Review

Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis

Shengming Zhang[1], MSc, BSc; Chaohai Zhang[1], BSc; Jiaxin Zhang[1,2], PhD

[1]School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

[2]Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology, School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen, Guangdong, China

Corresponding Author:
Jiaxin Zhang, PhD
School of Automation and Intelligent Manufacturing
Southern University of Science and Technology
1088 Xueyuan Avenue, Nanshan District
Shenzhen, Guangdong, 518055
China
Phone: 86 13416141690
Email: zhangjx@sustech.edu.cn

Abstract

Background: Depression affects people’s daily lives and even leads to suicidal behavior. Text-based depression estimation using natural language processing has emerged as a feasible approach for early mental health screening. However, most existing reviews often included studies with weak depression labels, which affected the reliability of the results and further limited the practical application of the automatic depression estimation models.

Objective: This review aimed to evaluate the predictive performance of text-based depression models that used standard labels, and to identify text resources, text representation, model architecture, annotation source, and reporting quality contributing to performance heterogeneity.

Methods: Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines, we systematically searched 4 main databases (PubMed, Scopus, IEEE Xplore, and Web of Science) for studies published between 2014 and 2025. The eligible studies were included: machine learning models were developed based on the text generated by the participants and used validated scales or clinical diagnoses as depression labels. Pooled effect sizes (*r*) were calculated using random-effects meta-analysis with Hartung-Knapp-Sidik-Jonkman correction, and subgroup and meta-regression analyses were conducted to explore potential moderators.

Results: We scanned 3067 articles and finally filtered 15 models from 11 studies for the meta-analysis. The overall pooled effect size was 0.605 (95% CI 0.498-0.693), indicating a large strength of association. Subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features (*r*=0.741, 95% CI 0.648-0.812 vs *r*=0.514, 95% CI 0.385-0.623; *P*<.001 for subgroup difference), and deep learning architectures outperformed shallow models (*r*=0.731, 95% CI 0.660-0.789 vs *r*=0.486, 95% CI 0.352-0.599; *P*<.001). Models trained with clinician diagnoses also outperformed better than those relying on self-report scales (*r*=0.688, 95% CI 0.554-0.787 vs *r*=0.500, 95% CI 0.340-0.631; *P*=.03). Reporting quality was positively associated with model performance (β=0.085, 95% CI 0.050-0.119; *P*<.001). Begg–Mazumdar and Egger tests provided no evidence of small-study effects. Begg–Mazumdar test (Kendall τ=0.17143, *P*=.37) and the Egger test (t_{14}=1.13401, 2-tailed *P*=.28) indicated no evidence of small-study effects.

Conclusions: Text-based depression estimation models trained with standard depression labels demonstrate solid predictive performance, with embedding features, deep model architectures, and clinician diagnosis labels showing significantly higher performance. Transparent reporting is also positively associated with model performance. This study highlights the importance of standard labels, feature representation, and reporting quality for improving model reliability. Unlike prior reviews that included weak or heterogeneous depression labels, this study offers more clinically reliable and comparable evidence. Moreover, this review provides clearer methodological guidance for developing more consistent and practically informative text-based depression screening models.

Trial Registration: PROSPERO CRD420251056902; https://www.crd.york.ac.uk/PROSPERO/view/CRD420251056902

(J Med Internet Res 2026;28:e82686) doi: 10.2196/82686

KEYWORDS

depression; natural language processing; standard labels; text; TRIPOD; Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Introduction

Background

Depression, a type of mental disorder, is clinically characterized by persistent and significant low mood [1]. Individuals with severe depression often experience impaired social functioning and may even exhibit suicidal behaviors [2,3]. Currently, the diagnosis of depression relies on face-to-face consultation by psychiatrists and refers to the patients' self-reports. Although the clinical diagnosis by psychiatrists is regarded as the gold standard of depression detection [4], due to the time-constrained consultations and subjective information in psychiatry [5], the lack of clinical resources has prevented the gold standard from being widely used. To alleviate clinical burden, standardized self-reported scales have been developed based on psychiatrists' clinical experience [6], and several of them have been demonstrated to have comparable performance to clinical diagnosis [7,8]. However, these tools partially address the shortage of clinical resources; limitations such as subjective bias and time-consuming assessments remain [9-11]. With the development of artificial intelligence technologies, the limitations of traditional depression diagnosis methods facilitate the emergence of automatic depression estimation (ADE) models based on various multimodal data sources [12,13].

Among these multimodal data, text-based features have become a popular target in depression estimation studies [14-17]. Unlike other features used in ADE models [18-20], text (language) serves as a natural medium of human communication, conveying not only semantic content but also affective states [21,22]. The ability to capture emotional cues from text aligns closely with the depression diagnostic and may provide references for estimation [23]. For example, Cariola et al [24] found that mothers with depression used more first-person singular pronouns and present-focused words in mother-child dialogues, possibly reflecting heightened self-focus and introspection. Another study demonstrated that the ratio of negative to positive language played a vital role in establishing emotional tone [25]. Additionally, text data can be sourced from various contexts, including social media posts [26,27], SMS messages [28], chat logs [17], clinical transcripts [29], and even electronic health records [30]. Wang et al [27] achieved 91% classification accuracy based on social media posts, while Liu et al [28] reported an area under the curve of 76% on SMS messages. These studies achieved notable performance and supported the feasibility of text-based ADEs. However, differences in the reported contents (eg, text sources, model construction and validation, and depression labels) across studies may lead to substantial variability in model performance, thereby affecting the reliability of the results and the further practical application.

The quality of training data is critical to model performance, and the difference in text sources affects the performance of the ADE models [31]. The datasets of existing studies could be broadly categorized into two types: (1) public datasets, such as the Distress Analysis Interview Corpus – Wizard-of-Oz (DAIC-WOZ) [32], the Audio/Visual Emotion Challenge 2014 dataset [33], the Extended Audio-Transcript Depression Corpus [34], and the Chinese Multimodal Depression Corpus [35], which are collected under standardized protocols and provide a unified benchmark for comparing algorithm performance. For example, text-only models trained on DAIC-WOZ have shown gradual improvements in recent years [36,37]. The others belong to (2) self-constructed datasets, which are more diverse compared to public datasets in collection, and are usually used for exploring model feasibility in specific populations or contexts rather than comparing model algorithms. They include transcripts of clinical interviews or therapeutic dialogues, personal essays, diaries, questionnaires, etc. (see the previous paragraph for details). These datasets tend to better reflect natural language use in real-world contexts, which enhances the practical application ability of the ADE models [38]. For example, Cheng et al [39] analyzed social media (Weibo [Sina Corporation]) texts to predict depression and found that linguistic and cultural factors can affect the model performance. While Dogrucu et al [40] used text data via smartphones for developing Moodable, an Android-based depression sensing application developed at Worcester Polytechnic Institute, and validated this application for depression estimation. Overall, these examples illustrated the importance of self-constructed datasets, while the differences in text sources and contexts may contribute to heterogeneity in model performance.

After determining the dataset, model construction and validation are the core stages of ADE studies. Previous studies mainly relied on statistical methods to manually extract linguistic features. Rude et al [41] noted that individuals with depression frequently used negations and first-person singular pronouns, while another research examined the frequency of affective words such as "anger" or "sadness" to derive participants' emotion [42]. However, with the rapid development of natural language processing (NLP), embedding-based representations (eg, Word2Vec [43], Global Vector [44]) and pretrained models (eg, BERT [Bidirectional Encoder Representations from Transformers] [45], GPT [46]) have enhanced the semantic modeling capacity of ADE models. Niu et al [47] applied graph attention networks to capture hierarchical contextual semantics, and some studies used BERT to encode transcribed speech into context-aware sentence embeddings [48,49]. In model architecture, traditional shallow models, such as support vector machines, random forests, and decision trees, were commonly used in previous works [50], while deep learning models have also performed well in the ADE tasks recently [51-53]. Dinkel et al [51] built a multitask model based on a bidirectional gated recurrent unit, achieving an F_1-score of 0.84 on DAIC-WOZ. Martinez et al [52] further improved classification performance using RoBERTa. Furthermore, the validation strategies also vary: while single-holdout testing is commonly used when datasets are large [54-56], small-sample medical texts often use repeated bootstrapping [57] or k-fold cross-validation [30,58] to ensure robustness. Overall, differences in feature representation, model architecture, and validation strategies all contribute to potential differences in model performance and generalization.

Another potential factor affecting the model performance is the quality of depression labels. Clinical diagnosis by certified psychiatrists is widely considered the gold standard of depression [1,59]. However, clinical diagnosis is costly and subject to the experience of the psychiatrists [60]. As a substitute, many text-based ADE studies used standardized self-report depression scales. Common scales include the Patient Health Questionnaire, 8 or 9 items (PHQ-8 or 9) [61], Beck Depression Inventory-II (BDI-II) [62], Zung Self-Rating Depression Scale (SDS) [63], 21-item Depression, Anxiety, and Stress Scale (DASS-21) [64], and Center for Epidemiological Studies Depression Scale (CES-D) [65]. The ADE tasks can be equal to predicting the self-reported scores. For example, Li et al [66] achieved an F_1-score of 78.3% on the DAIC-WOZ dataset using PHQ-8, and another study reported a classification accuracy of 69% using PHQ-9 labels [67]. However, structured scales demonstrate good reliability and validity, but their use is often limited by ethical constraints and annotation costs. Consequently, many studies resort to weak depression labels, such as keyword matching [68,69], sentiment analysis tools [70], or self-declared depressive status from social media [71]. These approaches facilitate large-scale data collection and exploration, but there may be deviations in the reliability and validity of practical applications. For example, keyword-based labels may ignore semantic context [72], and sentiment analysis tools may confuse sadness with clinical depression [73]. Therefore, increasing emphasis has been placed on adopting gold-standard labels for model training and validation [16,31,74]. In this study, we included only those studies that use either clinical diagnosis or the PHQ-9 scale as depression labels: Other scales, such as BDI-II and CES-D, though commonly used for screening, differ in target populations and sensitivity, which may impair cross-study comparability [75]. The potential impact of annotation source on model performance is examined in our Results section.

In recent years, several systematic reviews have investigated NLP-based depression estimation. Mao et al [31] provided a multimodal overview covering facial expression, speech, and text models, highlighting the lack of model transparency and limited practical application as critical challenges. Some reviews focus on specific methods. For example, Yao et al [16] summarized depression-related studies using social media text and found that machine learning (ML) and statistical analysis methods were more prevalent. Tahir et al [76] provided a comprehensive review of ML and deep learning approaches for ADE tasks based on social media data. Nanggala et al [77] conducted a systematic review of the performance of the transformer structure in ADE tasks and noted the competitive advantages of the text modalities. Many studies have examined the use of social media text for depression estimation, analyzing feature design and training strategies [16,78]. While these previous works provided valuable insights into this field, 2 major limitations remain: First, most reviews concentrate on specific text sources (eg, social media and public datasets), which may not generalize well to practical applications [79]. Second, annotation sources are often overlooked, and weak depression labels, such as self-declared or keyword matching, are often confused without discussing the impact of their differences from standard labels on model performance [72]. Therefore, this study uses label quality as a core inclusion criterion. We systematically assessed performance and sources of heterogeneity in text-based ADE models.

Research Aims and Structure

This review aims to evaluate the performance of text-based depression estimation models that use standard labels and to identify the potential moderators that may account for performance differences. We specifically analyzed variations in text sources, model architectures, validation strategies, annotation standards, as well as the quality of study reporting, to examine their potential influence on depression predictive performance. The structure of this review is as follows: the Introduction section outlines the background and research aims, the Methods section presents the study selection criteria, data extraction, and meta-analysis protocol, the Results section presents the outcomes of the meta-analysis, and the Discussion and Conclusion section interprets the findings and provides implications for future research and practical applications.

Methods

Study Design

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 and the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) for reporting literature searches guidelines [80] (Multimedia Appendix 1). We also registered our review on PROSPERO (CRD20251056902). There were no deviations from the registered protocol.

