Province, China, through a computer-based mental-health assessment and interview system; the source publication provides the broader multimodal study context.[14]

All records in the locked file were eligible if they contained a unique participant identifier, at least one transcript segment, and numeric PHQ-9 and GAD-7 scores. No sampling was performed after the public file was obtained. The final cohort therefore comprised all 1,275 records. The analytic file contained 9,417 segments. Counts, identifiers, transcript completeness, score ranges, segment counts, and transcript lengths were checked before modeling. Duplicate participant identifiers and identical concatenated transcripts were assessed directly. Because no demographic, school, site, or interviewer variables were available in the analytic file, these characteristics could not be used for descriptive stratification, covariate adjustment, or subgroup performance analysis.

The repository documentation contains two sample-size descriptions: a narrative statement referring to 1,000 shared students and a data-description statement referring to 1,275 users. The locked JSON file contains 1,275 unique records, and that directly observed count defined the analysis cohort. This distinction is a provenance clarification rather than an exclusion or imputation decision.

2.3. Variables and instruments

Outcomes. The two outcomes were defined before text processing. PHQ-9 screening positivity was a total score of at least 10, and GAD-7 screening positivity was a total score of at least 10. These thresholds identify clinically relevant symptom burden for screening analyses but were not treated as diagnostic reference standards. The original continuous scores were retained for cohort description; the binary outcomes were used for model development.

Outcome labels were not derived from the transcripts, and neither questionnaire score nor binary label entered any predictor matrix. The models therefore estimated concordance between language and self-reported screening status, not accuracy against a structured clinical interview. Because the same interview context may elicit content closely related to questionnaire domains, estimated performance may partly reflect alignment between two self-report processes conducted in one assessment system.

Transcript assembly and predictor representation. Segments were concatenated in their source order to produce one transcript per participant. No participant quotation was reproduced in the analysis outputs or manuscript. Whitespace was removed for character-level feature generation, while punctuation present in the source transcript was retained. Character n-grams of lengths two to four were counted. This representation was chosen because Mandarin word segmentation can introduce an additional model-dependent preprocessing step, whereas short character sequences capture local lexical patterns and are relatively tolerant of tokenization variation.

Within each outer-training set, candidate n-grams were ranked by total frequency and the 6,000 most frequent features were retained. The vocabulary was never constructed using the outer-test participants. Count matrices were transformed using term frequency-inverse document frequency (TF-IDF), defined for participant i and feature j as $TF\text{-}IDF_{i,j} = tf_{i,j} \times \{\log[(N + 1)/(df_j + 1)] + 1\}$, where N was the number of outer-training participants and df_j was the number containing feature j. Inverse-document-frequency weights were calculated from outer-training participants only, and rows were L2-normalized. The corresponding locked vocabulary and weights were then applied to the outer-test transcripts. The feature limit was fixed before outcome modeling to control dimensionality and computational cost; it was not selected separately for PHQ-9 and GAD-7.

No demographic or acoustic features were included. This transcript-only scope was deliberate: it created a lower-complexity model that could be compared directly with response quantity and prevented the analysis from attributing multimodal information to text. The approach does not imply that text is superior to audio or multimodal representations; it isolates a specific candidate signal.

Prediction models and comparators. The primary model was logistic regression with elastic-net penalization. For participant i, the predicted probability was $p_i = 1/\{1 + \exp[-(\beta_0 + x_i^T\beta)]\}$. The mixing parameter was fixed at 0.5, providing a balance between the sparse selection tendency of the lasso and the correlated-feature stability of ridge penalization. The penalty parameter was selected within the inner cross-validation loop by minimizing binomial deviance. Predictor standardization was performed by the modeling software inside the training process.

Two comparators were defined. The response-length model was an unpenalized logistic regression using log-transformed character count and log-transformed number of interview segments. It represented information available from response quantity without lexical content. The prevalence comparator assigned every outer-test participant the outcome prevalence observed in the corresponding outer-training set. The prevalence model provided a reference for Brier score and precision-recall interpretation but was not intended as an operational classifier.

Models for PHQ-9 and GAD-7 were fitted separately. This choice allowed outcome-specific probabilities, calibration, thresholds, and decision consequences. Although depression and anxiety frequently co-occur, a joint multivariate classifier would answer a different question and could obscure whether one outcome was clearly easier to identify than the other. The overlap between outcomes was therefore described but not used as a predictor.

2.4. Data collection procedure

Nested participant-wise validation, calibration, and threshold selection. Outer folds were stratified by the four combinations of PHQ-9 and GAD-7 screening status. This produced test-fold sizes of 256, 256, 256, 254, and 253 participants. PHQ-9-positive counts were 52, 52, 52, 51, and 51; GAD-7-positive counts were 28, 28, 28, 26, and 26. Each participant appeared in exactly one outer-test fold. The stratification maintained both outcome prevalences while preserving participant independence.

For every outer fold, the remaining participants formed the outer-training set. A four-fold inner stratified cross-validation was created within that training set. Elastic-net penalties were evaluated only on inner data. Prevalidated inner predictions at the selected penalty were converted to probabilities when required and used to fit a logistic Platt calibrator. This calibrator mapped model output to the observed event rate within the outer-training data. The fitted calibrator was then applied unchanged to predictions from the outer-test participants.

A classification threshold was selected from the calibrated inner predictions with a target sensitivity of at least 0.80. The highest threshold meeting that target was chosen, thereby maximizing specificity subject to the sensitivity constraint in the inner data. The fold-specific threshold was applied without modification to the corresponding outer-test probabilities. As a result, confusion matrices pooled decisions from five thresholds selected independently of their test outcomes. No global threshold was tuned on pooled out-of-fold predictions.

This validation design estimates performance when the complete modeling procedure is repeated for a new sample from the same underlying source. It does not estimate performance after deployment to another school, region, language, period, or interview platform. Empirical external-validation and sample-size studies show why a locked model, a defined target setting, adequate events, calibration assessment, and decision utility are required in independent data.[28–30]

2.5. Statistical analysis

Performance measures. Discrimination was evaluated using AUROC and area under the precision-recall curve (AUPRC). AUROC estimates the probability that a randomly selected positive participant receives a higher predicted probability than a randomly selected negative participant. AUPRC summarizes precision across recall levels and was interpreted relative to outcome prevalence. Overall probability accuracy was evaluated with the Brier score, where lower values indicate smaller mean squared error.

