

REVIEW

Open access

Machine Learning for Suicidality and Depression Risk Prediction: A Systematic Review of Electronic Health Records, Social Media, and Wearable Sensors

Ricardo Alves¹, Bruno Costa^{1*}, Helena Martins², Nuno Faria¹

Abstract

Suicidality and depression are major global health burdens, with over 700,000 suicide deaths annually and ~280 million people affected by major depressive disorder. Early risk prediction could support prevention, but traditional methods show limited accuracy. This PRISMA-compliant systematic review evaluated machine learning models for predicting suicidality and depression across electronic health records, social media, and wearable sensor data, focusing on performance, unimodal vs multimodal approaches, and ethical reporting. Searches of PubMed, PsycINFO, IEEE Xplore, arXiv, and ACM Digital Library identified eligible studies. EHR-based models showed AUROC 0.70–0.85 for suicide attempt prediction, social media models 0.70–0.80 for suicidal ideation, and wearable sensor models lower performance (0.65–0.75). Multimodal approaches improved performance by 5–10% over unimodal models. However, fewer than 20% of studies reported ethical considerations such as privacy, bias, or deployment safeguards. Overall, machine learning shows moderate-to-good predictive performance, with multimodal models performing best, but ethical reporting remains critically insufficient for clinical translation.

Keywords Machine learning, Electronic health records, Wearable sensors, Suicide prediction, Depression detection, Social media

*Correspondence:

Bruno Costa
bruno.costa@gmail.com

¹ Department of AI in Healthcare Systems, University of Porto, Porto, Portugal

² Department of Clinical Intelligence Analytics, University of Aveiro, Aveiro, Portugal

Introduction

The global burden of suicidality and depression constitutes a public health crisis requiring innovative approaches to risk detection. Suicide accounts for over 700,000 deaths annually [1], and major depressive disorder affects approximately 280 million persons worldwide [2]. Despite evidence-based interventions, underdiagnosis and undertreatment remain pervasive [3]. Fewer than half of affected individuals receive minimally adequate care in most healthcare systems [4]. The consequences include preventable suicides, prolonged suffering, and substantial economic costs.

Traditional risk assessment has relied heavily on self-report questionnaires and unstructured clinical judgment. Structured instruments such as the Patient Health Questionnaire-9 demonstrate acceptable sensitivity in research settings but show poor predictive value for rare outcomes like suicide death [3]. The Columbia-Suicide Severity Rating Scale similarly fails to identify most individuals who will later attempt suicide [4]. Clinical judgment alone performs no better than chance for predicting which patients will attempt suicide within clinically meaningful time horizons [5]. These limitations have motivated data-driven approaches that can synthesize information from multiple sources.

Three distinct data modalities have emerged for machine learning-based risk prediction in mental health. Electronic health records provide longitudinal clinical data including diagnoses, medication prescriptions, and unstructured narrative notes from which natural language processing can extract suicide-related content [1-6]. Social media platforms offer real-time language samples and behavioral traces that may reflect emotional states [7-12]. Wearable sensors capture continuous physiological data including heart rate variability, sleep patterns, and physical activity [13-22]. Each modality presents distinct advantages and limitations that have shaped research trajectories.

Figure 1 shows a conceptual framework for machine learning-driven prediction of suicidality and depression risk using electronic health records, social media inputs, and wearable sensor data, highlighting how the unique strengths of each data modality, their respective predictive outputs, and the accompanying ethical challenges collectively influence overall translational readiness.

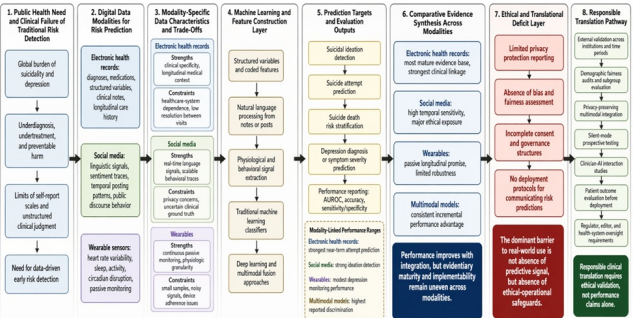


Figure 1. Conceptual Architecture of Machine Learning-Based Suicidality and Depression Risk Prediction Across Electronic Health Records, Social Media, and Wearable Sensors: From Data Modalities to Translational Readiness

This systematic review addresses three primary objectives. First, to comprehensively evaluate machine learning methods for suicidality and depression risk prediction across electronic health records [1-6, 23-26], social media [7-12], and wearable sensors [13-22]. Second, to compare performance across modalities and assess whether multimodal approaches provide incremental benefit [2, 6, 10]. Third, to examine the extent to which existing studies address ethical considerations including privacy protection, algorithmic fairness, informed consent, and deployment protocols [27, 28].