Information Sources and Search Strategy

To include as many studies of text-based ADE as possible at the beginning, we systematically searched 4 datasets: Scopus, IEEE Xplore, Web of Science, and PubMed. The search strategy followed the PRISMA-S reporting extension [81] and was developed based on common model architectures, text features, and outcome formats used in ADE research. The search strategy included combinations of the following terms: (“depression” OR “depressive” OR “clinical depression” OR “major depressive disorder” OR “MDD”) AND (“assessment” OR “measur*” OR “diagnos*” OR “predict*” OR “estimat*”) AND (“automated” OR “automatic” OR “AI” OR “machine learning” OR “deep learning” OR “large language model” OR “natural language processing” OR “NLP”) AND (“text” OR “linguistic analys*” OR “sentiment analys*” OR “semantic analys*” OR “lexical feature*” OR “transcribed speech” OR “written” OR “textual*”). The initial search was conducted on February 24, 2025, covering studies published from January 2014 onward, and was updated on December 3, 2025. Full database-specific search strings are provided in Multimedia Appendix 2. Additional eligible studies were identified by manually searching the reference lists of the included studies and previous reviews.

Eligibility Criteria

We included studies that fulfill the following four criteria: (1) used texts to investigate depression, (2) used ML to establish an ADE model, (3) reported the data collection process and the depression label sources, and (4) reported the metrics of the model for effect size calculation. Our exclusion criteria are the studies that (1) did not use the standard depression label (details in paragraph 5 of the introduction); (2) were not written in English; (3) published before 2024 and having fewer citations than years of publish were excluded (eg, articles published in 2021 with 3 citations, or articles published in 2019 with 5 citations would be included) [31], to ensure that full-text screening focused on studies that had been acknowledged and referenced within the field; and (4) were developed exclusively on public datasets, as these models often iterate rapidly, which could introduce potential bias.

Data Collection Process

Two independent researchers (SZ and JZ) excluded duplicates from the relevant studies retrieved, then screened and selected the relevant studies by inclusion and exclusion criteria. Disagreements were solved through discussion with a third author (CZ) when necessary. The relevant studies were evaluated through title and abstract screening. Thereafter, the remaining studies were screened through full texts to identify those for further analysis. Studies were managed using EndNote and Microsoft Excel. No automation tools were used.

Data Items

Two researchers (SZ and JZ) independently coded the included studies in Microsoft Excel. The following information was extracted from each article: (1) study characteristics (first author and publication year), (2) annotation source (clinician diagnosis and self-report scales [PHQ-9]), (3) population characteristics (sample size and positive rate), (4) modeling strategy (text sources [eg, interviews, social media, writing tasks, and diaries], text representation [eg, RoBERTa, transformer, TF-IDF, LIWC, and Empath], model architecture [deep learning vs shallow ML] and validation strategies [eg, cross-validation, hold-out, and external validation]), and (5) predictor values [sensitivity, specificity, F_1-score or data used to impute these values]. Furthermore, for studies that used different algorithms based on the same text dataset, only the model with the highest F_1-score was included in the meta-analysis. If a study reported results based on different text sources (eg, clinical transcripts vs written documents), the best-performing model from each dataset was included separately. For multimodal studies, only the performance metrics from the text modality were included.

In addition, we assessed the reporting quality of each included study using the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) tool [82]. TRIPOD was adapted to be more specific to NLP studies: The modifications were based on previously published research of NLP studies [83,84]. Two independent researchers (SZ and JZ) independently assessed each study based on TRIPOD items; disagreements were resolved through a third author (CZ).

Effect Measures

For each study, the sensitivity, specificity, and sample sizes were used to calculate the effect size (ES). Specifically, the odds ratio (OR) was first computed by sensitivity and specificity, followed by a log transformation to obtain log (OR) [85]. The SE of log (OR) was then calculated based on the sensitivity, specificity, and sample size. Log (OR) and SE (log OR) were converted into ES by using Comprehensive Meta-Analysis (version 4; Biostat Inc). To direct comparison of model performance across studies, Pearson correlation coefficient (r), transformed via Fisher z for analysis and subsequently back-transformed for interpretation, was selected as the ES indicator [86]. Furthermore, the Pearson correlation coefficient contributes to a general understanding of the relationship between text-based models' predictions and true labels. Values of r around 0.1, 0.3, and 0.5 are generally interpreted as indicating small, moderate, and large effect sizes, respectively. These thresholds have been widely applied in psychological and behavioral research, including prior reviews of depression prediction models [78,87].

Data Synthesis and Analysis

Meta-analysis was initially performed using Comprehensive Meta-Analysis (version 4; Biostat Inc). Between-study heterogeneity was assessed using the chi-square Q statistic, which tests the null hypothesis of homogeneity [88]. The extent of heterogeneity was further described using the inconsistency index (I^2), along with the between-study variance (τ^2) and its SD (τ) [89]. In addition, 95% prediction intervals were reported to reflect the expected range of effects in future studies [90]. We adopted the random-effects model to assess the heterogeneity among studies and estimate the pooled effect size from each study. Given the limited number of included studies, the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was applied to adjust the SEs [91]. Because CMA 4.0 does not implement the HKSJ procedure, we reestimated the pooled effects using the Meta package in R (version 4.3.2), and the 95% CI values reported in the main text were derived from the HKSJ-adjusted models.

To explore potential sources of heterogeneity, we performed a moderator analysis with a random effects model [92]. According to the extracted information from each study, four predefined groups were included: (1) text representation (embedding-based vs traditional features), (2) annotation source (clinician diagnosis vs self-reported scale), (3) model architecture (deep vs shallow), and (4) text source (documentation vs transcribed speech). For each subgroup, pooled ES (r) were estimated under a random-effects model, and between-group differences were assessed using the Q-test for heterogeneity. In addition, univariate meta-regression analyses were conducted to examine the potential moderating effects of 3 continuous covariates: reporting quality (TRIPOD score), sample size (log-transformed), and positive rate. Each covariate was analyzed separately with a random-effects model, and an analog R^2 indicated its explanatory power for study variance. Finally, we performed a sensitivity analysis using the leave-one-out method to assess the influence of each study on the overall results. We also conducted a cumulative meta-analysis to explore whether more recent studies have contributed to increased consistency.

Small-Study Effects

We assessed potential small-study effects, including the possibility of publication bias, by visually inspecting a funnel plot of the SE Fisher *z* [93]. Statistical evaluation of funnel plot asymmetry was conducted using the Begg–Mazumdar rank correlation test and the Egger regression test, with a *P* value <.05 indicating significant small-study effects [94]. In addition, we applied the Duval and Tweedie trim-and-fill procedure to explore the potential impact of missing studies on the pooled effect size [95].

Quality and Certainty Assessment

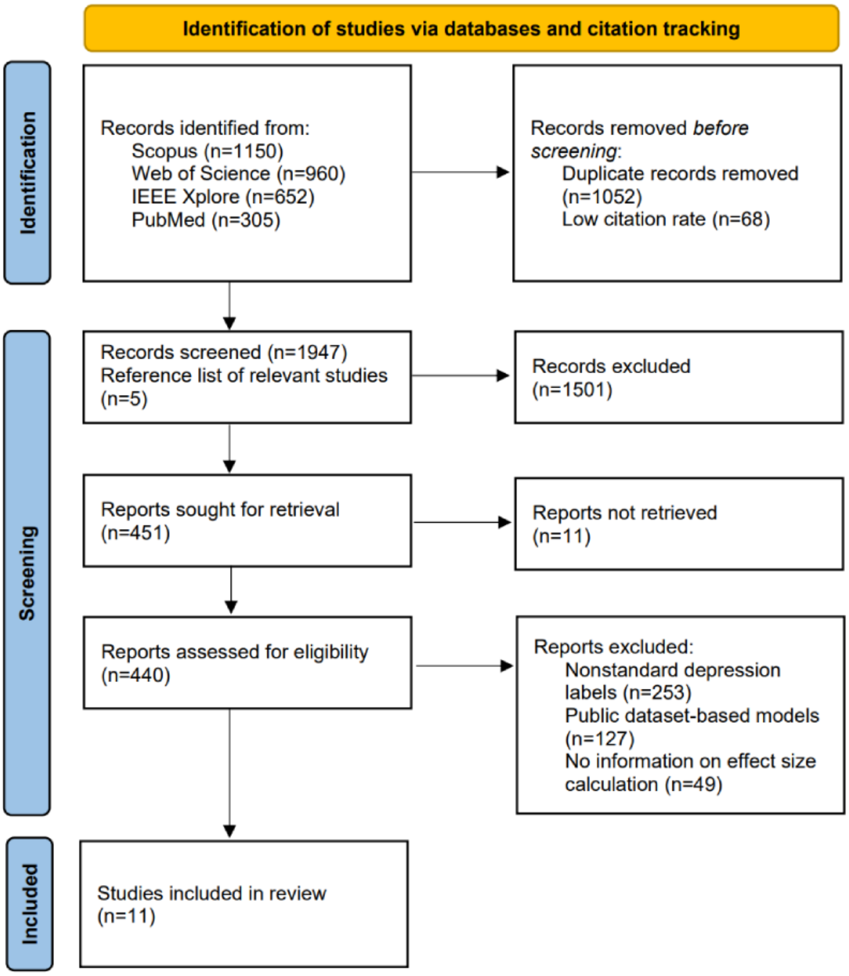
The methodological risk of bias of each included study was independently assessed by 2 authors (SZ and CZ) across five domains relevant to text-based ML research: (1) participant selection, (2) outcome labeling, (3) text acquisition and preprocessing, (4) model development and validation, and (5) reporting completeness. Each domain was rated as "low risk," "unclear risk," or "high risk" and an overall judgment was derived accordingly. The certainty of evidence for the primary pooled effect was evaluated using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework [96], considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Results

Study Selection Process

The flow diagram of studies screening and selection is presented in Figure 1. The initial database search was conducted on February 24, 2025, and was updated on December 3, 2025.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of studies included in this review.


Identification of studies via databases and citation tracking
Identification
Records identified from:
Scopus (n=1150)
Web of Science (n=960)
IEEE Xplore (n=652)
PubMed (n=305)
Records removed *before screening*:
Duplicate records removed (n=1052)
Low citation rate (n=68)
Screening
Records screened (n=1947)
Reference list of relevant studies (n=5)
Records excluded (n=1501)
Reports sought for retrieval (n=451)
Reports not retrieved (n=11)
Reports assessed for eligibility (n=440)
Reports excluded:
Nonstandard depression labels (n=253)
Public dataset-based models (n=127)
No information on effect size calculation (n=49)
Included
Studies included in review (n=11)

A total of 3067 records were identified through database searching, including Web of Science (n=960), PubMed (n=305), Scopus (n=1150), and IEEE Xplore (n=652). After removing 1052 duplicates and 68 low-citation records, 1947 records remained for title and abstract screening, together with 5 additional records identified through manual citation tracking. Among them, 1501 records were excluded because they were not relevant to text-based depression estimation, including that the model inputs were not text data and no quantitative results were reported. The remaining 451 records were included in the full-text screening, and 11 studies that were inaccessible or retracted were first excluded. Next, we excluded studies that used nonstandard depression labels (n=253), studies based on public datasets (n=127), and studies lacking sufficient information to compute effect sizes (n=49). Finally, 11 studies [17,24,26,67,97-103] were included in the meta-analysis.

Characteristics of Included Studies

A total of 11 studies were included in the final meta-analysis [17,24,26,67,97-103], contributing 15 independent text-based depression estimation models. A summary of included studies and model characteristics is presented in Table 1, and the detailed data extraction table is provided in Multimedia Appendix 3. The sample size ranged from 77 to 749 participants, with a median of 110. Positive rates (proportion of participants labeled as depressed) varied substantially across studies, ranging from 9.2% to 77.9%, reflecting differences in population selection and depression labeling strategies.

Table 1. Summary of study and model characteristics included in the meta-analysis (n=15).