Calibration-in-the-large was estimated by fitting a logistic model with the predicted log odds as an offset; its intercept is ideally zero. Calibration slope was estimated by regressing the outcome on predicted log odds; its ideal value is one. Slopes below one commonly indicate predictions that are too extreme, while slopes above one indicate insufficient spread, although cross-validated aggregation and finite samples also influence the estimate.[31] Calibration plots grouped predicted probabilities into quantile-based bins for display. Binned points were used for visualization rather than formal hypothesis testing.

At the fold-specific sensitivity-directed thresholds, the analysis calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and counts of true-positive, false-positive, true-negative, and false-negative classifications. PPV and NPV were treated as prevalence-dependent cohort properties, not invariant model characteristics. Particular attention was given to false positives because an adolescent triage system may impose additional interviews, concern, or stigma, and to false negatives because a screening pathway should not create false reassurance in the presence of acute clinical concern.

Incremental performance compared transcript and response-length models using paired differences in AUROC, AUPRC, and Brier score. For Brier score, a positive improvement indicated lower error for the transcript model. Uncertainty was estimated with 2,000 bootstrap samples drawn at the participant level. Each bootstrap retained the paired outcome, transcript probability, response-length probability, and threshold decision. Percentile 95% confidence intervals were calculated. Intervals were used to convey precision and compatibility rather than as binary declarations of effect.

Decision-curve analysis. Decision-curve analysis compared the transcript model, response-length model, treat-all strategy, and treat-none strategy across threshold probabilities from 0.05 to 0.50. At threshold probability p_t, net benefit was $NB(p_t) = TP/N - (FP/N) \times p_t/(1 - p_t)$, where TP and FP were the true-positive and false-positive counts and N was the cohort size. A threshold represents the risk at which the expected value of further assessment is considered sufficient to justify its burdens.

The evaluated range was exploratory because no service-specific threshold, referral capacity, cost ratio, or adolescent preference study was available. A positive net benefit relative to treat-none is not evidence that an intervention improves health outcomes. Similarly, a model may have favorable net benefit in retrospective data but remain unsuitable because of calibration drift, inequitable errors, privacy risks, workflow failure, or lack of an effective downstream service.

Software. Analyses were performed in R version 4.5.3 using glmnet 4.1–10, Matrix 1.7–4, and jsonlite 2.0.0. The analysis script fixed the random seeds used for resampling and bootstrap generation, wrote participant-level out-of-fold predictions, recorded fold sizes, penalty values, calibration coefficients, and thresholds, and generated machine-readable result tables.

2.6. Ethical approval

The analysis used a public, anonymized dataset and did not contact participants or attempt re-identification. Ethical approval for the secondary analysis was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/030). The source repository states that data were collected in collaboration with education authorities and anonymized before release; participant-level primary-consent documents were not included in the public materials. No transcript excerpts or possibly identifying language were published.

Free text from adolescents can contain sensitive information even after direct identifiers are removed. Reuse therefore requires minimization, access control, purpose limitation, and assessment of whether language could permit attribute inference or re-identification. Large-scale federated adolescent prediction research illustrates one technical route for reducing centralized data transfer, while also leaving consent, purpose limitation, and governance as separate requirements.[32] The current analysis stored the locked source and derived predictions within the case workspace and reported aggregate findings.

3. Results

3.1. Cohort and outcome distribution

All 1,275 records met the analytic inclusion criteria. Participant identifiers and concatenated transcripts were unique. PHQ-9, GAD-7, and transcript fields were complete. The dataset contained 9,417 segments, with a median of 7 segments per participant (interquartile range [IQR] 7–8; range 4–10), and the median transcript length was 214 characters (IQR 121–390; range 24–3,967). The median PHQ-9 score was 6 (IQR 4–9; range 2–24). The median GAD-7 score was 2 (IQR 1–6; range 0–21).

PHQ-9 screening positivity occurred in 258 participants (20.2%). GAD-7 screening positivity occurred in 136 (10.7%). A total of 118 participants were positive on both measures, 140 were positive only on PHQ-9, 18 were positive only on GAD-7, and 999 were negative on both. The overlap was asymmetric: 86.8% of GAD-7-positive participants were also PHQ-9 positive, whereas 45.7% of PHQ-9-positive participants were also GAD-7 positive. The full cohort distributions are given in Table 1, and outcome overlap is visualized in Figure 1.

Table 1. Participant and outcome characteristics in the locked MMPsy cohort.

Characteristic	Value
Participants	1,275
Interview segments	9,417
Segments per participant	Median 7 (IQR 7–8); range 4–10
Transcript length, characters	Median 214 (IQR 121–390); range 24–3,967
PHQ-9 score	Median 6 (IQR 4–9); range 2–24
GAD-7 score	Median 2 (IQR 1–6); range 0–21
PHQ-9 ≥ 10	258 (20.2%)
GAD-7 ≥ 10	136 (10.7%)
Both thresholds	118 (9.3%)

IQR, interquartile range; PHQ-9, Patient Health Questionnaire-9; GAD-7, Generalized Anxiety Disorder-7.