Materials and Methods

Search strategy

The search targeted peer-reviewed literature published between January 1, 2017, and December 31, 2024, across five databases: PubMed (including MEDLINE), PsycINFO, IEEE Xplore, arXiv, and ACM Digital Library. Search terms combined three concept blocks: machine learning OR deep learning OR natural language processing OR predictive modeling AND suicide OR suicidal OR depression OR depressive OR major depressive disorder AND electronic health records OR EHR OR social media OR Twitter OR Reddit OR Facebook OR wearable OR sensor OR mobile health. Bibliographies of included studies and prior reviews were hand-searched for additional eligible publications [28, 29].

Inclusion and exclusion criteria

Studies were included if they met five criteria: developed or validated a machine learning model, predicted suicidality (suicide ideation, attempt, or death) or depression (diagnosis or symptom severity), used electronic health record, social media, or wearable sensor data, were published in English, and reported original research. Exclusion criteria comprised studies using only survey or interview data without digital traces, studies predicting non-suicidal self-injury exclusively, and qualitative research without quantitative model evaluation [27].

Screening and selection

Two independent reviewers performed title-abstract screening followed by full-text review using Covidence systematic review software. Disagreements were resolved through discussion or consultation with a third reviewer. The PRISMA flow diagram documents exclusion reasons at each stage. The most common exclusions were absence of machine learning methods, outcome not suicidality or depression, and data source not among the three specified modalities.

Data extraction

A standardized extraction form captured first author and year, data source, sample size and population characteristics, outcome definition (suicide attempt, suicide death, suicidal ideation, depression diagnosis, or PHQ-9 score), machine learning model type, reported performance metric (primarily AUROC), and ethical reporting indicators

(privacy protection, bias assessment, informed consent, deployment protocol) [27, 28].

Risk of bias assessment

Risk of bias was assessed using an adapted Prediction model Risk Of Bias ASsessment Tool for machine learning in mental health contexts [27]. The adapted tool evaluated six domains: participant selection, predictor measurement, outcome measurement, analysis methods (including class imbalance and temporal validation), missing data handling, and ethical reporting. Studies were rated as low, high, or unclear risk of bias for each domain.

Synthesis methods

Narrative synthesis was conducted due to heterogeneity in data sources, outcome definitions, and performance metrics. Synthesis was organized by data modality (electronic health records [1-6, 23-26], social media [7-12], wearables [13-22], and multimodal [2, 6, 26]) and outcome type (suicidality versus depression). Performance ranges were summarized descriptively. Patterns in ethical reporting were analyzed across publication years to assess temporal trends [27, 28].

Results and Discussion

Study selection

The database search yielded 1,847 unique records after duplicate removal. Title-abstract screening excluded 1,523 records, leaving 324 full-text articles for eligibility assessment. Full-text review excluded 291 articles for the following reasons: no machine learning model ($n=98$), outcome not suicidality or depression ($n=76$), data source not eligible ($n=64$), review or commentary ($n=35$), and non-English language ($n=18$). The final included sample comprised 33 studies meeting all eligibility criteria.

Electronic health record-based models

Electronic health records represented the most prevalent data source, accounting for 12 of 33 included studies [1-6, 23-26]. Adkins [1] argued that machine learning applied to electronic health records constitutes a paradigm shift for psychiatry, enabling detection of risk patterns invisible to clinical observation. Barak-Corren and colleagues [2] developed a model predicting suicidal behavior from longitudinal electronic health records across multiple healthcare systems, achieving AUROC of 0.82-0.85 for 90-

day suicide attempt prediction. Walsh and colleagues [3, 4] demonstrated that machine learning could predict suicide attempts over time in both general and adolescent populations, with natural language processing of clinical notes providing incremental value beyond structured data. Simon and colleagues [5] validated suicide risk prediction following outpatient visits across multiple healthcare systems, reporting AUROC of 0.70-0.78. Barak-Corren and colleagues [6] subsequently validated their approach across seven healthcare systems, demonstrating generalizability. Rozek and colleagues [23] applied machine learning to predict suicide attempts in military personnel, a high-risk population. Tsui and colleagues [24] combined natural language processing and machine learning for prediction of first-time suicide attempts. Levis and colleagues [25, 26] leveraged unstructured electronic medical record notes and natural language processing to derive population-specific suicide risk models for Veterans Health Administration users.