Author (year)	Sample size[a]	Positive rate	TRIPOD	Text representation	Annotation source	Model architecture	Text source	Sensitivity	Specificity
Geraci et al (2017) [97]	366	0.243	17	Traditional features	Clinician diagnosis	Deep	Documentation	0.94	0.68
Ricard et al (2018) [26]	749	0.0921	13	Traditional features	Self-report scale	Shallow	Documentation	0.57	0.77
Tlachac et al (2020) [98]	162	0.3395	12	Traditional features	Self-report scale	Shallow	Documentation	0.93	0.63
Zhao et al (2021) [67]	110	0.4	11	Traditional features	Self-report scale	Shallow	Documentation	0.33	0.86
Zhao et al (2021) [67]	114	0.614	11	Traditional features	Self-report scale	Shallow	Documentation	0.81	0.53
Zhao et al (2021) [67]	341	0.463	11	Traditional features	Self-report scale	Shallow	Documentation	0.51	0.86
Shin et al (2022) [99]	166	0.5	16	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.7	0.97
Cariola et al (2022) [24]	140	0.514	9	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.68	0.66
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.83	0.85
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.88	0.93
Munthuli et al (2023) [100]	80	0.5	15	Embedding-based	Clinician diagnosis	Deep	Transcribed speech	0.9	0.83
Tlachac et al (2023) [101]	88	0.602	13	Traditional features	Self-report scale	Shallow	Documentation	0.79	0.74
Jihoon et al (2024) [102]	77	0.779	17	Traditional features	Clinician diagnosis	Shallow	Transcribed speech	0.96	0.25
Shin et al (2024) [17]	91	0.171	14	Embedding-based	Self-report scale	Deep	Documentation	0.929	0.761
Xu et al (2025) [103]	100	0.55	19	Embedding-based	Clinical diagnosis	Deep	Transcribed speech	0.955	0.933

[a] Validation strategies include k-fold cross-validation, leave-one-out, and nested validation.

Sample size refers to the total number of participants included in each model.

Positive rate refers to the proportion of participants labeled as depressed.

TRIPOD score was based on a modified 27-item checklist adapted for NLP-based studies.

Text representations include traditional features (eg, Lexical, TF-IDF, LWIC, and Emotional Dictionary) and embedding-based features (eg, BERT, Word2Vec, and RoBERTa).

Annotation source indicates whether the depression label was derived from clinician diagnosis or self-report scales (Patient Health Questionnaire-9, PHQ-9).

Model architecture refers to shallow learning (eg, SVM, Logistic Regression, and Random Forests) versus deep learning (eg, GRU, BERT, Transformers).

Text sources were classified as documentation (eg, Writing, Social media, and Messages) or transcribed speech (eg, Clinical Interviews and Psychological Conversations).

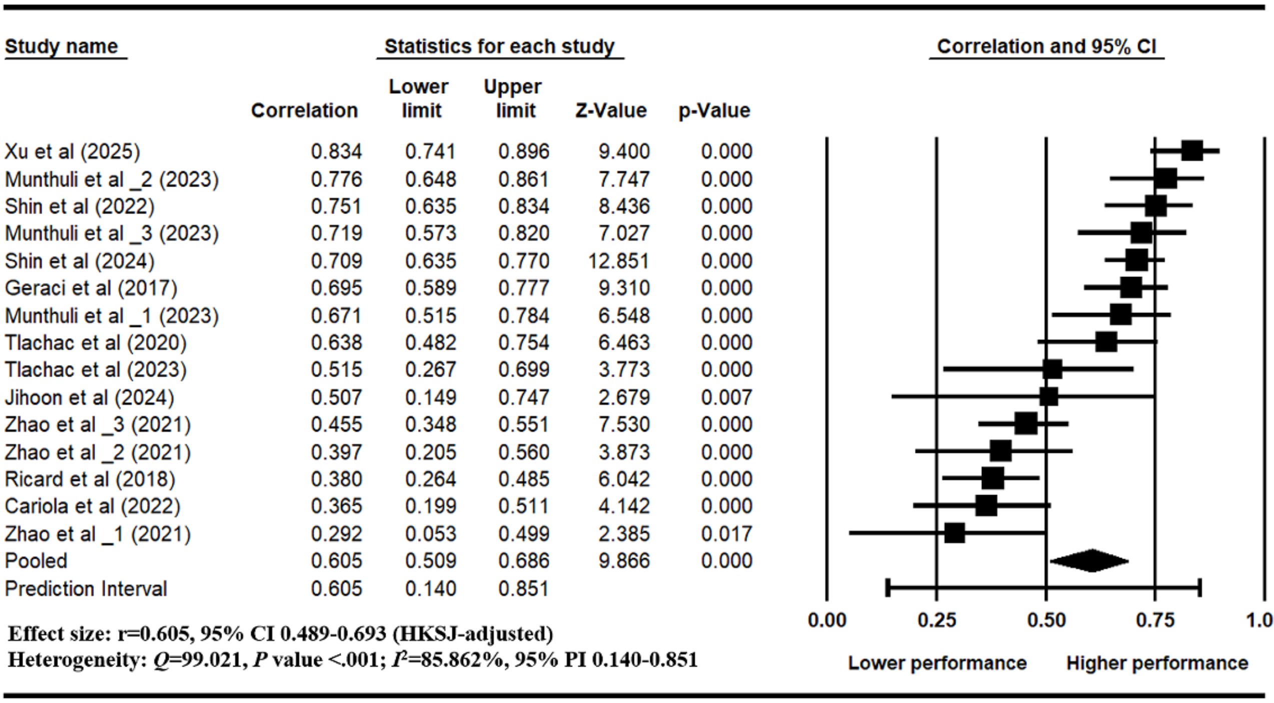
Regarding some characteristics of the modeling, 8 (53.3%) models used clinician diagnosis as depression labels, and the remaining 7 (46.7%) models relied on self-report scales such as the PHQ-9. In total, 10 (66.7%) models used traditional features to text representation, such as TF-IDF, LIWC, or emotional dictionaries, while 5 (33.3%) models used embedding-based techniques, including BERT and Word2Vec. 9 (60%) models used shallow learning models (eg, SVM, random forest, and logistic regression), and the remaining 6 (40%) models adopted deep learning architectures (eg, GRU and BERT-based classifiers). Text sources were classified as either documentation (n=8, 53.3%) or transcribed speech (n=7, 46.7%), the latter often derived from clinical interviews or therapeutic dialogues. Validation strategies also varied: 11 models implemented some form of k-fold cross-validation, while others applied nested or leave-group-out cross-validation.

The reporting quality (mean 13.9, SD 2.8; range 9-19) of all included models was assessed using an adapted 21-item TRIPOD checklist for NLP-based predictive modeling. Detailed scoring criteria and individual scores are provided in Multimedia Appendix 4.

Results of Meta-Analysis

A random-effects model was applied to synthesize the effect sizes of 15 text-based depression estimation models extracted from 11 eligible studies [17,24,26,67,97-103]. In studies reporting multiple models, each model was derived from a distinct text dataset with nonoverlapping participant samples and was therefore treated as a point estimate [104]. The pooled effect size (*r*) was 0.605 (95% CI 0.498-0.693) [87], indicating an overall high predictive ability in models using standard depression labels (Figure 2). Substantial between-model heterogeneity was observed (*Q*=99.02, *P*<.001; I^2=85.9%). The estimated between-study variance was τ^2=0.062 (τ=0.249). The 95% prediction interval ranged from 0.140 to 0.851, indicating considerable variability in the magnitude of effects expected across future studies. The forest plot revealed that each model achieved significant correlations (*P*<.05), the effect sizes varied across models (ranging from 0.292 to 0.776), further supporting the observed heterogeneity. The HKSJ-adjusted forest plot generated in R is provided in Multimedia Appendix 5. This variation highlights the influence of study-level characteristics such as model architecture, annotation source, and text representation on model performance, which were further analyzed in subsequent moderator and meta-regression analyses.

Figure 2. Forest plot presenting the pooled effect size (correlation r) of machine-learning models for text-based depression estimation trained with gold-standard labels [17,24,26,67,97-103]. For studies appearing more than once in the forest plot (eg, Zhao et al, 2021 [67]; Munthuli et al, 2023 [100]), indicate models trained and evaluated on independent text datasets derived from nonoverlapping participant samples.



Study name	Statistics for each study				
	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Xu et al (2025)	0.834	0.741	0.896	9.400	0.000
Munthuli et al _2 (2023)	0.776	0.648	0.861	7.747	0.000
Shin et al (2022)	0.751	0.635	0.834	8.436	0.000
Munthuli et al _3 (2023)	0.719	0.573	0.820	7.027	0.000
Shin et al (2024)	0.709	0.635	0.770	12.851	0.000
Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Munthuli et al _1 (2023)	0.671	0.515	0.784	6.548	0.000
Tlachac et al (2020)	0.638	0.482	0.754	6.463	0.000
Tlachac et al (2023)	0.515	0.267	0.699	3.773	0.000
Jihoon et al (2024)	0.507	0.149	0.747	2.679	0.007
Zhao et al _3 (2021)	0.455	0.348	0.551	7.530	0.000
Zhao et al _2 (2021)	0.397	0.205	0.560	3.873	0.000
Ricard et al (2018)	0.380	0.264	0.485	6.042	0.000
Cariola et al (2022)	0.365	0.199	0.511	4.142	0.000
Zhao et al _1 (2021)	0.292	0.053	0.499	2.385	0.017
Pooled	0.605	0.509	0.686	9.866	0.000
Prediction Interval	0.605	0.140	0.851		

Effect size: r=0.605, 95% CI 0.489-0.693 (HKSJ-adjusted)
Heterogeneity: *Q*=99.021, *P* value <.001; I^2=85.862%, 95% PI 0.140-0.851

Results of Subgroup Analysis for Moderators

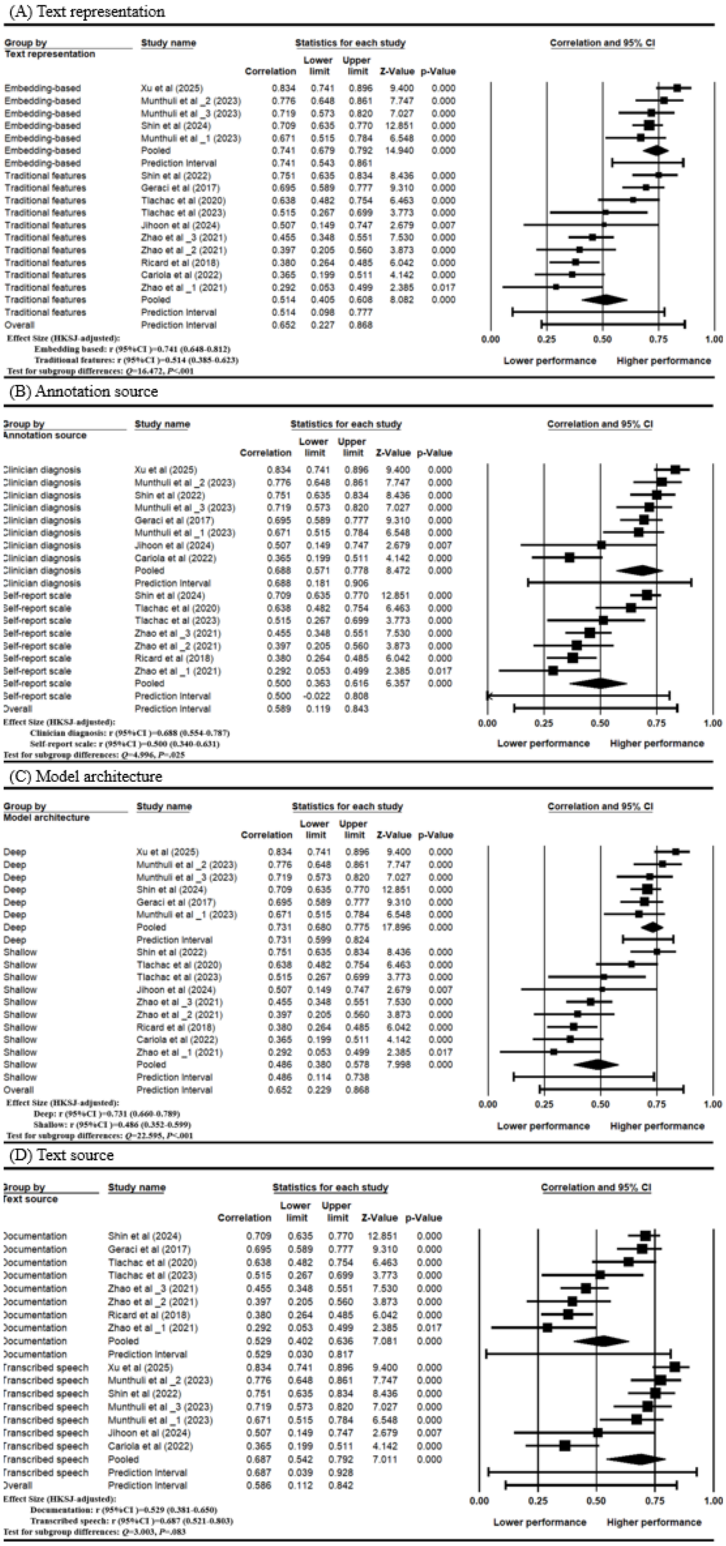
Subgroup analyses were conducted to examine the moderating effects of text representation, annotation source, model architecture, and text source (Table 2). Significant moderation effects were observed for text representation (*Q*=16.47, *P*<.001) and model architecture (*Q*=22.60, *P*<.001). Models using embedding-based features yielded higher performance (*r*=0.741, 95% CI 0.648-0.812) compared with those using traditional features (*r*=0.514, 95% CI 0.385-0.623). Deep learning architectures also outperformed shallow models (*r*=0.731, 95% CI 0.660-0.789 vs *r*=0.486, 95% CI 0.352-0.599). Annotation source showed a statistically significant moderation effect (*Q*=5.00, *P*=.03), with models using clinician diagnoses achieving higher pooled performance (*r*=0.688, 95% CI 0.554-0.787) than those using self-report scales (*r*=0.500, 95% CI 0.340-0.631). Text source did not reach statistical significance (*Q*=3.00, *P*=.08), although models using transcribed speech tended to outperform those using documentation (r=0.687 vs 0.529). Forest plots of these subgroup analyses are presented in Figure 3, and HKSJ-adjusted subgroup plots generated in R are provided in Multimedia Appendix 5.