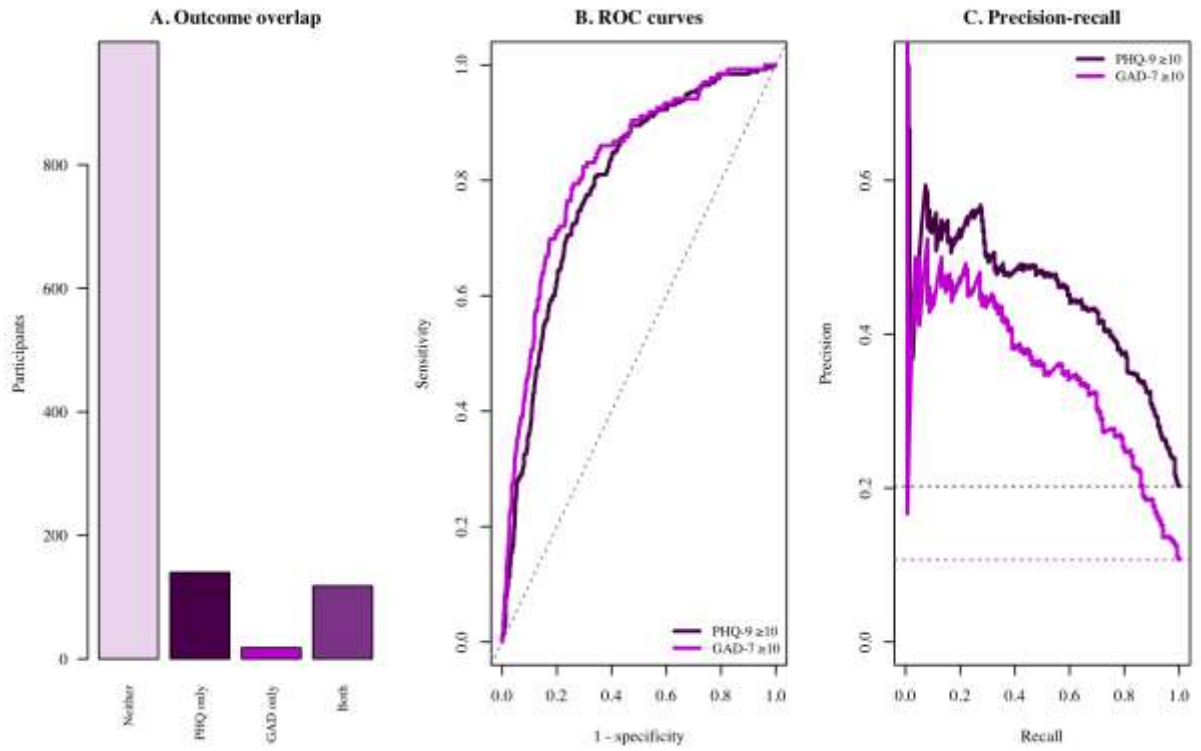


Figure 1. Cohort outcome overlap and participant-level out-of-fold discrimination. (A) Mutually exclusive PHQ-9 and GAD-7 screening-status groups: neither (n = 999), PHQ-9 positive only ("PHQ only", n = 140), GAD-7 positive only ("GAD only", n = 18), and both (n = 118). (B) Receiver operating characteristic (ROC) curves. (C) Precision-recall curves. Dotted precision baselines indicate outcome prevalence.

3.2. Discrimination and probability accuracy

All performance estimates were based on outer-fold predictions. These participants were excluded from vocabulary construction, TF-IDF weighting, penalty selection, model fitting, calibration, and threshold selection. The PHQ-9 transcript model achieved an AUROC of 0.791 (95% confidence interval [CI] 0.761–0.820) and an AUPRC of 0.453 (95% CI 0.394–0.517). The GAD-7 transcript model achieved an AUROC of 0.822 (95% CI 0.784–0.857) and an AUPRC of 0.356 (95% CI 0.284–0.435). AUPRC exceeded the corresponding prevalences of 0.202 and 0.107. Precision nevertheless decreased as recall increased, particularly for GAD-7.

The Brier score was 0.137 (95% CI 0.125–0.150) for PHQ-9 and 0.082 (95% CI 0.070–0.093) for GAD-7. Response-length models achieved AUROCs of 0.769 and 0.788, respectively. Complete discrimination, probability-error, and calibration estimates are shown in Table 2, while ROC and precision-recall curves are shown in Figure 1. These results show that response quantity alone contained substantial predictive information, especially for PHQ-9.

Table 2. Out-of-fold discrimination, probability error, and calibration.

Outcome	Model	AUROC	AUPRC	Brier	Cal. int./slope
PHQ-9 ≥ 10	Transcript	0.791 (0.761–0.820)	0.453 (0.394–0.517)	0.137 (0.125–0.150)	0.02 / 1.10
PHQ-9 ≥ 10	Length	0.769	0.439	0.135	0.00 / 0.99
GAD-7 ≥ 10	Transcript	0.822 (0.784–0.857)	0.356 (0.284–0.435)	0.082 (0.070–0.093)	−0.02 / 1.10
GAD-7 ≥ 10	Length	0.788	0.301	0.085	0.00 / 0.98

All transcript estimates are pooled outer-fold results. Values in parentheses are participant-bootstrap 95% confidence intervals. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; Cal., calibration; int., intercept.

3.3. Calibration

Calibration-in-the-large was close to zero for both transcript models: 0.02 for PHQ-9 and −0.02 for GAD-7. Calibration slopes were 1.10 for both outcomes, indicating that pooled calibrated probabilities were slightly less dispersed than the outcome relationship observed in the complete out-of-fold sample. The response-length models had intercepts near zero and slopes near one. Numerical calibration estimates appear in Table 2. Calibration curves are shown in Figure 2. They were broadly aligned with the diagonal across densely populated risk ranges. Observations at higher predicted probabilities were less precise, because fewer participants contributed to those bins.