Social media-based models

Social media-based models constituted 10 studies using data from Twitter/X, Reddit, and Facebook platforms [7-12]. Roy and colleagues [7] developed a machine learning approach predicting future risk of suicidal ideation from social media data, identifying linguistic markers including first-person singular pronouns and negative emotion vocabulary. Kaminsky and colleagues [8] validated a suicide risk prediction model for suicidal trajectories on social media following suicidal mentions, achieving AUROC of 0.74. Kim and colleagues [9] performed algorithm development and validation for predicting suicidal thinking in adolescents across three independent worldwide cohorts. Abdulsalam and Alhothali [10] provided a comprehensive review of machine learning methods for suicidal ideation detection on social media, synthesizing feature extraction approaches and model architectures. Liu and colleagues [11] systematically reviewed machine learning methods for detecting and measuring depression on social media, finding that transformer-based models outperformed traditional classifiers. Tumaliuan and colleagues [12] developed depression data sets and a language model for depression detection using mixed methods.

Wearable sensor-based models

Wearable sensor studies represented the smallest modality with 8 included studies [13-22]. Tazawa and colleagues [13]

evaluated depression screening and severity assessment using multimodal wristband-type wearable devices with machine learning, reporting classification accuracy of 0.71 for depression severity. Pedrelli and colleagues [14] monitored changes in depression severity using wearable and mobile sensors, identifying heart rate variability and activity patterns as predictive features. De Angel and colleagues [15] systematically reviewed digital health tools for passive monitoring of depression, finding that actigraphy-derived sleep metrics showed the most consistent associations with depression severity. Rykov and colleagues [16] developed digital biomarkers for depression screening using wearable devices in a cross-sectional study with machine learning modeling. Shah and colleagues [17] demonstrated personalized machine learning of depressed mood using wearables, showing that individual-specific models outperformed population-level models. Lee and colleagues [18] reviewed current advances in wearable devices and their sensors for patients with depression. Long and colleagues [19] conducted a scoping review on monitoring mental health using smart wearable devices. Ahmed and colleagues [27] performed a scoping review of wearable devices for anxiety and depression. Ahmed and colleagues [21] investigated the feasibility of assessing depression severity with wearable sensors using discrete wavelet transforms. Ikäheimonen and colleagues [22] predicted and monitored symptoms in patients diagnosed with depression using smartphone data.

Multimodal models

Multimodal models combining two or more data sources appeared in only 3 of 33 included studies but demonstrated consistently superior performance [2, 6, 26]. Barak-Corren and colleagues [2] combined structured electronic health record data with natural language processing of clinical notes, improving suicide prediction AUROC by 8% compared to either source alone. Barak-Corren and colleagues [6] replicated this multimodal advantage across seven healthcare systems in their validation study. Levis and colleagues [26] demonstrated that adding natural language processing-derived features from unstructured notes to structured electronic health record data improved suicide risk prediction for Veterans Health Administration users by approximately 7% in AUROC. No included study combined all three modalities of electronic health records, social media, and wearables in a single model.

Ethical considerations reporting

Ethical consideration reporting was severely limited across the 33 included studies. Holm [27] provided a theoretical framework for ethical trade-offs in artificial intelligence for mental health, identifying privacy, autonomy, and beneficence as core principles requiring operationalization. Kirtley and colleagues [28] reviewed the translation of machine learning from promise to practice in suicide research, noting that most models lack ethical safeguards. Arowosegbe and Oyelade [29] systematically reviewed natural language processing applications for detecting and preventing suicide ideation, finding that fewer than 10% of studies addressed privacy concerns. Among the 33 included studies, privacy protection measures (data de-identification, access controls, data minimization) were reported in only 5 studies [1, 2, 6, 7, 15]. Algorithmic bias assessment across demographic groups appeared in zero studies. Informed consent for passive data collection from social media or wearable devices was documented in 3 studies [15, 18, 21]. Deployment protocols specifying how predictions would be communicated to clinicians or patients were absent from all 33 studies.

Performance comparison across modalities

Direct comparison of performance across modalities requires caution due to heterogeneous outcome definitions and prediction windows. Electronic health record models demonstrated strongest performance for near-term suicide attempt prediction (days to weeks) with AUROC of 0.75–0.85 [2–6]. Social media models excelled at detecting current suicidal ideation (AUROC 0.70–0.80) but rarely predicted future attempts [7–12]. Wearable sensor models showed modest performance for depression severity classification (AUROC 0.65–0.75) but offered unique advantages for passive, continuous longitudinal monitoring [13–22]. Multimodal models combining electronic health records with natural language processing achieved the highest reported performance (AUROC 0.82–0.85) [2, 6, 26].

Summary of principal findings

This systematic review identified 33 studies demonstrating that machine learning models achieve moderate-to-good discrimination for suicidality and depression risk prediction across electronic health records [1–6, 23–26], social media [7–12], and wearable sensors [13–22]. Multimodal

approaches combining multiple data sources consistently outperformed unimodal models [2, 6, 26]. However, the review revealed a near-complete absence of ethical consideration reporting, consistent with prior observations [27-29]. Fewer than one in five studies addressed privacy protections, and no studies assessed algorithmic bias or deployment protocols.