Table 2. Subgroup analyses examining the influence of text representation, annotation source, model architecture, and text source on model performance in text-based depression estimation.

Moderators	Models, n (%)	Point estimation (95% CI)	*Q*-value (*df*)	*P* value
Text representation	15 (100)	—[a]	16.472 (1)	<.001
Embedding-based	5 (33.3)	0.741 (0.648-0.812)	—[a]	<.001
Traditional features	10 (66.7)	0.514 (0.385-0.623)	—*	<.001
Annotation source	15 (100)	—[a]	4.996 (1)	.03
Clinician diagnosis	8 (53.3)	0.688 (0.554-0.787)	—[a]	<.001
Self-report scale	7 (46.7)	0.500 (0.340-0.631)	—[a]	<.001
Model architecture	15 (100)	—[a]	22.595 (1)	<.001
Deep	6 (40)	0.731 (0.660-0.789)	—[a]	<.001
Shallow	9 (60)	0.486 (0.352-0.599)	—[a]	<.001
Text source	15 (100)	—[a]	3.003 (1)	.08
Documentation	8 (53.3)	0.529 (0.381-0.650)	—[a]	<.001
Transcribed speech	7 (46.7)	0.687 (0.521-0.803)	—[a]	<.001

[a] Not applicable.

Figure 3. Forest plots of subgroup analyses by (A) text representation, (B) annotation source, (C) model architecture, and (D) text source generated using CMA4.0 [17, 24, 26, 67, 97, 98, 99, 100, 101, 102, 103]. High-resolution versions of all subgroup forest plots are provided in Multimedia Appendix 5.



(A) Text representation

Group by Text representation	Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Embedding-based	Xu et al (2025)	0.834	0.741	0.896	9.400	0.000
Embedding-based	Munthuli et al _2 (2023)	0.776	0.648	0.861	7.747	0.000
Embedding-based	Munthuli et al _3 (2023)	0.719	0.573	0.820	7.027	0.000
Embedding-based	Shin et al (2024)	0.709	0.635	0.770	12.851	0.000
Embedding-based	Munthuli et al _1 (2023)	0.671	0.515	0.784	6.548	0.000
Embedding-based	Pooled	0.741	0.679	0.792	14.940	0.000
Embedding-based	Prediction Interval	0.741	0.543	0.861		
Traditional features	Shin et al (2022)	0.751	0.635	0.834	8.436	0.000
Traditional features	Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Traditional features	Tlachac et al (2020)	0.638	0.482	0.754	6.463	0.000
Traditional features	Tlachac et al (2023)	0.515	0.267	0.699	3.773	0.000
Traditional features	Jihoon et al (2024)	0.507	0.149	0.747	2.679	0.007
Traditional features	Zhao et al _3 (2021)	0.455	0.348	0.551	7.530	0.000
Traditional features	Zhao et al _2 (2021)	0.397	0.205	0.560	3.873	0.000
Traditional features	Ricard et al (2018)	0.380	0.264	0.485	6.042	0.000
Traditional features	Cariola et al (2022)	0.365	0.199	0.511	4.142	0.000
Traditional features	Zhao et al _1 (2021)	0.292	0.053	0.499	2.385	0.017
Traditional features	Pooled	0.514	0.405	0.608	8.082	0.000
Traditional features	Prediction Interval	0.514	0.098	0.777		
Overall	Prediction Interval	0.652	0.227	0.868		

Effect Size (HKSJ-adjusted):
Embedding based: r (95%CI)=0.741 (0.648-0.812)
Traditional features: r (95%CI)=0.514 (0.385-0.623)
Test for subgroup differences: *Q*=16.472, *P*<.001

(B) Annotation source

Group by Annotation source	Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Clinician diagnosis	Xu et al (2025)	0.834	0.741	0.896	9.400	0.000
Clinician diagnosis	Munthuli et al _2 (2023)	0.776	0.648	0.861	7.747	0.000
Clinician diagnosis	Shin et al (2022)	0.751	0.635	0.834	8.436	0.000
Clinician diagnosis	Munthuli et al _3 (2023)	0.719	0.573	0.820	7.027	0.000
Clinician diagnosis	Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Clinician diagnosis	Munthuli et al _1 (2023)	0.671	0.515	0.784	6.548	0.000
Clinician diagnosis	Jihoon et al (2024)	0.507	0.149	0.747	2.679	0.007
Clinician diagnosis	Cariola et al (2022)	0.365	0.199	0.511	4.142	0.000
Clinician diagnosis	Pooled	0.688	0.571	0.778	8.472	0.000
Clinician diagnosis	Prediction Interval	0.688	0.181	0.906		
Self-report scale	Shin et al (2024)	0.709	0.635	0.770	12.851	0.000
Self-report scale	Tlachac et al (2020)	0.638	0.482	0.754	6.463	0.000
Self-report scale	Tlachac et al (2023)	0.515	0.267	0.699	3.773	0.000
Self-report scale	Zhao et al _3 (2021)	0.455	0.348	0.551	7.530	0.000
Self-report scale	Zhao et al _2 (2021)	0.397	0.205	0.560	3.873	0.000
Self-report scale	Ricard et al (2018)	0.380	0.264	0.485	6.042	0.000
Self-report scale	Zhao et al _1 (2021)	0.292	0.053	0.499	2.385	0.017
Self-report scale	Pooled	0.500	0.363	0.616	6.357	0.000
Self-report scale	Prediction Interval	0.500	-0.022	0.808		
Overall	Prediction Interval	0.589	0.119	0.843		

Effect Size (HKSJ-adjusted):
Clinician diagnosis: r (95%CI)=0.688 (0.554-0.787)
Self-report scale: r (95%CI)=0.500 (0.340-0.631)
Test for subgroup differences: *Q*=4.996, *P*=.025

(C) Model architecture

Group by Model architecture	Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Deep	Xu et al (2025)	0.834	0.741	0.896	9.400	0.000
Deep	Munthuli et al _2 (2023)	0.776	0.648	0.861	7.747	0.000
Deep	Munthuli et al _3 (2023)	0.719	0.573	0.820	7.027	0.000
Deep	Shin et al (2024)	0.709	0.635	0.770	12.851	0.000
Deep	Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Deep	Munthuli et al _1 (2023)	0.671	0.515	0.784	6.548	0.000
Deep	Pooled	0.731	0.680	0.775	17.896	0.000
Deep	Prediction Interval	0.731	0.599	0.824		
Shallow	Shin et al (2022)	0.751	0.635	0.834	8.436	0.000
Shallow	Tlachac et al (2020)	0.638	0.482	0.754	6.463	0.000
Shallow	Tlachac et al (2023)	0.515	0.267	0.699	3.773	0.000
Shallow	Jihoon et al (2024)	0.507	0.149	0.747	2.679	0.007
Shallow	Zhao et al _3 (2021)	0.455	0.348	0.551	7.530	0.000
Shallow	Zhao et al _2 (2021)	0.397	0.205	0.560	3.873	0.000
Shallow	Ricard et al (2018)	0.380	0.264	0.485	6.042	0.000
Shallow	Cariola et al (2022)	0.365	0.199	0.511	4.142	0.000
Shallow	Zhao et al _1 (2021)	0.292	0.053	0.499	2.385	0.017
Shallow	Pooled	0.486	0.380	0.578	7.998	0.000
Shallow	Prediction Interval	0.486	0.114	0.738		
Overall	Prediction Interval	0.652	0.229	0.868		

Effect Size (HKSJ-adjusted):
Deep: r (95%CI)=0.731 (0.660-0.789)
Shallow: r (95%CI)=0.486 (0.352-0.599)
Test for subgroup differences: *Q*=22.595, *P*<.001

(D) Text source

Group by Text source	Study name	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Documentation	Shin et al (2024)	0.709	0.635	0.770	12.851	0.000
Documentation	Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Documentation	Tlachac et al (2020)	0.638	0.482	0.754	6.463	0.000
Documentation	Tlachac et al (2023)	0.515	0.267	0.699	3.773	0.000
Documentation	Zhao et al _3 (2021)	0.455	0.348	0.551	7.530	0.000
Documentation	Zhao et al _2 (2021)	0.397	0.205	0.560	3.873	0.000
Documentation	Ricard et al (2018)	0.380	0.264	0.485	6.042	0.000
Documentation	Zhao et al _1 (2021)	0.292	0.053	0.499	2.385	0.017
Documentation	Pooled	0.529	0.402	0.636	7.081	0.000
Documentation	Prediction Interval	0.529	0.030	0.817		
Transcribed speech	Xu et al (2025)	0.834	0.741	0.896	9.400	0.000
Transcribed speech	Munthuli et al _2 (2023)	0.776	0.648	0.861	7.747	0.000
Transcribed speech	Shin et al (2022)	0.751	0.635	0.834	8.436	0.000
Transcribed speech	Munthuli et al _3 (2023)	0.719	0.573	0.820	7.027	0.000
Transcribed speech	Munthuli et al _1 (2023)	0.671	0.515	0.784	6.548	0.000
Transcribed speech	Jihoon et al (2024)	0.507	0.149	0.747	2.679	0.007
Transcribed speech	Cariola et al (2022)	0.365	0.199	0.511	4.142	0.000
Transcribed speech	Pooled	0.687	0.542	0.792	7.011	0.000
Transcribed speech	Prediction Interval	0.687	0.039	0.928		
Overall	Prediction Interval	0.586	0.112	0.842		

Effect Size (HKSJ-adjusted):
Documentation: r (95%CI)=0.529 (0.381-0.650)
Transcribed speech: r (95%CI)=0.687 (0.521-0.803)
Test for subgroup differences: *Q*=3.003, *P*=.083

Univariable meta-regression analysis was used to test 3 continuous moderators: TRIPOD score, the positive rates, and the log-transformed sample size (log n). All meta-regressions were conducted using Fisher *z*-transformed correlation coefficients under a random-effects model. As shown in Table 3, only the TRIPOD score was significantly associated with model performance (β=0.085, 95% CI 0.050-0.119; *P*<.001), indicating a positive association between reporting quality and effect size. This model explained 71% of the between-study variance (R^2 analog=0.71). In contrast, neither the positive rate (β=–0.027, 95% CI –0.830 to 0.884; *P*=.95) nor the sample size (β=–0.012, 95% CI –0.214 to 0.190; *P*=.91) showed significant associations with effect size. Corresponding regression plots are included in Multimedia Appendix 6 (CMA-based) and Multimedia Appendix 5 (HKSJ-adjusted).

Table 3. Univariable meta-regression assessing reporting quality (TRIPOD), positive rate, and sample size as moderators of model performance.

Moderators	N	β (95% CI)	SE	R^2	*P* value
TRIPOD[a] NLP[b]	15	0.085 (0.050-0.119)	0.018	0.71	<.001
Positive Rate	15	–0.027 (–0.830 to 0.884)	0.437	<0.2	.95
Log n	15	–0.012 (–0.214-0.190)	0.103	<0.2	.91

[a] TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

[b] NLP: natural language processing.