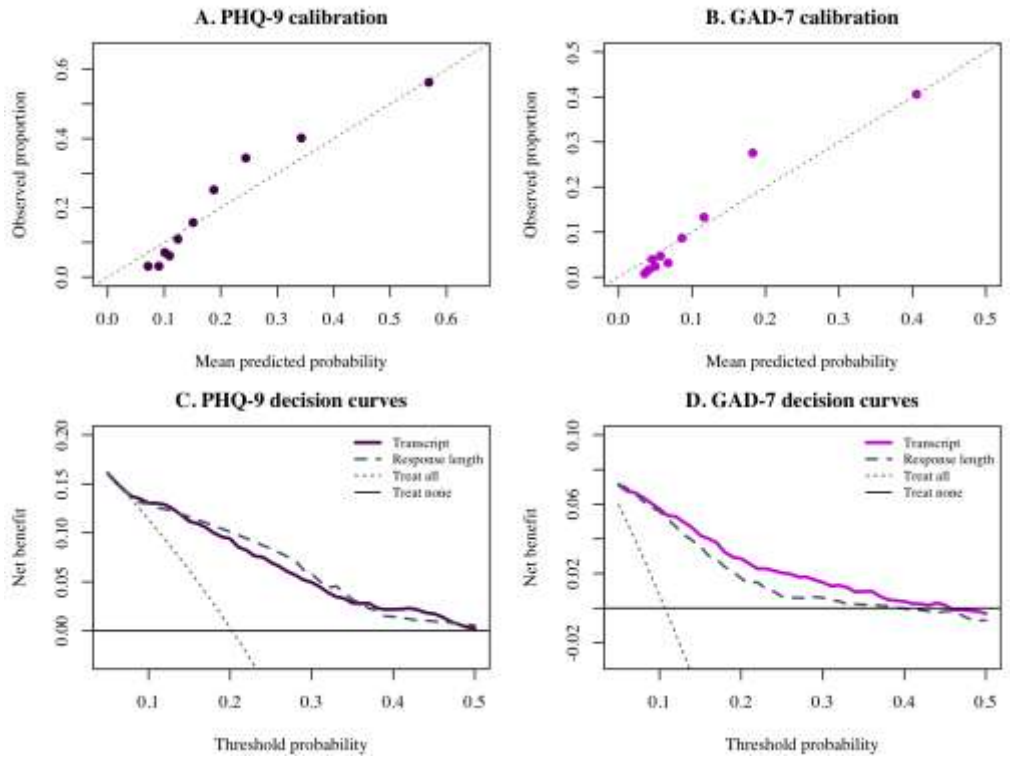


Figure 2. Participant-level out-of-fold calibration and decision curves. (A, B) Quantile-group calibration for PHQ-9 and GAD-7. (C, D) Net benefit of the transcript, response-length, treat-all, and treat-none strategies across threshold probabilities for PHQ-9 and GAD-7.

3.4. Threshold-dependent triage characteristics

The five PHQ-9 thresholds selected in inner cross-validation ranged from 0.143 to 0.156. When applied to outer-test participants, pooled sensitivity was 81.4% (95% CI 76.6%–85.9%), specificity 63.5% (95% CI 60.5%–66.4%), PPV 36.1%, and NPV 93.1% (95% CI 91.1%–94.9%). The model identified 210 of 258 positive participants and classified 371 of 1,017 negative participants as positive.

The five GAD-7 thresholds ranged from 0.074 to 0.091. Pooled sensitivity was 80.1% (95% CI 73.2%–86.8%), specificity 68.9% (95% CI 66.3%–71.5%), PPV 23.5%, and NPV 96.7% (95% CI 95.4%–97.9%). The model identified 109 of 136 positive participants and classified 354 of 1,139 negative participants as positive. The high NPVs reflect both model ranking and the low outcome prevalences. They should not be read as permission to disregard warning signs, functional deterioration, or safety concerns. Fold-specific threshold ranges, operating characteristics, and all four confusion-matrix counts are shown in Table 3.

Table 3. Performance at fold-specific sensitivity-directed thresholds.

Outcome	Threshold range	Sensitivity (95% CI)	Specificity (95% CI)	PPV	NPV (95% CI)	TP / FP / TN / FN
PHQ-9 ≥ 10	0.143–0.156	81.4% (76.6–85.9)	63.5% (60.5–66.4)	36.1%	93.1% (91.1–94.9)	210 / 371 / 646 / 48
GAD-7 ≥ 10	0.074–0.091	80.1% (73.2–86.8)	68.9% (66.3–71.5)	23.5%	96.7% (95.4–97.9)	109 / 354 / 785 / 27

Thresholds were selected within each outer-training set using inner predictions and applied unchanged to the outer-test fold. CI, 95% participant-bootstrap confidence interval; threshold range, the five fold-specific thresholds (minimum–maximum); PPV intervals were not reported; PPV, positive predictive value; NPV, negative predictive value; TP, true positive; FP, false positive; TN, true negative; FN, false negative.

3.5. Incremental value beyond response length

For PHQ-9, the difference between transcript and response-length models was 0.021 in AUROC (95% CI −0.009 to 0.054) and 0.014 in AUPRC (95% CI −0.046 to 0.070). The Brier-score improvement was −0.002 (95% CI −0.010 to 0.005), indicating no reliable probability-error advantage for the transcript model. For GAD-7, differences were 0.033 in AUROC (95% CI −0.004 to 0.071), 0.055 in AUPRC (95% CI −0.024 to 0.132), and 0.003 in Brier-score improvement (95% CI −0.003 to 0.009). The paired estimates and 95% CIs are summarized in Table 4. Point estimates favored transcript content more consistently for GAD-7. Even so, every paired interval included no incremental improvement.

Table 4. Paired incremental performance of transcript content beyond response length.

Outcome	Comparison	Difference	Bootstrap 95% CI
PHQ-9 ≥ 10	AUROC	0.021	−0.009 to 0.054
PHQ-9 ≥ 10	AUPRC	0.014	−0.046 to 0.070
PHQ-9 ≥ 10	Brier improvement	−0.002	−0.010 to 0.005
GAD-7 ≥ 10	AUROC	0.033	−0.004 to 0.071
GAD-7 ≥ 10	AUPRC	0.055	−0.024 to 0.132
GAD-7 ≥ 10	Brier improvement	0.003	−0.003 to 0.009

Differences are transcript minus response-length model, except Brier improvement, which is response-length minus transcript Brier score; positive values favor transcript content. CI, confidence interval.

3.6. Decision-curve patterns

Both transcript models had positive net benefit relative to treat-none across much of the evaluated range, but comparison with the response-length model was outcome-dependent. For PHQ-9, the two models produced similar net benefit at low thresholds, and the response-length model was higher across many thresholds between approximately 0.10 and 0.35. The transcript model became higher in selected portions of the upper range but did not dominate overall. For GAD-7, the transcript model exceeded the response-length model at 44 of 46 evaluated thresholds, including clinically plausible thresholds from 0.10 through 0.40, although absolute net benefit declined as thresholds increased. Figure 2 shows these patterns. The retrospective curves support further evaluation of GAD-7 triage. They do not establish service impact.

4. Discussion

4.1. Principal findings

This participant-wise analysis found that short-range character patterns in Mandarin digital-interview transcripts carried internal predictive information about PHQ-9 and GAD-7 screening positivity. Discrimination was moderate for PHQ-9 and stronger for GAD-7. Global calibration after nested Platt scaling was close to ideal in intercept. Slopes were slightly above one. Sensitivity-directed thresholds reached the intended sensitivity range in outer-test participants and gave high NPVs, but PPVs were low. The models therefore behaved as broad triage tools, reducing the number of screening-positive participants given a low probability at the price of sending many screening-negative participants toward further assessment.