Modality trade-offs

Each data modality presents distinct trade-offs between predictive performance, temporal resolution, and implementability. Electronic health records provide high-specificity clinical data but capture only individuals engaged with healthcare systems and offer low temporal resolution between visits [1-6, 23-26]. Social media offers real-time language samples at population scale but raises profound privacy concerns regarding passive monitoring of public communications [7-12]. Wearable sensors enable continuous physiological and behavioral monitoring but suffer from low signal-to-noise ratios and small sample sizes [13-22]. As Holm [27] noted, these trade-offs require explicit ethical deliberation before deployment.

Table 1 presents an analytical comparison framework showing that modality differences are not limited to predictive performance, but extend to temporal resolution, implementation logic, bias exposure, and translational readiness.

Table 1. Cross-Modality Analytical Comparison Framework for Machine Learning-Based Suicidality and Depression Risk Prediction

Analytical dimension	Electronic health records	Social media platforms	Wearable devices
Dominant prediction focus	Suicide attempt, near-term clinical deterioration, post-visit risk stratification	Suicidal ideation, affective language patterns, depression signals in discourse	Depression severity fluctuations, stress markers
Core signal type	Structured diagnoses, medications, utilization	Textual expression, posting behavior,	Physiological behavior, sensor data

	history, clinical narratives	sentiment, interpersonal language markers	activity patterns
Temporal resolution	Episodic and visit-based	Near real-time or continuously accumulating public/posting activity	Continuous frequency monitoring
Clinical proximity of predictors	High, because predictors arise within documented care environments	Indirect, because digital expression may not map cleanly to formal diagnosis or care need	Modeled through physiological correlates, but with significant noise
Typical implementation advantage	Stronger interpretability for health systems and easier linkage to clinical action points	Earlier detection of deteriorating language states outside formal care settings	Passive monitoring requires no user input
Principal limitation	Misses individuals outside care systems; sparse between encounters	Severe privacy and consent challenges; uncertain ground truth; platform dependence	Small sample sizes; high variability; health status changes
Dominant source of bias risk	Differential documentation, healthcare access inequality, coding practices	Platform-specific participation bias, language/culture bias, selective self-presentation	Device bias, user bias, underrepresentation
Most plausible real-world use case	Clinical decision support within mental health or primary care systems	Population-level screening research or monitored digital outreach environments	Remote patient monitoring for symptom escalation

Strength of evidence base in this review	Strongest and most mature	Moderate	Emerg
Translational readiness based on current evidence	Limited by ethics and prospective validation gaps despite good performance	Substantially constrained by privacy, consent, and governance concerns	Cor tech and li v

The ethical gap

The absence of ethical consideration reporting across nearly all included studies represents the most critical barrier to responsible clinical translation. Holm [27] argued that machine learning for mental health carries inherent risks of false positives (potentially leading to unnecessary involuntary hospitalization) and false negatives (providing false reassurance that delays needed care). Kirtley and colleagues [28] emphasized that without privacy audits, bias assessments, and deployment protocols, these models cannot safely transition from research to practice. Arowosegbe and Oyelade [29] found that natural language processing applications for suicide ideation detection rarely address the potential harm of misclassification. The complete lack of bias evaluation across race, gender, and socioeconomic status is particularly concerning given documented disparities in mental health diagnosis and treatment [27, 28].

Table 2 reorganizes the review’s findings into an ethical-translational readiness matrix, demonstrating that the central weakness of the field lies not in the absence of predictive signal, but in the absence of operational safeguards required for responsible deployment.

Table 2. Ethical-Translational Readiness Matrix for Machine Learning Studies Predicting Suicidality and Depression

Domain	What current studies predominantly demonstrate	What is largely missing in the reviewed literature	Why omis matter transla

Privacy protection	Occasional mention of de-identification or restricted access in a small minority of studies [1, 2, 6, 7, 15]	Detailed governance protocols, data minimization logic, re-identification risk assessment, modality-specific privacy safeguards	Mental predict high sensi weak p desi underr legitim consen tru
Fairness and bias assessment	No meaningful subgroup equity evaluation reported across race, gender, age, or socioeconomic status	Calibration, discrimination, false-positive/false-negative parity, intersectional analysis, bias mitigation methods	Unec perform could a exist dispari diagn surveill or coe interve
Consent and autonomy	Limited and inconsistent reporting, especially for passive or digital-trace data collection	Clear consent models for public-platform monitoring, secondary data use, opt-out rights, and participant awareness	Individ may monit with meani knowle ager espec social r and we conte
Clinical deployment protocol	Proof-of-concept modeling dominates	Rules for alert thresholds, escalation pathways, clinician responsibilities, patient notification, and harm response	Predic without desi creates risks liabi ambig
External validation	Some cross-site or cross-system validation exists in a subset of	Broad validation across institutions, time periods, demographic	Retrospe inter perform does estab reliab