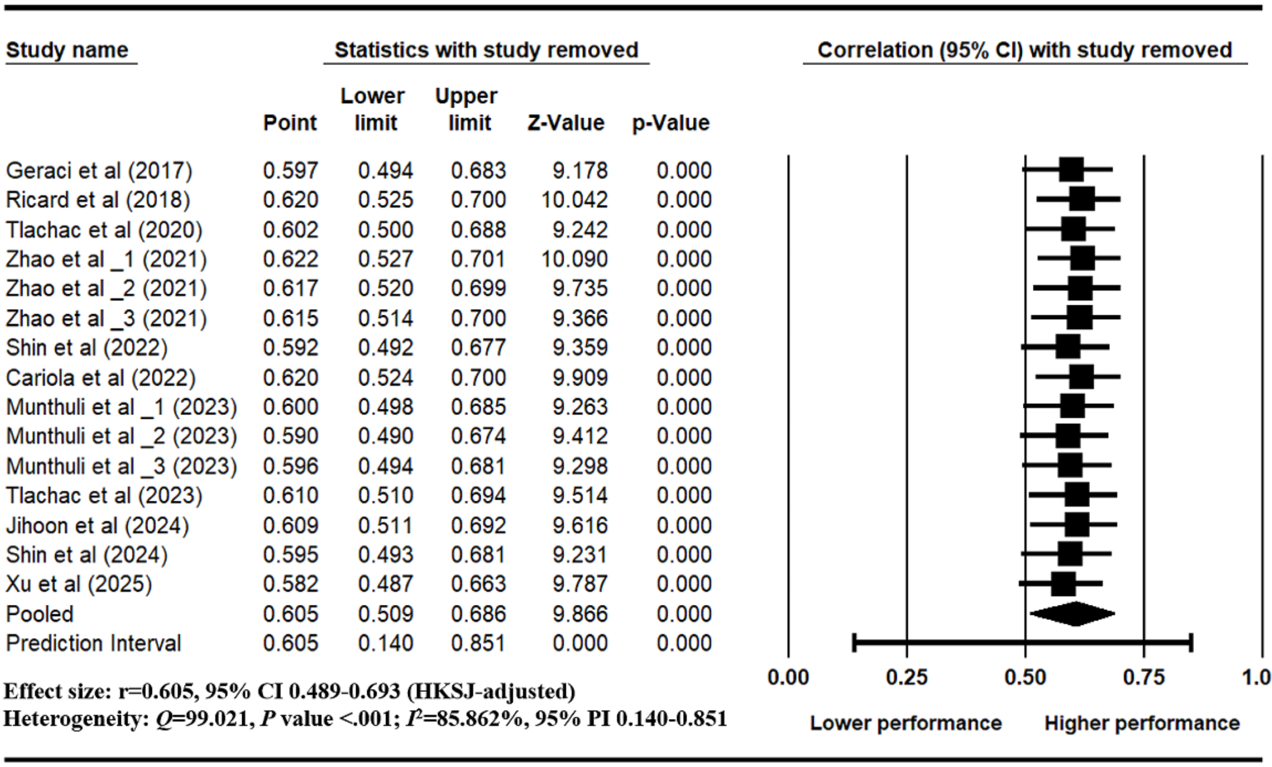
Risk of Bias and Certainty Assessment

Risk-of-bias assessment showed that most included models demonstrated low methodological concerns. Of the 15 models, 7 were rated as low risk, 6 as unclear risk, and 2 as high risk, with the latter typically related to concerns in outcome measurement or reporting (Multimedia Appendix 7). The overall certainty of evidence for the primary pooled effect was rated as moderate under the GRADE framework. This rating was driven mainly by the substantial heterogeneity observed across models, while concerns regarding imprecision and publication bias were minimal. Detailed domain ratings and justification for GRADE decisions are provided in Supplementary Table S4 (Multimedia Appendix 7).

Sensitivity Analysis and Cumulative Analysis

Leave-one-out analysis indicated that the pooled effect size remained stable across all iterations (*r*=0.605, 95% CI 0.498-0.693), suggesting that no single study had a disproportionate impact on the overall result (Figure 4). Cumulative meta-analysis showed that the effect estimates gradually stabilized over time, indicating consistency across publication years (Figure 5).

Figure 4. Leave-one-out sensitivity analysis [17,24,26,67,97-103].

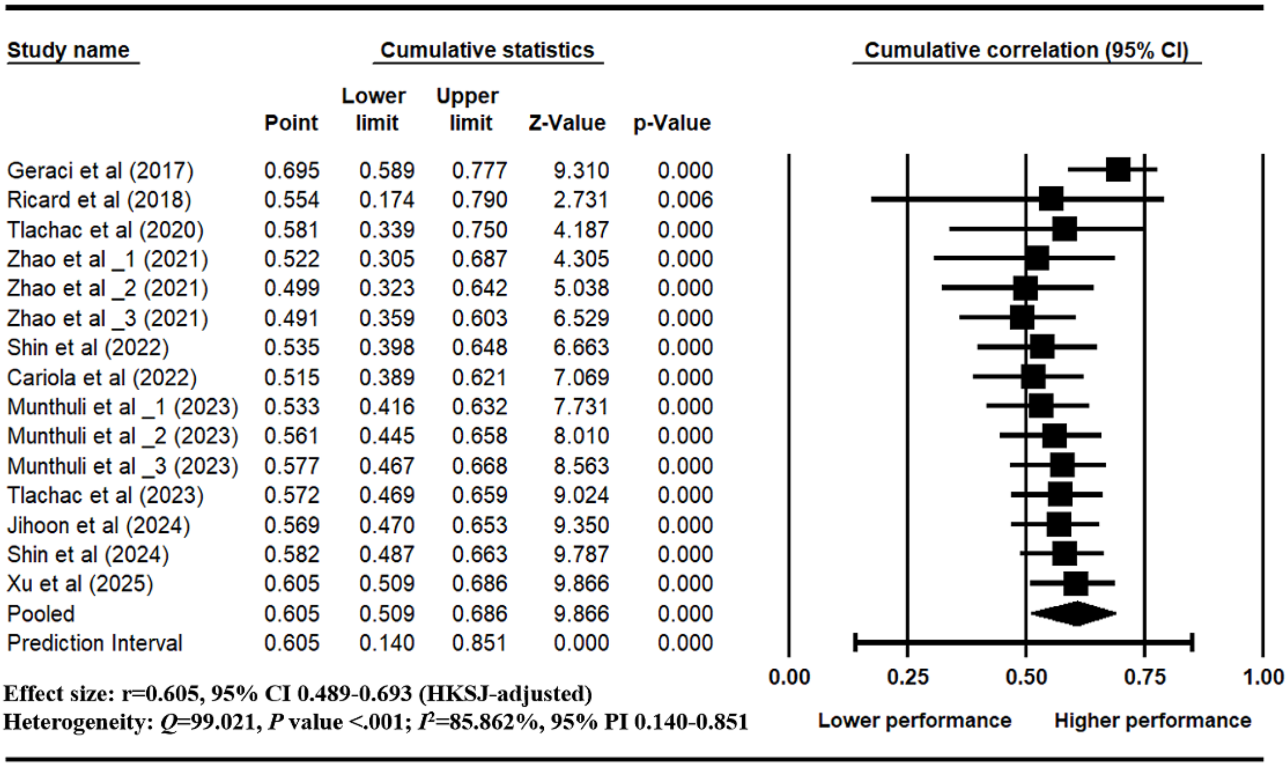


Study name	Statistics with study removed				
	Point	Lower limit	Upper limit	Z-Value	p-Value
Geraci et al (2017)	0.597	0.494	0.683	9.178	0.000
Ricard et al (2018)	0.620	0.525	0.700	10.042	0.000
Tlachac et al (2020)	0.602	0.500	0.688	9.242	0.000
Zhao et al _1 (2021)	0.622	0.527	0.701	10.090	0.000
Zhao et al _2 (2021)	0.617	0.520	0.699	9.735	0.000
Zhao et al _3 (2021)	0.615	0.514	0.700	9.366	0.000
Shin et al (2022)	0.592	0.492	0.677	9.359	0.000
Cariola et al (2022)	0.620	0.524	0.700	9.909	0.000
Munthuli et al _1 (2023)	0.600	0.498	0.685	9.263	0.000
Munthuli et al _2 (2023)	0.590	0.490	0.674	9.412	0.000
Munthuli et al _3 (2023)	0.596	0.494	0.681	9.298	0.000
Tlachac et al (2023)	0.610	0.510	0.694	9.514	0.000
Jihoon et al (2024)	0.609	0.511	0.692	9.616	0.000
Shin et al (2024)	0.595	0.493	0.681	9.231	0.000
Xu et al (2025)	0.582	0.487	0.663	9.787	0.000
Pooled	0.605	0.509	0.686	9.866	0.000
Prediction Interval	0.605	0.140	0.851	0.000	0.000

Effect size: r=0.605, 95% CI 0.489-0.693 (HKSJ-adjusted)

Heterogeneity: *Q*=99.021, *P* value <.001; I^2=85.862%, 95% PI 0.140-0.851

Figure 5. Cumulative meta-analysis [17,24,26,67,97-103].



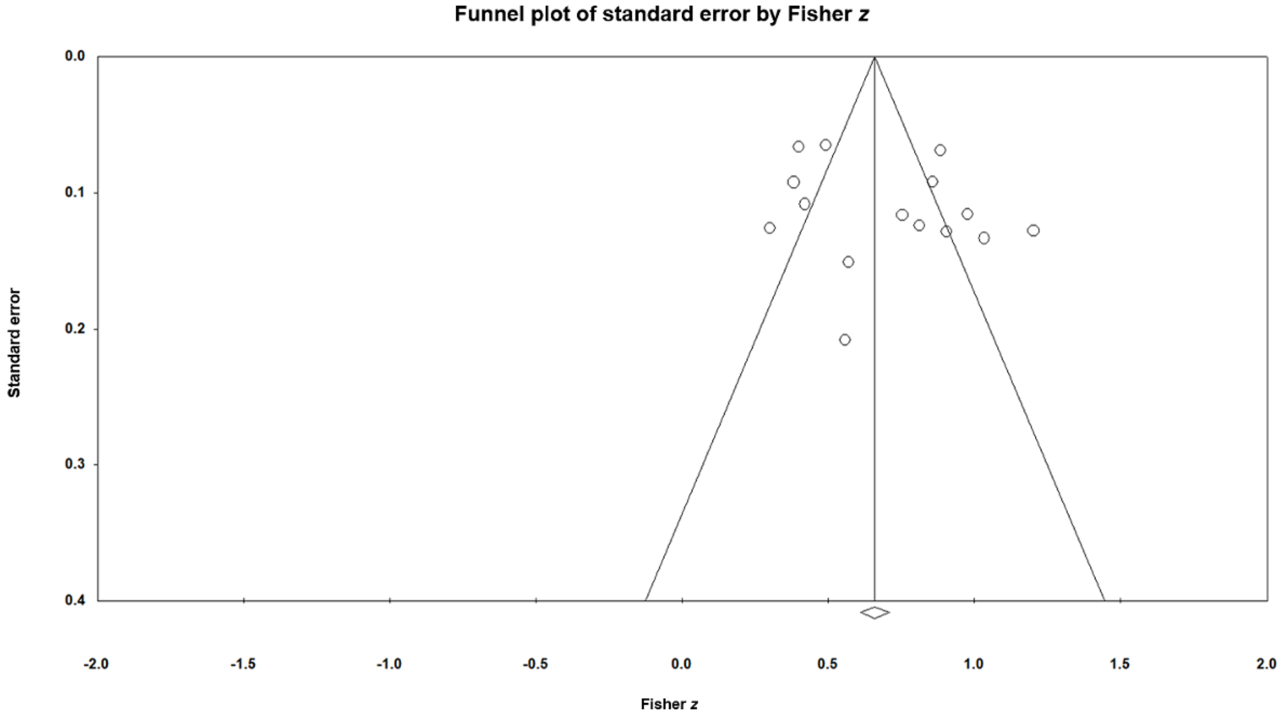
Study name	Point	Lower limit	Upper limit	Z-Value	p-Value
Geraci et al (2017)	0.695	0.589	0.777	9.310	0.000
Ricard et al (2018)	0.554	0.174	0.790	2.731	0.006
Tlachac et al (2020)	0.581	0.339	0.750	4.187	0.000
Zhao et al _1 (2021)	0.522	0.305	0.687	4.305	0.000
Zhao et al _2 (2021)	0.499	0.323	0.642	5.038	0.000
Zhao et al _3 (2021)	0.491	0.359	0.603	6.529	0.000
Shin et al (2022)	0.535	0.398	0.648	6.663	0.000
Cariola et al (2022)	0.515	0.389	0.621	7.069	0.000
Munthuli et al _1 (2023)	0.533	0.416	0.632	7.731	0.000
Munthuli et al _2 (2023)	0.561	0.445	0.658	8.010	0.000
Munthuli et al _3 (2023)	0.577	0.467	0.668	8.563	0.000
Tlachac et al (2023)	0.572	0.469	0.659	9.024	0.000
Jihoon et al (2024)	0.569	0.470	0.653	9.350	0.000
Shin et al (2024)	0.582	0.487	0.663	9.787	0.000
Xu et al (2025)	0.605	0.509	0.686	9.866	0.000
Pooled	0.605	0.509	0.686	9.866	0.000
Prediction Interval	0.605	0.140	0.851	0.000	0.000

Effect size: r=0.605, 95% CI 0.489-0.693 (HKSJ-adjusted)
Heterogeneity: *Q*=99.021, *P* value <.001; I^2=85.862%, 95% PI 0.140-0.851

Small-Study Effects

The funnel plot (Figure 6) showed a generally symmetrical distribution of effect sizes. Both the Begg–Mazumdar test (Kendall τ=0.17143, *P*=.37) and the Egger regression test (t_{14}=1.13401, 2-tailed *P*=.28) were nonsignificant, indicating no evidence of small-study effects. The trim-and-fill procedure did not impute any additional studies, and the adjusted pooled effect was identical to the original estimate.