The most consequential result was the comparison with response length. Transcript models showed no large or statistically precise advantage. For PHQ-9, probability error was slightly higher than for the length model, and decision-curve performance did not consistently favor lexical content. For GAD-7, point estimates and most decision thresholds favored the transcript model, yet the paired bootstrap intervals still included no difference. The interpretation shifts accordingly, from "language clearly adds value" to a more useful conclusion: language holds a candidate signal, but in this corpus its incremental contribution is limited and depends on the outcome. Future studies should keep a response-quantity comparator. They should also show that added complexity brings a practically meaningful benefit.

4.2. Comparison with the existing literature

The findings agree with recent evidence that language, speech, sensor, and longitudinal data can encode correlates of depressive and anxiety symptoms without yet working as universally portable clinical biomarkers.[15–25] Those studies differ widely in labels, platforms, populations, and validation procedures. The present analysis adds a participant-wise adolescent transcript example and shows how much apparent performance a two-variable response-length model can recover. In an adaptive interview, participants with greater distress may give longer answers, receive extra prompts, or produce punctuation and segmentation patterns that carry signal within the platform but do not travel beyond it.

The source MMPsy study used multimodal representations and a specialized architecture, and the present study did not try to reproduce or compete with that architecture.[14] It addressed a narrower question. Using only publicly available transcripts and scores, a transcript-only elastic-net model offers transparent regularization, stable computation, and auditable nested validation, and its purpose is to set a defensible baseline against which claims that lexical content has independent value can be judged.

Recent large adolescent studies have applied artificial neural networks and random forests to identify depression-related patterns, and they often report strong apparent discrimination in the source sample.[33,34] Longitudinal, multimodal imaging, and connectome-based investigations extend that work by estimating future onset or risk instead of reproducing a contemporaneous questionnaire threshold.[35–37] Their questions are clinically different. Performance estimates therefore cannot be carried directly across outcomes or time horizons.

Machine-learning studies of treatment response have predicted remission after cognitive behavioral therapy and outpatient medication management.[38,39] Other work has modeled major depressive disorder among the offspring of depressed parents, as well as longitudinal trajectories in adolescents already at high risk.[40,41] Such studies draw on richer clinical or temporal information than MMPsy provides. They offer a roadmap from cross-sectional concordance toward prospective decision support.

Recent research has also addressed adolescent suicide risk and nationwide smartphone-derived correlates of anxiety and depression.[42,43] Studies of self-injurious behavior and primary-care prediction show why rare but consequential outcomes should be defined separately from general symptom elevation.[44,45] One contemporary interpretable model shows the continuing emphasis on explanation. Yet explanation methods do not replace leakage control, calibration, or external validation.[46]

GAD-7 showed a stronger AUROC than PHQ-9 in the present study. Differences in linguistic expression, score distribution, interview prompts, or separability within this corpus may explain the gap, but the lower prevalence of GAD-7 positivity produced a PPV of only 23.5% at the sensitivity-directed threshold, and a model can rank participants well while still raising several false-positive flags for every true-positive flag. The burden is explicit. The AUPRC, threshold metrics, and decision curves in Table 2, Table 3, and Figure 2 make it visible.

The study also complements recent PHQ-9 and GAD-7 research in adolescents and young adults.[7–11] A screening score has a different status from a diagnostic interview, in line with the original purposes of both instruments.[12,13] The transcript models were trained to reproduce questionnaire threshold status, so they cannot get past the conceptual limits of that target. Future diagnostic work should use an independent clinical reference standard and prospectively specified timing. It should also test whether language adds information beyond the questionnaire itself.

4.3. Interpretation and mechanism

Calibration, thresholds, and practical importance. Calibration matters whenever predicted probabilities influence who receives further assessment, how urgently a case is reviewed, or whether a model abstains.[28–31] The near-zero calibration intercepts show that average predicted and observed risks were similar in the pooled internal validation data. Slopes of 1.10 suggest modest under-dispersion, not the exaggerated extremes often linked to overfitting. These results reassure for the internal sample but should not be extrapolated. Calibration can shift quickly when prevalence, interview design, transcription quality, or participant characteristics change, and a model moved to another school or country may need recalibration even if its rank discrimination stays acceptable.

The sensitivity-directed thresholds are an operational illustration, not a recommended clinical cut-off. Targeting approximately 0.80 sensitivity left 48 PHQ-9-positive and 27 GAD-7-positive participants below threshold, and it also generated 371 and 354 false-positive classifications. Is the trade-off acceptable? That depends on the intended action. A brief supportive review may tolerate more false positives than referral to a scarce specialist service, whereas a system meant to identify acute risk would need a different outcome, much higher sensitivity, explicit red-flag logic, and immediate human escalation. Thresholds should be chosen with stakeholders and service data, not copied from the current folds.

Statistical and practical importance also come apart in the incremental results. Even the favorable GAD-7 point estimate of 0.033 AUROC does not say how many additional adolescents would benefit or how many extra assessments would be needed. Paired intervals were compatible with no gain. The decision-curve advantage for GAD-7 was consistent but modest and depended on thresholds. A prospective study should measure referrals, time to assessment, missed cases, workload, participant experience, and downstream care. Improved prediction should not be assumed to improve outcomes.

Interpretation of transcript signals. Character n-grams can capture words, fragments, negation, intensifiers, discourse markers, and punctuation patterns without a Mandarin word segmenter. They can also capture non-clinical artifacts such as prompt wording, transcription conventions, regional expressions, hesitation marks, and boundaries between segments. Coefficients are spread across correlated features. A large elastic-net coefficient therefore does not show that a phrase is a stable symptom marker. For that reason, the final analysis did not publish a list of participant-derived n-grams, because such lists risk decontextualizing sensitive speech, encouraging keyword profiling, and implying a mechanism that the design cannot establish.