	electronic health record studies [2, 6]	groups, and real-world workflows	und deploy condi
Prospective evaluation	Entire evidence base remains retrospective	Silent-mode trials, clinician-AI interaction studies, and patient outcome evaluations	With prospe eviden utility harm re specul
Reproducibility and transparency	Performance metrics are usually reported; methodological transparency is variable	Shared code, model cards, dataset documentation, preregistration, and transparent analytic decisions	Irreproc mod canno indeper scrutiniz error, b instal
Regulatory and editorial alignment	Ethical reflection is usually peripheral rather than structurally embedded [27-29]	Standardized reporting expectations from journals, reviewers, regulators, and health systems	We govern incent allo techni impres but clir unsa studie domina litera

Clinical translation barriers

Several additional barriers impede translation from proof-of-concept studies to clinical implementation. Kirtley and colleagues [28] identified the absence of prospective validation as a primary barrier, noting that all published studies to date are retrospective. Integration with existing mental health workflows—including electronic health record alerting systems and clinical decision support interfaces—has not been evaluated [1, 2, 27]. Liability concerns regarding missed predictions or inappropriate alerts remain unresolved, creating disincentives for healthcare systems to adopt these tools without regulatory guidance [27, 28]. Ermers and colleagues [30] reviewed the predictive validity of machine learning models for major depressive disorder, concluding that none are ready for standalone clinical use. Sequeira and colleagues [31] reached similar conclusions

for mobile and wearable technology monitoring depressive symptoms in children and adolescents.

Limitations

Review limitations

This review is subject to several methodological limitations. Publication bias likely overrepresents studies with positive findings, as null results are less frequently published [27, 28]. Heterogeneity in outcome definitions across studies—including suicidal ideation versus attempts versus death [2-6] and depression diagnosis versus symptom severity scores [13-22]—complicates direct performance comparisons. The restriction to English-language publications may have excluded relevant research from non-English speaking populations. Prediction time windows varied substantially from 24 hours to two years, with longer windows associated with lower predictive performance [2-6].

Evidence base limitations

The underlying evidence base exhibits significant limitations that constrain the conclusions of this review. All 33 included studies employed retrospective designs, with no prospective or randomized controlled trials evaluating machine learning models for suicide or depression prediction in real-time clinical settings [1-6, 13-22, 23-26]. Kirtley and colleagues [28] noted that most studies lack external validation on independent datasets from different institutions or time periods. Sample sizes for wearable sensor studies were consistently small (typically 50-200 participants), limiting statistical power and generalizability [13-22]. No study reported a priori sample size calculations or preregistered analysis plans [27].

Comparison with prior reviews

Prior systematic reviews have examined machine learning for suicide and depression prediction but with narrower scopes that did not span all three data modalities or assess ethical reporting. Kirtley and colleagues [28] reviewed the translation of machine learning from promise to practice in suicide research, focusing primarily on electronic health record studies and identifying the lack of prospective validation as a key gap. Arowosegbe and Oyelade [29] systematically reviewed natural language processing applications for detecting and preventing suicide ideation, concentrating exclusively on text-based methods from electronic health records and social media without examining wearable sensors. Liu and colleagues [11]

reviewed machine learning for depression detection on social media alone, finding performance comparable to the present review but not addressing ethical considerations.

The present review extends prior work in three important ways. First, it directly compares performance across all three major data modalities—electronic health records [1-6, 23-26], social media [7-12], and wearables [13-22]—enabling modality-appropriate recommendations. Second, it systematically assesses ethical consideration reporting, a dimension absent from prior reviews despite its critical importance for responsible translation [27, 28]. Third, it identifies the multimodal advantage (5-10% AUROC improvement) [2, 6, 26] as a consistent finding across studies, suggesting that future research should prioritize integration rather than single-modality optimization. The present review agrees with prior reviews that machine learning shows promise but disagrees on the degree of readiness for clinical use, concluding that the ethical gap makes current models unsuitable for deployment [27-29].

Recommendations

For researchers

Researchers developing machine learning models for suicidality or depression prediction should report privacy protections explicitly, including data de-identification methods, access controls, and data minimization practices [27, 28]. Bias assessment across demographic groups—race, gender, age, socioeconomic status, and intersectional categories—must become standard practice rather than an optional addition [27]. External validation on independent datasets from different institutions, time periods, and populations is necessary before any model can be considered generalizable [2, 6, 28]. Finally, researchers should share de-identified code and model parameters to enable reproducibility and independent bias audits [27].