Figure 6. Funnel plot.



Discussion

Principal Findings

This study evaluated the performance of text-based depression estimation models that were trained and validated using standard depression labels. Our findings indicate that text-based ADE models demonstrate promising predictive performance. A previous review study mentioned the importance of a uniform standard label when evaluating the consistency of model performance across studies [78]. In this study, we focused on the studies using clinical diagnoses or PHQ-9 and excluded weak depression labels datasets (eg, keyword matching and self-declared), thereby enhancing the clinical reliability of the results [72,73]. These findings support the potential of text, especially that in transcribed speech and personal documentation, as informative signals for depression detection [23-25]. Nonetheless, the high between-study heterogeneity implies the differences in study-level characteristics, including modeling approaches, corpus sources, and reporting quality.

For the text representation and model architecture, the results of subgroup analyses showed that models using embedding-based text representations achieved higher performance than those using traditional features, and deep learning architectures outperformed shallow models. Consistent results have been reported on commonly used public benchmarks [105,106]. On the DAIC-WOZ dataset, the high-performing models often use embedding-based representation and deep model architecture [37,107]. This result aligns with the development of context-aware models like BERT [45] and GPT [46], which have superior semantic representation capabilities. A previous review study on emotion detection also reported similar conclusions [22]. However, the traditional linguistic features remain critical in the field of ADEs [16]. Linguistic features can be extracted from depressive texts by topic modeling techniques [108], which provide insights into underlying cognitive and emotional states and contribute to clinical understanding [109,110]. Overall, embedding-based and deep model architectures are superior in prediction performance, and the traditional features and shallow models remain valuable for enhancing interpretability and depression understanding.

For the annotation sources, the results of subgroup analyses indicated that models trained with clinician diagnoses also outperformed better than those relying on self-report scales. This supported the view expressed by Mao et al [31] that self-reported measures may not align with clinical diagnoses, potentially leading to inconsistency in model performance. Another study also highlighted the heterogeneity of depression estimation caused by subjectivity [111]. In addition, the results indicate that while models trained with clinically validated scales like PHQ-9 remain usable, their performance is generally lower than models trained with clinician diagnoses. For the text sources, models using transcribed speech were slightly higher than those using personal documentation, but this difference did not reach a statistically meaningful level. This may reflect the more information and affective cues embedded in spoken language [38], but some studies have pointed out that there are also noises such as stop words that are not directly related to emotion [112,113]. Overall, these results suggest that the annotation source demonstrated a determinative impact on model performance, with clinician diagnoses outperforming PHQ-9, whereas the text source showed only directional effects. We suggest constructing the ADE models based on the gold standard and natural environment-related text data. However, when clinician diagnoses are unavailable, validated self-report scales represent feasible alternatives.

It is worth noting that there is a significant positive correlation between the model report quality (TRIPOD score) and the model performance. This indicates that studies with higher report quality exhibit better performance in the ADE model. The TRIPOD checklist has been previously adopted in other automatic estimation domains [114,115]. This result further emphasized the necessity for comprehensive reporting of study-level characteristics, including modeling approaches and corpus sources, to enhance reliability and validity in ADE models [82]. In contrast, sample size and positive rates were not significantly associated with model performance. It is consistent with the previous review studies' results [76,78], suggesting that in the ADE tasks, the data quality, annotation, and methodological transparency may be more crucial than the quantity of the samples. This finding underscores the importance of rigorous methodological reporting for ADE studies.

Taken together, the findings of this meta-analysis highlight several methodological and reporting considerations that are relevant for the development and evaluation of text-based ADEs. Specifically, models using embedding-based features and deep architectures are superior in predictive performance, suggesting that future studies could prioritize richer text representations combined with more sophisticated model architectures. Adopting standardized depression labels, particularly clinician diagnoses or widely validated scales, may further facilitate comparability across studies. Notably, we found that there is a significant positive correlation between the TRIPOD score and the model performance. This indicates that improving adherence to established reporting guidelines could enhance the transparency and reproducibility of ADE models. The overall certainty of evidence was rated as moderate using the GRADE approach, reflecting acceptable confidence in the pooled findings despite substantial between-study heterogeneity.

Strengths and Limitations

The primary strength of this review lies in its strict focus on text-based ADE models trained and validated using standard depression labels. By excluding studies based on weak or inconsistent annotation sources, such as keyword matching or self-declared labels, we reduced label-related heterogeneity and improved the clinical interpretability of the synthesized findings. In addition, this meta-analysis systematically examined multiple methodological sources of heterogeneity, including text representation, model architecture, annotation source, and reporting quality, using a unified analytical framework combining subgroup analyses and meta-regression.

Nevertheless, this study also has some limitations. First, our inclusion criteria strictly require the use of standard depression labels. Although this enhanced the reliability of the labels, it also reduced the number of eligible studies (57.5% of the literature was excluded in the full-text assessment). Some of the latest studies were also not included, including large language models [116] and the prompt learning strategy [36]. Studies on conversational agents with NLP-based automatic depression detection were also not included [117]. Second, although the Egger test indicated no significant publication bias, the exclusion of non-English and low-visibility studies raises the possibility of study omission. Future work may cautiously broaden the inclusion scope to better capture emerging methodological trends, including hybrid approaches that integrate text with other modalities, with the aim of improving the practical applicability of ADE models.

Implications

The results of this review indicate that multiple methodological factors, including text source, text representation, model architecture, annotation source, and reporting quality, influence the performance of text-based depression estimation models. These findings not only provide methodological guidance for future text-based ADE research but also offer important insights into the sources of variation observed in existing studies. To our knowledge, this study is among the first systematic reviews to explicitly restrict inclusion to models trained and validated using standard depression labels. Previous systematic reviews in this field have largely focused on aspects such as model architecture or the use of social media data, with little systematic attention given to the impact of depression label quality. The inclusion of weak depression labels in prior research may have affected the reliability of reported results and further limited their practical applicability. By treating standard depression labels as a core inclusion criterion, this review demonstrates that annotation source is one of the key determinants of model performance. This observation suggests that performance differences across studies may reflect variations in labeling standards and reporting quality, rather than differences in intrinsic model capability alone. More broadly, the present findings support a shift in text-based ADE research toward evaluation frameworks that emphasize label rigor, transparent reporting, and methodological consistency, thereby enabling more meaningful comparison across studies and ultimately facilitating the clinical translation of ADE models.

Conclusions

In summary, this systematic review and meta-analysis surveyed the last decade of text-based ADE using standard depression labels. The overall pooled effect size (*r*=0.605) suggests that ADE models perform well and have the potential for practical application. However, substantial heterogeneity across studies was observed. Models using embedding-based features and deep architectures generally achieved superior performance, whereas the influence of annotation source and text source was comparatively limited. Models trained with clinical diagnoses and transcribed speech tended to outperform those using self-report scales and documentation, though the difference was not statistically significant. Moreover, reporting quality, as assessed by TRIPOD, was positively associated with performance, highlighting the need for transparent reporting. Future studies should consider richer text representations, standardized labels, integration with other modalities, and transparent reporting to enhance the reliability and practical applicability of ADE models. In our future work, we will also aim to capture emerging approaches and cautiously broaden the inclusion scope, thereby providing a more comprehensive view of the ADE field.

Acknowledgments

The authors would like to thank the editor and reviewers for their constructive comments, which strengthened the quality of this manuscript. The authors declare the use of generative artificial intelligence (GenAI) to assist with language editing, proofreading, and improving clarity of expression. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: proofreading and editing, and adapting and adjusting emotional tone. The GenAI tool used was ChatGPT (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

All data and search strategies are available in the additional files accompanying this article. Further details in this study are available from the corresponding author upon reasonable request.

Funding

This study was funded by the Second Batch of the 2023 Ministry of Education of the People’s Republic of China Industry–University Collaborative Education Program (Program number 230905329052215; Kingfar-CES “Human Factors and Ergonomics” Program); the Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology (Grant number 2024B1212010002); the Startup Fund of Southern University of Science and Technology (grant numbers Y01966117 and Y01966223).

Authors' Contributions

SZ contributed to the conceptualization, literature screening, data extraction, meta-analysis, and writing of the original draft. CZ contributed to the refinement of the review methodology, quality assessment, preparation of supplementary materials, and revision of the manuscript. JZ provided guidance on methodological design, supervised the review process, and contributed to conceptualization, data extraction, and manuscript revision. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA 2020 checklist.
[DOCX File , 274 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategies.
[DOCX File , 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Detailed data extraction table of included studies.
[DOCX File , 27 KB-Multimedia Appendix 3]

Multimedia Appendix 4

TRIPOD-AI_checklist.
[DOCX File , 86 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Hartung–Knapp–Sidik–Jonkman corrected analyses using the R meta package.
[DOCX File , 1457 KB-Multimedia Appendix 5]

Multimedia Appendix 6

CMA4.0-based subgroup and meta-regression analyses.
[DOCX File , 1513 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Risk of Bias and GRADE Assessment Tables.
[DOCX File , 141 KB-Multimedia Appendix 7]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. In: Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). United States. American Psychiatric Association Publishing DSM-5, 5th ed; 2013.
2. Tolentino JC, Schmidt SL. DSM-5 criteria and depression severity: implications for clinical practice. Front Psychiatry. 2018;9:450. [FREE Full text] [doi: 10.3389/fpsyt.2018.00450] [Medline: 30333763]
3. Hawton K, Casañas I C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord. 2013;147(1-3):17-28. [FREE Full text] [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
4. Gabbard GO, Crisp-Han H. The early career psychiatrist and the psychotherapeutic identity. Academic Psychiatry. 2017;41(1):30-34. [doi: 10.1007/s40596-016-0627-7] [Medline: 27882522]
5. Beck A, Alford B. Depression: Causes and Treatment, 2nd ed. US. University of Pennsylvania Press; 2009.
6. Breedvelt JJF, Zamperoni V, South E, Uphoff EP, Gilbody S, Bockting CLH, et al. et al. A systematic review of mental health measurement scales for evaluating the effects of mental health prevention interventions. Eur J Public Health. 2020;30(3):539-545. [doi: 10.1093/eurpub/ckz233] [Medline: 32236548]
7. Maust D, Cristancho M, Gray L, Rushing S, Tjoa C, Thase M. Aminoff MJ, Boller F, Swaab D, editors. Chapter 13 - Psychiatric rating scales. Cambridge University Press, Springer. Indian Psychiatric Society; 2012.
8. Nuevo R, Lehtinen V, Reyna-Liberato PM, Ayuso-Mateos JL. Usefulness of the beck depression inventory as a screening method for depression among the general population of finland. Scand J Public Health. 2009;37(1):28-34. [doi: 10.1177/1403494808097169] [Medline: 19141552]
9. Stockings E, Degenhardt L, Lee Y, Mihalopoulos C, Liu A, Hobbs M, et al. et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. Journal of Affective Disorders. 2015;174:447-463. [FREE Full text] [doi: 10.1016/j.jad.2014.11.061] [Medline: 25553406]

10. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: a meta-analysis. Canadian Medical Association Journal. 2008;178(8):997-1003. [FREE Full text] [doi: 10.1503/cmaj.070281] [Medline: 18390942]
11. Ren Y, Yang H, Browning C, Thomas S, Liu M. Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review. Med Sci Monit. 2015;21:646-653. [FREE Full text] [doi: 10.12659/MSM.892537] [Medline: 25725615]
12. Ray A, Bhardwaj A, Malik Y, Singh S, Gupta R. Artificial intelligence and Psychiatry: an overview. Asian J Psychiatr. 2022;70:103021. [FREE Full text] [doi: 10.1016/j.ajp.2022.103021] [Medline: 35219978]
13. Sun J. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian Journal of Psychiatry. 2023:103705. [FREE Full text] [doi: 10.1016/j.ajp.2023.103705]
14. Wu C, Kuo C, Su C, Wang S. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. Journal of Affective Disorders. 2020:617-623. [FREE Full text] [doi: 10.1016/j.jad.2019.09.044]
15. Yao X, Yu G, Tang J, Zhang J. Extracting depressive symptoms and their associations from an online depression community. Computers in Human Behavior. Jul 2021;120:106734. [FREE Full text] [doi: 10.1016/j.chb.2021.106734]
16. Yao S, Wang F, Chen J, Lu Q. Utilizing health-related text on social media for depression research: themes and methods. Library Hi Tech. 2023;43(1):274-294. [doi: 10.1108/lht-02-2023-0076]
17. Shin D, Kim H, Lee S, Cho Y, Jung W. Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: instrument validation study. J Med Internet Res. 2024;26:e54617. [FREE Full text] [doi: 10.2196/54617] [Medline: 39292502]
18. Haque A, Guo M, Miner A, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv. Nov 21, 2018. [doi: 10.48550/arXiv.1811.08592]
19. Zhang B, Wei D, Yan G, Lei T, Cai H, Yang Z. Feature-level fusion based on spatial-temporal of pervasive EEG for depression recognition. Comput Methods Programs Biomed. 2022;226:107113. [doi: 10.1016/j.cmpb.2022.107113] [Medline: 36103735]
20. Zhang Z, Huang P, Li S, Liu Z, Zhang J, Li Y, et al. et al. Neural mechanisms underlying the processing of emotional stimuli in individuals with depression: an ALE meta-analysis study. Psychiatry Res. 2022;313:114598. [doi: 10.1016/j.psychres.2022.114598] [Medline: 35544984]
21. Berger J, Packard G. Using natural language processing to understand people and culture. American Psychologist. 2022;77(4):525-537. [doi: 10.1037/amp0000882] [Medline: 34914405]
22. Maruf AA, Khanam F, Haque MM, Jiyad ZM, Mridha MF, Aung Z. Challenges and opportunities of text-based emotion detection: a survey. IEEE Access. 2024;12:18416-18450. [doi: 10.1109/access.2024.3356357]
23. Jackson J, Watts J, List J, Drabble R, Lindquist K. From text to thought: how analyzing language can advance psychological science. PsyArXiv. Jul 01, 2020. [doi: 10.31234/osf.io/nws35]
24. Cariola L, Hinduja S, Bilalpur M, Sheeber L, Allen N, Morency L, et al. et al. Language use in mother-adolescent dyadic interaction: preliminary results. Int Conf Affect Comput Intell Interact Workshops. 2022;2022. [doi: 10.1109/acii55700.2022.9953886] [Medline: 39161704]
25. Olson G, Damme K, Cowan HR, Alliende LM, Mittal VA. Emotional tone in clinical high risk for psychosis: novel insights from a natural language analysis approach. Front Psychiatry. 2024;15:1389597. [FREE Full text] [doi: 10.3389/fpsyt.2024.1389597] [Medline: 38803678]
26. Ricard B, Marsch L, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. JMIR Preprints. Aug 06, 2018. [FREE Full text] [doi: 10.2196/preprints.11817]
27. Wang L, Zhang Y, Zhou B, Cao S, Hu K, Tan Y. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. Comput Electr Eng. 2024;118:109413. [FREE Full text] [doi: 10.1016/j.compeleceng.2024.109413]
28. Liu T, Meyerhoff J, Eichstaedt J, Karr C, Kaiser S, Kording K, et al. et al. The relationship between text message sentiment and self-reported depression. J Affect Disord. 2022;302:7-14. [FREE Full text] [doi: 10.1016/j.jad.2021.12.048] [Medline: 34963643]
29. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. 2023. Presented at: Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]
30. Feng W, Wu H, Ma H, Yin Y, Tao Z, Lu S, et al. et al. Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records. Int J Med Inform. 2025;196:105801. [FREE Full text] [doi: 10.1016/j.ijmedinf.2025.105801] [Medline: 39889672]
31. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. Npj Ment Health Res. 2023;2(1):20. [FREE Full text] [doi: 10.1038/s44184-023-00040-z] [Medline: 38609509]
32. Gratch J, Arstein R, Lucas G, Stratou G, Morency L. The distress analysis interview corpus of human and computer interviews. 2014. Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); May 28-30, 2014; Paris, France.

33. Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. et al. AVEC 2014: 3D dimensional affect and depression recognition challenge. 2014. Presented at: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge; Nov 7, 2014:3; Orlando Florida. [doi: 10.1145/2661806.2661807]
34. Shen Y, Yang H, Lin L. Automatic depression detection: an emotional audio-textual corpus and a Gru/Bilstm-based model. 2022. Presented at: Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 23-27, 2022; Singapore, Singapore. [doi: 10.1109/icassp43922.2022.9746569]
35. Zou B, Han J, Wang Y, Liu R, Zhao S, Feng L, et al. et al. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. IEEE Trans. Affective Comput. 2023;14(4):2823-2838. [doi: 10.1109/taffc.2022.3181210]
36. Guo Y, Liu J, Wang L, Qin W, Hao S, Hong R. A prompt-based topic-modeling method for depression detection on low-resource data. IEEE Trans Comput Soc Syst. 2024;11(1):1430-1439. [doi: 10.1109/tcss.2023.3260080]
37. Senn S, Tlachac M, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. 2022. Presented at: Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); May 11-15, 2022:11-15; New York City. [doi: 10.1109/embc48229.2022.9871120]
38. Sikström S, Valavičiūtė I, Kuusela I, Evors N. Question-based computational language approach outperforms rating scales in quantifying emotional states. Commun Psychol. 2024;2(1):45. [doi: 10.1038/s44271-024-00097-2] [Medline: 39242812]
39. Cheng Q, Li T, Kwok C, Zhu T, Yip P. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res. 2017;19(7):e243. [FREE Full text] [doi: 10.2196/jmir.7276] [Medline: 28694239]
40. Dogrucu A, Perucic A, Isaro A, Ball D, Toto E, Rundensteiner E, et al. et al. Moodable: on feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. Smart Health. 2020;17:100118. [FREE Full text] [doi: 10.1016/j.smhl.2020.100118]
41. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognition & Emotion. 2004;18(8):1121-1133. [doi: 10.1080/02699930441000030]
42. Gong Y, Poellabauer C. Topic modeling based multi-modal depression detection. 2017. Presented at: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge; Oct 23, 2017; Mountain View California USA. [doi: 10.1145/3133944.3133945]
43. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Computer Science. 2020. [doi: 10.3126/jiee.v3i1.34327]
44. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. 2014. Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 26-28, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
45. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
46. Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training. San Francisco, California. OpenAI; 2018.
47. Niu M, Chen K, Chen Q, Yang L. HCAG: a hierarchical context-aware graph attention model for depression detection. 2021. Presented at: Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP);; June 06-11, 2021:6-11; Toronto, ON, Canada. [doi: 10.1109/icassp39728.2021.9413486]
48. Makiuchi M, Warnita T, Uto K, Shinoda K. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. Association for Computing Machinery; 2019. Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Nice, France: Association for Computing Machinery; Oct 21, 2019:55-63; Nice, France. [doi: 10.1145/3347320.3357694]
49. Fan W, He Z, Xing X, Cai B, Lu W. Multi-modality depression detection via multi-scale temporal dilated CNNs. Association for Computing Machinery; 2019. Presented at: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 19, 2019:73-80; Nice, France. [doi: 10.1145/3347320.3357695]
50. Yang L, Jiang D, He L, Pei E, Oveneke M, Sahli H. Decision tree based depression classification from audio video and language information. Association for Computing Machinery; 2016. Presented at: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge; Amsterdam, The Netherlands: Association for Computing Machinery; Oct 16, 2016; Amsterdam, The Netherlands. [doi: 10.1145/2988257.2988269]
51. Dinkel H, Wu M, Yu K. Text-based depression detection: what triggers an alert. arXiv. Apr 08, 2019. [doi: 10.48550/arXiv.1904.05154]
52. Martinez AE, Cuadrado J, Peña D, Martinez SJ, Puertas E. Automated depression detection in text data. Leveraging lexical features, phonesthemes embedding, and RoBERTa transformer model? 2023.
53. Wu Y, Liu Z, Yuan J, Chen B, Cai H, Liu L. PIE: a personalized information embedded model for text-based depression detection. Inf Process Manag. 2024:103830. [FREE Full text] [doi: 10.1016/j.ipm.2024.103830]
54. Dai H, Su C, Lee Y, Zhang Y, Wang C, Kuo C, et al. et al. Deep learning-based natural language processing for screening psychiatric patients. Front Psychiatry. 2020;11:533949. [FREE Full text] [doi: 10.3389/fpsyt.2020.533949] [Medline: 33584354]