The response-length comparator helped separate lexical content from quantity, but it could not remove all interview-process information. If an adaptive branch asks different questions after a concerning answer, the transcript model may identify the branch itself. That is not necessarily useless, since workflow signals can predict screening status. It does limit transportability. A model may fail when deployed with a different interview script, human interviewer, speech-to-text engine, or conversational policy. Future data collection should preserve prompt identifiers and timestamps, so that content, process, and duration can be evaluated separately.

4.4. Implications for practice and policy

Language, culture, and transportability. The model operates on Mandarin character sequences from one education and interview context. It cannot be moved into an Indonesian application by translating individual n-grams or swapping the vocabulary. Language embodies culturally patterned expressions of distress, norms of disclosure, politeness, school pressures, family relationships, and metaphors. The prevalence and meaning of questionnaire scores may differ as well. A valid Indonesian study would need locally designed and ethically reviewed data collection, validated questionnaire versions, an interview protocol suited to the intended use, and either a new model or a carefully evaluated transfer-learning strategy.

External validation is not one generic test. It should target the population, setting, and decision where the model is meant to operate.[28–30] A useful sequence would begin with temporal validation within the same system.

Geographic or school-level validation in Mandarin-speaking cohorts would follow, then platform validation under a changed interview or transcription pipeline, and later cross-language development. Where these variables can be collected ethically, performance should be reported by age, sex or gender, socioeconomic context, school, region, language variety, and transcript quality. Adequate sample size is needed to estimate both overall and subgroup calibration with useful precision.[29]

Recent adolescent multimodal and longitudinal studies illustrate a broader movement toward richly characterized mental-health resources.[17,18,35–37] Such initiatives can support cross-cohort evaluation. Open or reusable data, however, do not remove the need for consent-compatible reuse and community governance. Cross-dataset evaluation is most useful when outcome definitions, timing, and participant contexts are explicitly mapped, not treated as interchangeable.

Fairness, privacy, and human oversight. No fairness audit was possible, because demographic and site attributes were absent from the analytic file. Absence is not evidence of equal performance. Language models can perform differently across dialects, socioeconomic groups, neurodevelopmental profiles, speech or language impairments, and levels of digital familiarity, and an aggregate AUROC can stay stable while calibration or false-negative rates differ in a smaller group. Future protocols should prespecify protected and clinically relevant subgroups, justify data collection, report uncertainty, and define actions when performance disparities appear.

Privacy risk is unusually salient, since a transcript can reveal relationships, events, locations, health conditions, or identity through context even when names are removed. Privacy-aware model development should minimize retained text, restrict access, separate identifiers, record use, and ask whether features or model outputs permit sensitive inference. Federated adolescent modeling may reduce centralized data transfer. It does not resolve purpose limitation, consent, or governance.[32] Adolescents should receive understandable information about what is analyzed, who can see the result, how long data are retained, and whether a human can override or contest the score.

Human oversight has to be defined as a workflow. A slogan is not enough. A reviewer needs the model's intended use, its probability, an uncertainty or abstention signal, relevant questionnaire information, and a clear action pathway. The model should never suppress an assessment triggered by suicidal ideation, abrupt deterioration, psychosis, mania, abuse, violence, or another urgent concern, and reporting and appraisal guidance, together with empirical external-validation studies, supports assessing calibration, decision utility, human-model interaction, and performance across groups alongside discrimination.[26–30]

Implications and next research steps. The immediate implication is methodological. Transcript-based adolescent mental-health studies should separate participants before any feature engineering, compare language with response quantity and workflow variables, report calibration and precision-recall results, and carry uncertainty into claims of incremental value. A model that performs similarly to length may still motivate research. It does not justify the privacy and governance burden of analysing content.

The next study should lock the present pipeline and evaluate it in a temporally or geographically separate Mandarin cohort without retuning. If performance falls, recalibration and model updating should be prespecified and tested on a further independent sample. Prompt identifiers and transcription-quality measures should be retained. A diagnostic substudy should compare predictions with structured clinical assessment and ask whether the model adds value beyond questionnaires and readily available contextual variables.

For Indonesian development, the appropriate route is a new prospective, multisite protocol. The protocol should define whether the tool supports universal screening, appointment prioritization, or clinician review, and select outcomes and time horizons to match; it should use validated local measures and, where feasible, an independent clinical reference; and it should recruit a sufficiently diverse sample and involve adolescents, families, clinicians, school personnel, and ethics experts in threshold and workflow design. Fairness, privacy, abstention, human override, red-flag escalation, calibration monitoring, and incident response belong in the protocol from the start. They should not wait for deployment.

4.5. Strengths and limitations

Strengths. The study has several strengths. The participant was the unit of every split. This eliminated leakage across segments from the same person. Vocabulary construction, TF-IDF weighting, penalty selection, calibration, and threshold selection were all restricted to outer-training data, and the analysis used a locked public-data fingerprint and reproducible code that allow record-level reconciliation. Performance was reported across discrimination, precision-recall, probability error, calibration, threshold metrics, and decision curves. Paired participant bootstraps quantified uncertainty in incremental performance, and the response-length comparator tested a plausible low-complexity explanation for predictive performance. Finally, interpretation was tied to screening rather than diagnosis and kept internal validation distinct from transportability.

Limitations. The principal limitation is the absence of external validation. Nested cross-validation reduces resubstitution optimism, but it cannot reproduce shifts in geography, school systems, language, time, prevalence, interview design, or transcription. The reported intervals quantify sampling uncertainty. They are conditional on the dataset and analysis and do not capture distributional uncertainty in deployment.

Second, the outcomes were self-reported PHQ-9 and GAD-7 thresholds, not diagnoses from a structured

clinical interview. These measures are useful for screening. Even so, the labels contain measurement error and do not distinguish symptom causes, and the analyzed file lacks item-level responses, which prevented an evaluation of whether specific high-risk items or response patterns drove the association.

Third, demographic, school, site, socioeconomic, developmental, treatment, and clinical-history variables were unavailable. The study could not describe representativeness, adjust for context, evaluate fairness, or determine whether performance arose from confounding by participant or interview characteristics. The prevalence estimates therefore describe the file, not the underlying student population.

Fourth, adaptive interview prompts and transcription practices may have contributed to model performance. The response-length comparator captured character and segment counts. It did not capture prompt identity, branch sequence, pause structure, or automatic-transcription confidence. Lexical features may encode the assessment process as well as participant language.

Fifth, the feature space was limited to the 6,000 most frequent character 2–4-grams and a fixed elastic-net mixing parameter. This approach, fixed before outcome modeling, supports reproducibility. It does not establish optimality. Other representations might improve performance, while more flexible models could raise overfitting or reduce interpretability, so model comparison should rely on nested evaluation and practical endpoints, not on benchmark maximization alone.

Sixth, the pooled calibration assessment combines predictions from five related fitted models. That suits the evaluation of a development procedure. A deployable system, however, would need a single locked final model, an explicit training set, and validation in an untouched cohort. The current analysis provides neither a deployment-ready coefficient set nor a universal threshold.

Finally, decision-curve analysis was retrospective and used a generic threshold range. No patient, family, clinician, school, or service data were available to define acceptable trade-offs. Net benefit should therefore be read as a comparative analytic signal and not as evidence of improved care.

5. Conclusion

Transcript-only elastic-net models gave internally validated information about PHQ-9 and GAD-7 screening positivity in the MMPsy cohort, with stronger discrimination and a more consistent but modest decision-curve advantage for GAD-7. Calibration was acceptable internally, and sensitivity-directed thresholds yielded high NPVs. PPVs were low, however, and incremental gains beyond response length were small and uncertain. These models should be viewed as research-stage triage candidates, not as diagnostic or autonomous decision tools. External and prospective validation, subgroup evaluation, privacy governance, and a defined human assessment pathway are necessary before use in schools or health services.

Declarations

Use of Generative Artificial Intelligence. The authors declare that no generative artificial intelligence tool was used in the conception, research, analysis, drafting, revision or preparation of this manuscript.

Author Contributions. W.P.: conceptualization, methodology, software, validation, formal analysis, data curation, visualization, writing – original draft. M.C.A.: investigation, validation, writing – review & editing. A.M.: resources, writing – review & editing. H.S.: supervision, project administration, writing – review & editing. All authors approved the final manuscript and accept accountability for the work.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare that they have no conflict of interest.

Ethical Approval. Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/030). The secondary analysis used a public anonymized dataset and involved no new participant contact.

Informed Consent. No new participants were recruited for this secondary analysis. The source repository states that data were collected with education-authority involvement and anonymized before release; participant-level primary-consent documentation was not included in the public materials.

Data Availability. The MMPsy data are publicly available from the official repository and distribution described in the source publication. The locked analytic file is identified by SHA-256 4f19f1d0b0518ebaedbc0e2a067214b5d3c33d68d6706b7d19d4d3f38df24db7. Reanalysis code and aggregate outputs accompany the study package.

Acknowledgements. None.

References

1. Zhang J, Liu D, Ding L, et al. Prevalence of depression in junior and senior adolescents. Front Psychiatry. 2023;14:1182024. doi:10.3389/fpsyt.2023.1182024
2. Young KS, Purves KL, Hübel C, et al. Depression, anxiety and PTSD symptoms before and during the COVID-19 pandemic in the UK. Psychol Med. 2023;53(12):5428-5441. doi:10.1017/S0033291722002501
3. Mkhize M, van der Westhuizen C, Sorsdahl K. Prevalence and factors associated with depression and anxiety among young school-going adolescents in the Western Cape Province of South Africa. Compr Psychiatry. 2024;131:152469. doi:10.1016/j.comppsych.2024.152469
4. Rodriguez M, Osborn TL, Gan JY, et al. Loneliness in Kenyan adolescents: socio-cultural factors and network association with depression and anxiety symptoms. Transcult Psychiatry. 2022;59(6):797-809. doi:10.1177/13634615221099143
5. Davis M, Jones JD, So A, et al. Adolescent depression screening in primary care: who is screened and who is at risk? J Affect Disord. 2022;299:318-325. doi:10.1016/j.jad.2021.12.022

6. Coats T, Zurmehly J. Implementation of an evidence-based clinical guideline for depression screening of the adolescent. Issues Ment Health Nurs. 2023;44(7):610-619. doi:10.1080/01612840.2023.2205523
7. Fonseca-Pedrero E, Díez-Gómez A, Pérez-Albéniz A, et al. Youth screening depression: validation of the Patient Health Questionnaire-9 in a representative sample of adolescents. Psychiatry Res. 2023;328:115486. doi:10.1016/j.psychres.2023.115486
8. Liang K, Chen S, Zhao Y, et al. A new PHQ-2 for Chinese adolescents: identifying core items of the PHQ-9 by network analysis. Child Adolesc Psychiatry Ment Health. 2023;17(4):11. doi:10.1186/s13034-023-00559-1
9. Pranckeviciene A, Saudargiene A, Gecaite-Stonciene J, et al. Validation of the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder-7 in a Lithuanian student sample. PLoS One. 2022;17(4):e0263027. doi:10.1371/journal.pone.0263027
10. Martínez-Vázquez S, Martínez-Galiano JM, Peinado-Molina RA, et al. Validation of the General Anxiety Disorder-7 questionnaire in Spanish nursing students. PeerJ. 2022;10:e14296. doi:10.7717/peerj.14296
11. Cotton SM, Menssink J, Rickwood D, et al. What do the PHQ-9, the GAD-7 and their variants miss in assessing young people presenting to youth mental health services? J Affect Disord. 2026;392:120188. doi:10.1016/j.jad.2025.120188
12. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613. doi:10.1046/j.1525-1497.2001.016009606.x
13. Spitzer RL, Kroenke K, Williams JBW, et al. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092-1097. doi:10.1001/archinte.166.10.1092
14. Qin J, Liu C, Tang T, et al. Mental-Perceiver: audio-textual multi-modal learning for estimating mental disorders. Proc AAAI Conf Artif Intell. 2025;39(23):25029-25037. doi:10.1609/aaai.v39i23.34687
15. Zhang L, Fan Y, Jiang J, et al. Adolescent depression detection model based on multimodal data of interview audio and text. Int J Neural Syst. 2022;32(11):2250045. doi:10.1142/S0129065722500459
16. Hawes MT, Schwartz HA, Son Y, et al. Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning. Psychol Med. 2023;53(13):6205-6211. doi:10.1017/S0033291722003452
17. Mullick T, Radovic A, Shaaban S, et al. Predicting depression in adolescents using mobile and wearable sensors: multimodal machine learning-based exploratory study. JMIR Form Res. 2022;6(6):e35807. doi:10.2196/35807
18. Yoo A, Li F, Youn J, et al. Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning. Sci Rep. 2024;14(1):23282. doi:10.1038/s41598-024-72158-9
19. Hutto A, Zikry TM, Bohac B, et al. Using a natural language processing toolkit to classify electronic health records by psychiatric diagnosis. Health Informatics J. 2024;30(4):14604582241296411. doi:10.1177/14604582241296411
20. Bahar L, Rego SA, Sadeh-Sharvit S. Detecting climate anxiety in therapy through natural language processing. Sci Rep. 2024;14(1):25976. doi:10.1038/s41598-024-75269-5
21. Zarate D, Ball M, Prokofieva M, et al. Identifying self-disclosed anxiety on Twitter: a natural language processing approach. Psychiatry Res. 2023;330:115579. doi:10.1016/j.psychres.2023.115579
22. Ohse J, Hadžić B, Mohammed P, et al. GPT-4 shows potential for identifying social anxiety from clinical interview data. Sci Rep. 2024;14(1):30498. doi:10.1038/s41598-024-82192-2
23. Luo J, Wu Y, Liu M, et al. Differentiation between depression and bipolar disorder in children and adolescents by voice features. Child Adolesc Psychiatry Ment Health. 2024;18(1):19. doi:10.1186/s13034-024-00708-0
24. Mangalik S, Eichstaedt JC, Giorgi S, et al. Robust language-based mental health assessments in time and space through social media. NPJ Digit Med. 2024;7(4):109. doi:10.1038/s41746-024-01100-0
25. Aragón ME, López-Monroy AP, Montes-Y-Gómez M, et al. Adapting language models for mental health analysis on social media. Artif Intell Med. 2025;168:103217. doi:10.1016/j.artmed.2025.103217
26. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
27. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505
28. Gulati G, Upshaw J, Wessler BS, et al. Generalizability of cardiovascular disease clinical prediction models: 158 independent external validations of 104 unique models. Circ Cardiovasc Qual Outcomes. 2022;15(4):e008487. doi:10.1161/CIRCOUTCOMES.121.008487
29. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. Stat Med. 2022;41(7):1280-1295. doi:10.1002/sim.9275
30. Arribas M, de Micheli A, Krakowski K, et al. Joint detection of risk for psychotic disorders or bipolar disorders in clinical practice in the UK: development and validation of a clinical prediction model. Lancet Psychiatry. 2026;13(1):14-23. doi:10.1016/S2215-0366(25)00307-4
31. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7
32. Kuang Y, Liao X, Jiang Z, et al. Federated learning-based prediction of depression among adolescents across multiple districts in China. J Affect Disord. 2025;369:625-632. doi:10.1016/j.jad.2024.10.027
33. Zhou Y, Zhang X, Gong J, et al. Identifying the risk of depression in a large sample of adolescents: an artificial neural network based on random forest. J Adolesc. 2024;96(7):1485-1497. doi:10.1002/jad.12357
34. Gohari MR, Doggett A, Patte KA, et al. Using random forest to identify correlates of depression symptoms among adolescents. Soc Psychiatry Psychiatr Epidemiol. 2024;59(11):2063-2071. doi:10.1007/s00127-024-02695-1
35. Chavanne AV, Paillère Martinot ML, Penttilä J, et al. Anxiety onset in adolescents: a machine-learning prediction. Mol Psychiatry. 2023;28(2):639-646. doi:10.1038/s41380-022-01840-z
36. Gracia-Tabuenca Z, Barbeau EB, Xia Y, et al. Predicting depression risk in early adolescence via multimodal brain imaging. Neuroimage Clin. 2024;42:103604. doi:10.1016/j.nicl.2024.103604
37. Morfini F, Kucyi A, Zhang J, et al. Brain functional connectivity predicts depression and anxiety during childhood and adolescence: a connectome-based predictive modeling approach. Imaging Neurosci. 2025;3:IMAG.a.145. doi:10.1162/IMAG.a.145

38. Bertie LA, Quiroz JC, Berkovsky S, et al. Predicting remission following cognitive behavioural therapy for childhood anxiety disorders: a machine learning approach. Psychol Med. 2024;54(16):4612-4622. doi:10.1017/S0033291724002654
39. Balzen KM, Hensley JK, Nakonezny PA, et al. Predicting remission from depression in youth receiving outpatient medication management. J Clin Psychiatry. 2023;84(4):22m14581. doi:10.4088/JCP.22m14581
40. Stephens A, Allardyce J, Weavers B, et al. Developing and validating a prediction model of adolescent major depressive disorder in the offspring of depressed parents. J Child Psychol Psychiatry. 2023;64(3):367-375. doi:10.1111/jcpp.13704
41. Park J, Choi EK, Choi M. Longitudinal analysis of adolescents at high risk of depression: prediction models. Appl Nurs Res. 2025;82:151927. doi:10.1016/j.apnr.2025.151927
42. Wang H, Yuan H, Zhang Y, et al. Suicide risk prediction for Korean adolescents based on machine learning. Sci Rep. 2025;15(1):14921. doi:10.1038/s41598-025-99626-0
43. Weiß M, Gutzeit J, Pryss R, et al. Common and differential variables of anxiety and depression in adolescence: a nationwide smartphone-based survey. Child Adolesc Psychiatry Ment Health. 2024;18(1):103. doi:10.1186/s13034-024-00793-1
44. Zhang Z, Chen H, Ye Y, et al. Entropy-based risk network identification in adolescent self-injurious behavior using machine learning and network analysis. Transl Psychiatry. 2025;15(1):299. doi:10.1038/s41398-025-03511-3
45. Jones JD, Davis M, Reagan S, et al. Predicting adolescent depression and suicide risk based on preadolescent behavioral health screening in primary care. Acad Pediatr. 2025;25(6):102833. doi:10.1016/j.acap.2025.102833
46. Chen X, Yang B, Han L, et al. Predicting adolescent depression: an interpretable machine learning model. BMC Med Inform Decis Mak. 2026;26(1):218. doi:10.1186/s12911-026-03482-5