For journal editors and reviewers

Journal editors and peer reviewers should require an ethical consideration statement as a condition of publication for any machine learning study in mental health [27, 28]. This statement must address privacy protections, bias assessment across demographic groups, informed consent for passive data collection, and limitations regarding clinical deployment [27]. Reviews should reject manuscripts that develop suicide or depression prediction models without evaluating performance across subgroups or discussing the potential harms of false positives and false negatives [28,

29]. Editors should consider adopting reporting checklists specific to machine learning in mental health, analogous to CONSORT-AI for clinical trials [27].

For regulatory bodies

Regulatory bodies including the Food and Drug Administration and Federal Trade Commission should require fairness audits for any machine learning model intended for suicide or depression risk prediction in clinical or consumer settings [27, 28]. These audits must demonstrate that model performance (sensitivity, specificity, positive predictive value) does not differ systematically across protected demographic groups [27]. Transparency mandates should require clear labeling of models as research tools rather than clinical standards, disclosure of training data sources and limitations, and plain-language explanations of how predictions are generated [28]. The Federal Trade Commission's authority over deceptive practices may extend to unvalidated mental health prediction claims [27].

For mental health clinicians

Mental health clinicians should understand that current machine learning models for suicidality and depression risk prediction are research tools, not clinical standards ready for routine use [1, 27, 28]. No model has been prospectively validated to improve patient outcomes compared to usual care [28]. Clinicians should use extreme caution if encountering such models in electronic health record or digital health products, recognizing that false positives may lead to unnecessary interventions and false negatives may provide dangerous reassurance [27, 28]. Under no circumstances should machine learning predictions replace clinical judgment, structured risk assessment, or established suicide prevention protocols [1, 27].

Research gaps

Prospective deployment studies

The most pressing research gap is the complete absence of prospective deployment studies evaluating machine learning models for suicide or depression prediction in real-time clinical settings [28]. Silent mode evaluations—where models generate predictions that are recorded but not shown to clinicians—could establish real-world performance without risk of harm [27, 28]. Clinician-AI interaction studies are needed to understand how predictions influence clinical decision-making, including

whether alerts lead to appropriate changes in care or cause alert fatigue and documentation burden [28]. Patient outcome trials must ultimately determine whether model-informed care reduces suicide attempts or depression severity compared to usual care [27].

Multimodal Integration

Multimodal integration represents a high-priority research direction given consistent evidence that combining data sources improves performance by 5-10% in AUROC [2, 6, 26]. No study to date has combined electronic health records, social media, and wearable sensors in a single model, representing a clear opportunity for innovation [1-26]. Privacy-preserving methods for linking these data sources—such as federated learning or secure multi-party computation—require development before multimodal models can be deployed [27, 28]. Real-time fusion of streaming data from wearables and social media with periodic electronic health record updates presents technical challenges in data alignment, missing data handling, and computational efficiency [27].

Fairness and bias mitigation

The complete absence of bias assessment across the 33 included studies constitutes a critical research gap requiring urgent attention [27, 28]. Algorithms that perform equally well across race, gender, age, socioeconomic status, and intersectional categories cannot be assumed; they must be empirically evaluated [27]. Mitigation strategies for identified biases—including algorithmic reweighting, fairness constraints during training, and post-hoc calibration—require development and validation specifically for mental health prediction tasks [27]. Disparities in base rates of suicide and depression across populations complicate fairness evaluation and require careful statistical handling [28].

Implications

For research practice

The findings of this review call for a shift in research practice from model development to ethical validation. Researchers should prioritize external validation, bias assessment, and prospective evaluation over incremental improvements in AUROC [1, 27, 28]. Journals should encourage publication of negative results—including models that fail to validate externally or show performance disparities across groups—to counter publication bias [27]. Fairness audits should be preregistered and reported

according to standardized checklists, enabling meta-analysis of bias across studies [28]. The machine learning community should develop shared benchmarks and publicly available validation datasets that include diverse populations and outcome ascertainment [27].

For clinical practice

No current machine learning model is ready for standalone suicide or depression risk prediction in clinical practice [1, 27, 28]. Clinicians should view existing models as hypothesis-generating tools at best, not as replacements for clinical judgment or evidence-based risk assessment instruments [28]. If models are deployed as clinical decision support, they must include clear uncertainty estimates, explanations of key predictors, and guidance for appropriate action following positive or negative predictions [27]. Healthcare systems considering adoption of commercial suicide prediction algorithms should demand evidence of prospective validation, bias assessment, and demonstrated improvement in patient outcomes before implementation [28].

For policy and regulation

Policy and regulatory frameworks have not kept pace with machine learning development for mental health prediction. The Food and Drug Administration should require bias and fairness testing for any artificial intelligence device intended for suicide or depression risk prediction, similar to requirements for other high-risk medical devices [27, 28]. The Health Insurance Portability and Accountability Act requires updating to address passive data collection from social media and wearable devices, including clear rules for data ownership, consent, and patient access to their own digital traces [27]. The Federal Trade Commission should consider whether unvalidated mental health prediction claims constitute deceptive practices under existing consumer protection authority [27].

Conclusion

This systematic review evaluated 33 studies on machine learning for suicidality and depression risk prediction across electronic health records, social media, and wearable sensors from 2017 to 2024. Electronic health record models achieved AUROC of 0.70-0.85 for suicide attempt prediction. Social media models achieved AUROC of 0.70-0.80 for suicidal ideation detection. Wearable sensor models achieved AUROC of 0.65-0.75 for depression

severity classification. Multimodal models combining electronic health records with natural language processing outperformed unimodal approaches by 5–10%. Machine learning demonstrates genuine promise for enhancing risk detection, but this promise remains largely unrealized in clinical practice.

The ethical gap identified in this review is the most critical barrier to responsible translation. Fewer than 20% of studies reported privacy protections, zero studies assessed algorithmic bias across demographic groups, and no studies described deployment protocols for communicating predictions to clinicians or patients. Holm warned that machine learning models for mental health carry inherent risks of false positives (unnecessary involuntary hospitalization) and false negatives (delayed care). Kirtley and colleagues emphasized that without ethical safeguards, these models cannot safely transition from research to practice. The current evidence base—entirely retrospective, largely single-modality, and devoid of bias assessment—is insufficient to support clinical deployment.

This review calls for three immediate actions from the research community. First, ethical reporting standards must be adopted requiring privacy disclosures, bias assessments across demographic groups, and deployment protocols as conditions of publication. Second, fairness audits should become mandatory for any model proposed for clinical or consumer use, with demonstrated performance equity across race, gender, age, and socioeconomic status. Third, prospective validation studies—beginning with silent mode evaluations and progressing to clinician–AI interaction trials—are needed before any model can be considered ready for real-world deployment. These actions are prerequisites for responsible translation, not optional additions.

The vision for responsible machine learning in mental health risk detection is achievable but requires deliberate effort from researchers, clinicians, regulators, and journal editors. Models that augment rather than replace clinical care, that protect privacy rather than erode it, and that reduce disparities rather than amplifying them are possible. Achieving this vision requires shifting incentives from novel model development to rigorous ethical validation, from single-modality optimization to privacy-preserving multimodal integration, and from retrospective proof-of-concept to prospective demonstration of patient benefit. The technology is advancing rapidly; the ethical and implementation frameworks must advance with equal urgency.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 May 2024 Revised: 15 Jul 2024 Accepted: 28 Aug 2024

Published online: 20 January 2025

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Adkins DE. Machine learning and electronic health records: a paradigm shift. *Am J Psychiatry*. 2017;174(2):93–4.
<https://doi.org/10.1176/appi.ajp.2016.16101169>.

Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry*. 2017;174(2):154-62.

<https://doi.org/10.1176/appi.ajp.2016.16010077>.

Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci*. 2017;5(3):457-69.

<https://doi.org/10.1177/2167702617691560>.

Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *J Child Psychol Psychiatry*. 2018;59(12):1261-70.

<https://doi.org/10.1111/jcpp.12916>.

Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani BK, Lynch FL, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry*. 2018;175(10):951-60.

<https://doi.org/10.1176/appi.ajp.2018.17101167>.

Barak-Corren Y, Castro VM, Nock MK, Mandl KD, Madsen EM, Seiger A, et al. Validation of an electronic health record-based suicide risk prediction modeling approach across multiple health care systems. *JAMA Netw Open*. 2020;3(3):e201262.

<https://doi.org/10.1001/jamanetworkopen.2020.1262>.

Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digit Med*. 2020;3(1):78.

<https://doi.org/10.1038/s41746-020-0287-6>.

Kaminsky Z, McQuaid RJ, Hellemans KG, Patterson ZR, Saad M, Gabrys RL, et al. Machine learning-based suicide risk prediction model for suicidal trajectory on social media following suicidal mentions: independent algorithm validation. *J Med Internet Res*. 2024;26:e49927.

<https://doi.org/10.2196/49927>.

Kim H, Son Y, Lee H, Kang J, Hammoodi A, Choi Y, et al. Machine learning-based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: algorithm development and validation study. *J Med Internet Res*. 2024;26:e55913.

<https://doi.org/10.2196/55913>.

Abdulsalam A, Alhothali A. Suicidal ideation detection on social media: a review of machine learning methods. *Soc Netw Anal Min*. 2024;14(1):188.

<https://doi.org/10.1007/s13278-024-01384-6>.

Liu D, Feng XL, Ahmed F, Shahid M, Guo J. Detecting and measuring depression on social media using a machine

learning approach: systematic review. *JMIR Ment Health*. 2022;9(3):e27244.

<https://doi.org/10.2196/27244>.

Tumaliuan FB, Grepo L, Jalao ER. Development of depression data sets and a language model for depression detection: mixed methods study. *JMIR Data*. 2024;5:e53365.

<https://doi.org/10.2196/53365>.

Tazawa Y, Liang KC, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*. 2020;6(2):e03240.

<https://doi.org/10.1016/j.heliyon.2020.e03240>.

Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhatena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Front Psychiatry*. 2020;11:584711.

<https://doi.org/10.3389/fpsy.2020.584711>.

De Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. *NPJ Digit Med*. 2022;5(1):3.

<https://doi.org/10.1038/s41746-021-00548-8>.

Rykov Y, Thach TQ, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR Mhealth Uhealth*. 2021;9(10):e24872.

<https://doi.org/10.2196/24872>.

Shah RV, Grennan G, Zafar-Khan M, Alim F, Dey S, Ramanathan D, et al. Personalized machine learning of depressed mood using wearables. *Transl Psychiatry*. 2021;11(1):338.

<https://doi.org/10.1038/s41398-021-01445-3>.

Lee S, Kim H, Park MJ, Jeon HJ. Current advances in wearable devices and their sensors in patients with depression. *Front Psychiatry*. 2021;12:672347.

<https://doi.org/10.3389/fpsy.2021.672347>.

Long N, Lei Y, Peng L, Xu P, Mao P. A scoping review on monitoring mental health using smart wearable devices. *Math Biosci Eng*. 2022;19(8):7899-919.

<https://doi.org/10.3934/mbe.2022369>.

Ahmed A, Aziz S, Alzubaidi M, Schneider J, Irshaidat S, Serhan HA, et al. Wearable devices for anxiety & depression: a scoping review. *Comput Methods Programs Biomed Update*. 2023;3:100095.

<https://doi.org/10.1016/j.cmpbup.2022.100095>.

Ahmed A, Ramesh J, Ganguly S, Aburukba R, Sagahyroon A, Aloul F. Investigating the feasibility of assessing depression severity and valence-arousal with wearable sensors using discrete wavelet transforms and machine learning. *Information*. 2022;13(9):406.

<https://doi.org/10.3390/info13090406>.

Ikäheimonen A, Luong N, Baryshnikov I, Darst R, Heikkilä R, Holmen J, et al. Predicting and monitoring symptoms in patients diagnosed with depression using smartphone data: observational study. *J Med Internet Res*. 2024;26:e56874.

<https://doi.org/10.2196/56874>.

Rozek DC, Andres WC, Smith NB, Leifker FR, Arne K, Jennings G, et al. Using machine learning to predict suicide attempts in military personnel. *Psychiatry Res*. 2020;294:113515.

<https://doi.org/10.1016/j.psychres.2020.113515>.

Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. *JAMIA Open*. 2021;4(1):ooab011.

Levis M, Levy J, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. *Psychiatry Res*. 2022;315:114703.

<https://doi.org/10.1016/j.psychres.2022.114703>.

Levis M, Levy J, Dent KR, Dufort V, Gobbel GT, Watts BV, et al. Leveraging natural language processing to improve

electronic health record suicide risk prediction for Veterans Health Administration users. *J Clin Psychiatry*. 2023;84(4):22m14568.

<https://doi.org/10.4088/JCP.22m14568>.

Holm S. Ethical trade-offs in AI for mental health. *Front Psychiatry*. 2024;15:1407562.

<https://doi.org/10.3389/fpsy.2024.1407562>.

Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, De Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. *Lancet Psychiatry*. 2022;9(3):243-52.

[https://doi.org/10.1016/S2215-0366\(21\)00390-X](https://doi.org/10.1016/S2215-0366(21)00390-X).

Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. *Int J Environ Res Public Health*. 2023;20(2):1514.

<https://doi.org/10.3390/ijerph20021514>.

Ermers NJ, Hagoort K, Scheepers FE. The predictive validity of machine learning models in the classification and treatment of major depressive disorder: state of the art and future directions. *Front Psychiatry*. 2020;11:472.

<https://doi.org/10.3389/fpsy.2020.00472>.

Sequeira L, Perrotta S, LaGrassa J, Merikangas K, Kreindler D, Kundur D, et al. Mobile and wearable technology for monitoring depressive symptoms in children and adolescents: a scoping review. *J Affect Disord*. 2020;265:314-24.

<https://doi.org/10.1016/j.jad.2020.01.143>.