55. Jarvers I, Ecker A, Donabauer P, Kampa K, Weißenbacher M, Schleicher D, et al. et al. M.I.N.I.-KID interviews with adolescents: a corpus-based language analysis of adolescents with depressive disorders and the possibilities of continuation using Chat GPT. Front Psychiatry. 2024;15:1425820. [FREE Full text] [doi: 10.3389/fpsyt.2024.1425820] [Medline: 39748903]
56. Zhang Z, Zhang S, Ni D, Wei Z, Yang K, Jin S, et al. et al. Multimodal sensing for depression risk detection: integrating audio, video, and text data. Sensors (Basel). 2024;24(12):3714. [FREE Full text] [doi: 10.3390/s24123714] [Medline: 38931497]
57. He Z. Prediction mechanism of depression tendency among college students under computer intelligent systems. J Intell Syst. 2024;33(1). [doi: 10.1515/jisys-2023-0209]
58. Liu Z, Wu Y, Zhang H, Li G, Ding Z, Hu B. Stimulus-response patterns: the key to giving generalizability to text-based depression detection models. IEEE J Biomed Health Inform. 2024;28(8):4925-4936. [doi: 10.1109/JBHI.2024.3393244] [Medline: 38656850]
59. He L, Niu M, Tiwari P, Marttinen P, Su R, Jiang J, et al. et al. Deep learning for depression recognition with audiovisual cues: a review. Inf Fusion. 2022;80:56-86. [FREE Full text] [doi: 10.1016/j.inffus.2021.10.012]
60. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, et al. et al. The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry. 2013;170(1):1-5. [doi: 10.1176/appi.ajp.2012.12091189] [Medline: 23288382]
61. Kroenke K, Spitzer R. The PHQ-9: A new depression diagnostic and severity measure. San Francisco, California. US: SLACK; 2002.
62. Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of beck depression inventories -IA and -II in psychiatric outpatients. J Pers Assess. 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
63. Zung W. A self-rating depression scale. Arch Gen Psychiatry. 1965;12:63-70. [doi: 10.1001/archpsyc.1965.01720310065008] [Medline: 14221692]
64. Lovibond S, Lovibond P. Manual for the Depression Anxiety Stress Scales. Australia. S. H. Lovibond, Peter F. Lovibond; 1995.
65. Radloff L. The CES-D Scale: a self-report depression scale for research in the general population. US. Sage Publications; 1977:385-401.
66. Li M, Xu H, Liu W, Liu J. Bidirectional LSTM and attention for depression detection on clinical interview transcripts. 2022. Presented at: Proceedings of the 2022 IEEE 10th International Conference on Information, Communication and Networks (ICICN); Aug 23-24, 2022; Zhangye, China. [doi: 10.1109/icicn56848.2022.10006532]
67. Zhao T, Tlachac M. Optimization with tree ensembles to improve depression screening on textual datasets. IEEE Transactions on Affective Computing. 2025;16(2):573-585. [doi: 10.1109/taffc.2024.3442557]
68. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. 2018. Presented at: Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS); June 21, 2018:858-862; Palladam, India. [doi: 10.1109/iss1.2017.8389299]
69. Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, et al. et al. Deep learning for depression detection from textual data. Electronics. 2022;11(5):676. [doi: 10.3390/electronics11050676]
70. Raza G, Butt Z, Latif S, Wahid A. Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. 2021. Presented at: Proceedings of the 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2); May 20-21, 2021; Islamabad, Pakistan. [doi: 10.1109/icodt252288.2021.9441508]
71. Kour H, Gupta MK. Predicting the language of depression from multivariate twitter data using a feature‐rich hybrid deep learning model. Concurr Comput. Jul 22, 2022;34(24):107363. [doi: 10.1002/cpe.7224]
72. Milintsevich K, Sirts K. Your model is not predicting depression well and that is why: a case study of PRIMATE dataset. arXiv. Mar 01, 2024. [doi: 10.48550/arXiv.2403.00438]
73. Arias J, Williams C, Raghvani R, Aghajani M, Baez S, Belzung C, et al. The neuroscience of sadness: a multidisciplinary synthesis and collaborative review. Neurosci Biobehav Rev. 2020;111:199-228. [FREE Full text] [doi: 10.1016/j.neubiorev.2020.01.006] [Medline: 32001274]
74. Kaushik P, Bansal K, Kumar Y, Changela A. Mental disorders prognosis and predictions using artificial intelligence techniques: a comprehensive study. SN Comput Sci. Nov 14, 2024;5(8):1048. [doi: 10.1007/s42979-024-03416-w]
75. Levis B, Benedetti A, Ioannidis J, Sun Y, Negeri Z, He C, et al. et al. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. J Clin Epidemiol. 2020:115-128. [FREE Full text] [doi: 10.1016/j.jclinepi.2020.02.002] [Medline: 32105798]
76. Tahir WB, Khalid S, Almutairi S, Abohashrh M, Memon SA, Khan J. Depression detection in social media: a comprehensive review of machine learning and deep learning techniques. IEEE Access. 2025;13:12789-12818. [doi: 10.1109/access.2025.3530862]
77. Nanggala K, Elwirehardja GN, Pardamean B. Systematic literature review of transformer model implementations in detecting depression. 2023. Presented at: Proceedings of the 6th International Conference of Computer and Informatics Engineering (IC2IE) 2023; Sep 14-15, 2023; Lombok, Indonesia. [doi: 10.1109/ic2ie60547.2023.10331448]

78. Phiri D, Makowa F, Amelia VL, Phiri YVA, Dlamini LP, Chung M. Text-based depression prediction on social media using machine learning: systematic review and meta-analysis. J Med Internet Res. 2025;27:e59002. [FREE Full text] [doi: 10.2196/59002] [Medline: 40215481]
79. Kelley SW, Mhaonaigh CN, Burke L, Whelan R, Gillan CM. Machine learning of language use on Twitter reveals weak and non-specific predictions. NPJ Digit Med. 2022;5(1):35. [FREE Full text] [doi: 10.1038/s41746-022-00576-y] [Medline: 35338248]
80. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):89. [FREE Full text] [doi: 10.1186/s13643-021-01626-4] [Medline: 33781348]
81. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
82. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj.q902] [Medline: 38636956]
83. Lam B, Chrysafi P, Chiasakul T, Khosla H, Karagkouni D, McNichol M, et al. et al. Machine learning natural language processing for identifying venous thromboembolism: systematic review and meta-analysis. Blood advances. 2024;8(12):2991-3000. [FREE Full text] [doi: 10.1182/bloodadvances.2023012200] [Medline: 38522096]
84. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021;21(1):179. [FREE Full text] [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
85. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of Clinical Epidemiology. 2003;56(11):1129-1135. [doi: 10.1016/s0895-4356(03)00177-x]
86. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science. A practical primer for t-tests and ANOVAs. 2013;4:863. [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
87. Cohen J. Statistical power analysis for the behavioral sciences (Rev. ed). Hillsdale, NJ. US. Lawrence Erlbaum Associates, Inc; 1977.
88. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539-1558. [doi: 10.1002/sim.1186] [Medline: 12111919]
89. Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I is not an absolute measure of heterogeneity. Res Synth Methods. 2017;8(1):5-18. [doi: 10.1002/jrsm.1230] [Medline: 28058794]
90. Borenstein M. How to understand and report heterogeneity in a meta-analysis: the difference between I-squared and prediction intervals. Integr Med Res. 2023:101014. [FREE Full text] [doi: 10.1016/j.imr.2023.101014] [Medline: 38938910]
91. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25. [FREE Full text] [doi: 10.1186/1471-2288-14-25] [Medline: 24548571]
92. Bown MJ, Sutton AJ. Quality control in systematic reviews and meta-analyses. Eur J Vasc Endovasc Surg. 2010;40(5):669-677. [FREE Full text] [doi: 10.1016/j.ejvs.2010.07.011] [Medline: 20732826]
93. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2004;56(2):455-463. [FREE Full text] [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
94. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ. 2001;323(7304):101-115. [FREE Full text] [doi: 10.1136/bmj.323.7304.101] [Medline: 11451790]
95. Duval S. The trim and fill method. In: Publication Bias in Meta-Analysis. Chichester, UK. John Wiley & Sons, Ltd; 2005:127-144.
96. Guyatt G, Oxman A, Akl E, Kunz R, Vist G, Brozek J, et al. et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-394. [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
97. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid Based Ment Health. 2017;20(3):83-87. [FREE Full text] [doi: 10.1136/eb-2017-102688] [Medline: 28739578]
98. Tlachac ML, Rundensteiner E. Screening for depression with retrospectively harvested private versus public text. IEEE J Biomed Health Inform. 2020;24(11):3326-3332. [doi: 10.1109/JBHI.2020.2983035] [Medline: 32224470]
99. Shin D, Kim K, Lee S, Lee C, Bae YS, Cho WI, et al. et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Frontiers in Psychiatry. 2022;13:801301. [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]

100. Munthuli A, Pooprasert P, Klangpornkun N, Phienphanich P, Onsuwan C, Jaisin K, et al. et al. Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. PLoS One. 2023;18(3):e0283095. [FREE Full text] [doi: 10.1371/journal.pone.0283095] [Medline: 36996118]
101. Tlachac M, Shrestha A, Shah M, Litterer B, Rundensteiner EA. Automated construction of lexicons to improve depression screening with text messages. IEEE J Biomed Health Inform. 2023;27(6):2751-2759. [doi: 10.1109/JBHI.2022.3203345] [Medline: 36044503]
102. Oh J, Lee T, Chung ES, Kim H, Cho K, Kim H, et al. et al. Development of depression detection algorithm using text scripts of routine psychiatric interview. Front Psychiatry. 2023;14:1256571. [FREE Full text] [doi: 10.3389/fpsyt.2023.1256571] [Medline: 38239906]
103. Xu C, Chen Y, Tao Y, Xie W, Liu X, Lin Y. Deep learning-based detection of depression by fusing auditory, visual and textual clues. Journal of Affective Disorders. 2025:119860. [FREE Full text] [doi: 10.1016/j.jad.2025.119860]
104. Senn SJ. Overstating the evidence: double counting in meta-analysis and related problems. BMC Med Res Methodol. 2009;9(1):10. [FREE Full text] [doi: 10.1186/1471-2288-9-10] [Medline: 19216779]
105. Zhang Y, He Y, Rong L, Ding Y. A hybrid model for depression detection with transformer and bi-directional long short-term memory. 2022. Presented at: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 6-8, 2022; Las Vegas, NV, USA. [doi: 10.1109/bibm55620.2022.9995184]
106. Kaynak EB, Dibeklio?lu H. Systematic Analysis of Speech Transcription Modeling for Reliable Assessment of Depression Severity. Sakarya University Journal of Computer and Information Sciences. 2024 April. 2024;7(1):77-91. [doi: 10.35377/saucis...1381522]
107. Yadav U, Sharma A. A novel automated depression detection technique using text transcript. Int J Imaging Syst Tech. 2022;33(1):108-122. [FREE Full text] [doi: 10.1002/ima.22793]
108. Avasthi S, Chauhan R, Acharjya D. Topic modeling techniques for text mining over a large-scale scientific and biomedical text corpus. IJACI. 2022;13(1):1-18. [doi: 10.4018/ijaci.293137]
109. Du X, Sun Y. Linguistic features and psychological states: a machine-learning based approach. Front Psychol. 2022;13:955850. [FREE Full text] [doi: 10.3389/fpsyg.2022.955850] [Medline: 35936260]
110. Yang K, Kim J, Chun M, Ahn MS, Chon E, Park J, et al. et al. Factors to improve distress and fatigue in cancer survivorship; further understanding through text analysis of interviews by machine learning. BMC Cancer. 2021;21(1):741. [FREE Full text] [doi: 10.1186/s12885-021-08438-8] [Medline: 34176470]
111. Ghosh CC, McVicar D, Davidson G, Shannon C, Armour C. What can we learn about the psychiatric diagnostic categories by analysing patients' lived experiences with machine-learning? BMC Psychiatry. 2022;22(1):427. [FREE Full text] [doi: 10.1186/s12888-022-03984-2] [Medline: 35751077]
112. Tang SX, Kriz R, Cho S, Park SJ, Harowitz J, Gur RE, et al. et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. NPJ Schizophr. 2021;7(1):25. [FREE Full text] [doi: 10.1038/s41537-021-00154-3] [Medline: 33990615]
113. Teleki M, Dong X, Caverlee J. Quantifying the impact of disfluency on spoken content summarization. 2024. Presented at: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); 2024 May; Torino, Italia.
114. Shang J, Xue K, Yang C, Shi H, Pan L, Zeng Y. Prediction models for health care workers' exposure to type II workplace violence: a systematic review and meta-analysis. J Nurs Care Qual. 2026;41(1):29-36. [doi: 10.1097/NCQ.0000000000000897] [Medline: 40658938]
115. Zhang M, Nieuwenhuys A, Zhang Y. Posture prediction models in digital human modeling for ergonomic design: a systematic review. Med Eng Phys. 2025;143:104391. [FREE Full text] [doi: 10.1016/j.medengphy.2025.104391] [Medline: 40835359]
116. Gu Z, Zhu Q. MENTALER: toward professional mental health support with LLMs via multi-role collaboration. 2024. Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 3-6, 2024:3-6; Lisbon, Portugal. [doi: 10.1109/bibm62325.2024.10822643]
117. Otero-González I, Pacheco-Lorenzo M, Fernández-Iglesias MJ, Anido-Rifón LE. Conversational agents for depression screening: a systematic review. Int J Med Inform. 2024;181:105272. [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105272] [Medline: 37979500]

Abbreviations

ADE: automatic depression estimation
BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
CES-D: Center for Epidemiological Studies Depression Scale
DAIC-WOZ: Distress Analysis Interview Corpus – Wizard-of-Oz
DASS-21: 21-item Depression, Anxiety, and Stress Scale
ES: effect size
GRADE: Grading of Recommendations Assessment, Development and Evaluation

HKSJ: Hartung-Knapp-Sidik-Jonkman
ML: machine learning
NLP: natural language processing
OR: odds ratio
PHQ-8: Patient Health Questionnaire-8
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
SDS: Self-Rating Depression Scale
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Edited by S Brini; submitted 20.Aug.2025; peer-reviewed by M Ibrahim, U Pasupulety; comments to author 28.Nov.2025; accepted 10.Jan.2026; published 11.Feb.2026

Please cite as:
Zhang S, Zhang C, Zhang J
Text-Based Depression Estimation Using Machine Learning With Standard Labels: Systematic Review and Meta-Analysis
J Med Internet Res 2026;28:e82686
URL: https://www.jmir.org/2026/1/e82686
doi: 10.2196/82686
PMID: