Journal of Rare Cardiovascular Diseases
ISSN: 2299-3711 (Print) | e-ISSN: 2300-5505 (Online)
www.jrcd.eu
RESEARCH ARTICLE

Predictive Modeling for Mental Health: A Review of Machine Learning Methods for Early Detection

Suman Vashist[1], Nitesh Kumar Sain[2], Shivangi Gupta[3], Jaivin Jaisingh J[4], Chanderveer[5], Tanuja Semwal[6], Ruchi[7]

[1]Professor, Mental Health Nursing, Teerthanker Mahaveer College of Nursing, Teerthanker Mahaveer University, Moradabad, U.P.
[2]Assistant Professor, Mental Health Nursing, Jaipur Nursing College Maharaj Vinayak Global University, Jaipur, RJ.
[3]Assistant Professor, Obstetrics and Gynecology Nursing, Teerthanker Mahaveer College of Nursing, Teerthanker Mahaveer University, Moradabad, U.P.
[4]Associate Professor, Medical surgical nursing, T.S Misra College of nursing, Amausi, Uttar Pradesh.
[5]Senior nursing officer, all india institute of medical sciences, Patna.
[6]Assistant Professor, Beehive Nursing College, Selaqui, Dehradun.
[7]Nursing Tutor, ANM TC, Khirsu, Pauri, Garhwal, UK.

*Corresponding Author

Article History
Received: 16.09.2025
Revised: 14.10.2025
Accepted: 05.11.2025
Published: 01.12.2025

Abstract: Early detection of mental-health risk can avert crises and enable timely, lower-intensity care.We compare methods from regularized regression and gradient boosting to transformers for NLP and multimodal time series, emphasizing calibration and clinical utility. Across health-system studies, parsimonious EHR-based models achieve actionable short-term suicide-risk discrimination, with performance strengthened by text-derived features. Digital phenotyping shows promise for relapse monitoring but requires external validation, engagement strategies, and humane alerting. Social-media signals support population-level surveillance while demanding rigorous consent and governance. Implementation studies indicate risk-guided prompts can improve assessment and safety planning when integrated into workflow. We recommend task-first model design, local recalibration, transparent reporting (TRIPOD+AI), bias appraisal (PROBAST+AI), and evaluation beyond AUCs to patient-centered outcomes.

Keywords: Mental Health, Machine Learning, Early Detection, Suicide Risk Prediction, Digital Phenotyping.

INTRODUCTION

Around the world, clinicians, payors, and public-health agencies are asking whether machine learning (ML) can help identify mental-health risk earlier—before crises or costly care escalations occur. Over the past decade, predictive modeling has progressed from proof-of-concept studies on social media and electronic health records (EHRs) to pragmatic trials that embed risk scores into clinical workflows. The promise is straightforward: earlier detection of risk for suicide attempts, relapse in psychosis or mood disorders, or deterioration in depression could trigger more timely, targeted interventions. Yet the field has also learned hard lessons about transportability, fairness, and the need for rigorous evaluation that goes beyond headline AUCs.[1–5]

Three trends motivate this review. First, data sources for mental-health prediction have diversified. In addition to structured EHR variables (diagnoses, utilization, medications), unstructured clinical notes and patient communications are now routinely mined with natural language processing (NLP), and passive sensing ("digital phenotyping") from smartphones and wearables can capture behavioral change at high temporal resolution.[1] Second, methods have matured. While early work leaned on regularized logistic regression and random forests, transformer-based NLP models and sequence models for time-series now compete with, and sometimes complement, classical baselines; importantly, sophisticated models do not always outperform well-tuned simpler ones.[2] Third, the evidence base is moving from retrospective modeling toward implementation science: randomized and quasi-experimental studies are beginning to test whether risk-guided care improves clinical processes and outcomes.[3]

At the same time, earlier detection raises non-trivial clinical and societal questions. Algorithmic bias may reinforce inequities across race/ethnicity, age, or geography if models are trained on skewed data; external validity often drops when models travel to new sites; and privacy risks are amplified by the sensitivity of psychiatric data and the granularity of sensor streams. Addressing these requires transparent reporting and risk-of-bias appraisal (TRIPOD+AI; PROBAST+AI), attention to explainability for end-users, and governance that matches the velocity of technical innovation.[6,7]

Why early detection matters

In child and adolescent care, timely identification of self-injury has clear implications for triage and follow-up; work using EHR phenotyping in emergency departments shows that coding alone under-captures risk-relevant behaviors, whereas informatics approaches can surface latent risk earlier in the pathway.[4] Early detection also matters for

historically underserved populations. When investigators evaluated how existing ML models perform in a majority American Indian population, they found that—used thoughtfully—models added value beyond standard screening, underscoring the need for local validation and adaptation rather than one-size-fits-all deployment.[5] Prospective validation within large integrated systems similarly shows that well-calibrated models can anticipate short-term suicide attempts after mental-health visits, allowing teams to target outreach, safety planning, and same-day assessment more efficiently.[8],[9]

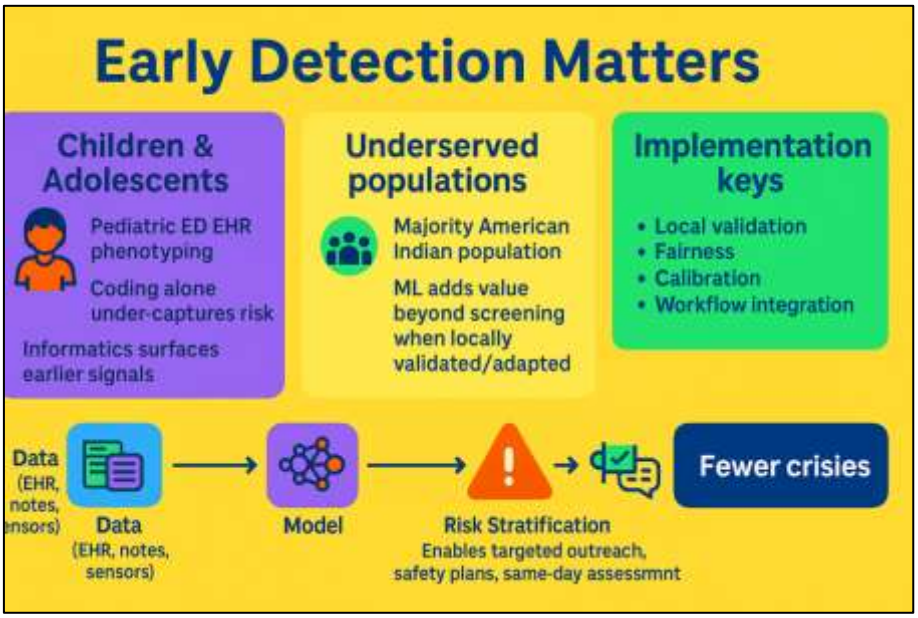

Figure 1. Early Detection Matters

Data Streams for Prediction: EHR, Notes, Sensors, and Social Data

EHRs remain the workhorse for predictive modeling: encounters, problem lists, psychotropic medications, and prior utilization are consistent high-value features for suicide-attempt prediction and relapse forecasting.[8],[9] Unstructured text augments these signals: NLP applied to mental-health notes can recover symptoms (e.g., suicidal ideation, anhedonia) and temporal trajectories that are often missing from structured fields, improving risk stratification when layered onto existing models.[10],[11],[12] Outside the clinic, continuous passive sensing from smartphones and wearables—GPS variance, activity, sleep regularity, screen and app usage—offers high-granularity behavioral data that track prodromal shifts in psychosis, bipolar disorder, and depression.[13–16] Finally, social media provides a population-level window: linguistic markers in user posts can signal depressive symptoms weeks or months ahead of clinical diagnosis, though deployment raises important consent and governance questions.[1]

Algorithms in Use: From Regression Baselines to Transformers

A recurring finding in mental-health prediction is that complexity is not automatically a strength. In large-scale comparisons, complex learners (ensembles, neural nets with thousands of temporally-defined predictors) sometimes match, but do not consistently exceed, well-specified logistic regression—particularly when design choices around time windows, leakage control, and class imbalance are optimized.[2] For unstructured text, transformer models pre-trained on clinical corpora (e.g., Bio_ClinicalBERT, GatorTron) have shown clear gains over rules and bag-of-words for extracting psychiatric history and risk factors from notes, and these models are increasingly used to build features for downstream risk prediction.[10] Methodologically, the field benefits from a "horses for courses" approach: choose the simplest model that meets the clinical requirement (discrimination, calibration, latency, maintainability), and reserve deep architectures for modalities where they are decisively better (long-range language or multimodal time-series).[2],[10]

Suicide Risk Prediction in Health Systems

Among use cases, suicide risk modeling is the most developed. Prospective studies in large systems (e.g., Kaiser Permanente) have validated models that predict short-term suicide attempts following mental-health visits, with performance sufficient to guide outreach at practical alert thresholds.[8] External, multicenter validations of parsimonious logistic and penalized logistic models further support transportability when development controls overfitting and aligns predictors with routinely available data.[9] In a randomized clinical trial, interruptive decision support triggered by model-identified high-risk visits increased in-person suicide risk assessments and use of safety planning early evidence that predictive models can change clinician behavior at the point of care.[3] Notably, evaluating models in populations that experience the greatest inequities is crucial: in a majority American Indian population, machine-learning models added signal

beyond standard screeners, but careful local calibration and workflow integration were essential to realize benefit without harm.[5]

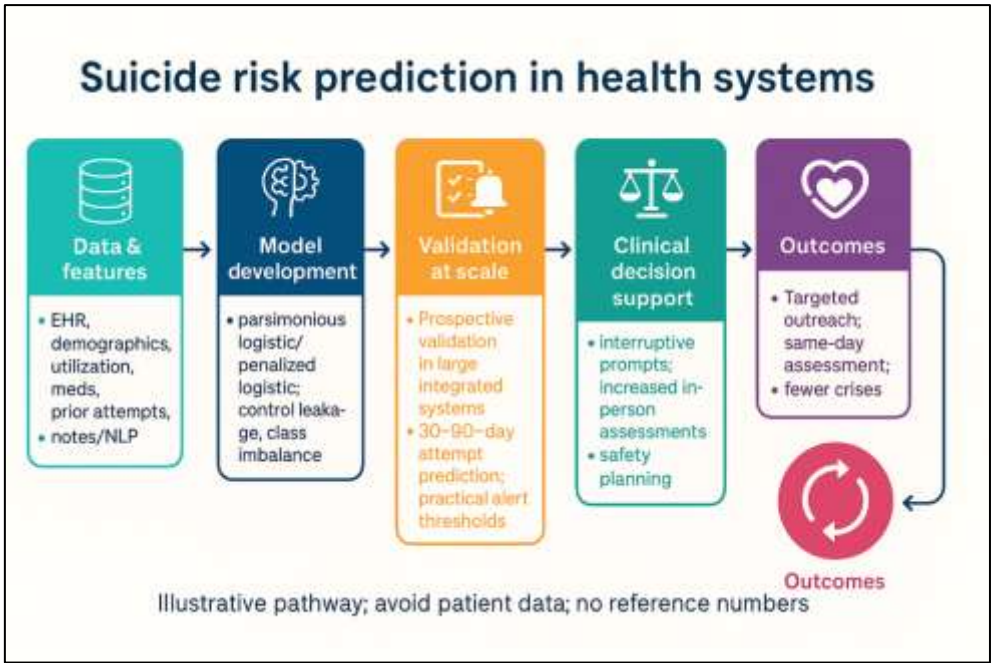

Figure 2. Suicide Risk Prediction in Health Systems

NLP on Clinical Text: Adding What Structured Data Miss

Unstructured clinical notes encode critical phenomena—recent loss, hopelessness, access to means that structured fields rarely capture. When NLP-derived features (e.g., mentions of suicidal ideation over time) are layered onto EHR baselines, several studies report incremental gains in discrimination for suicide-related outcomes and improved timeliness of detection.[11] Beyond feature extraction, scoping reviews document that modern NLP pipelines (including deep models) can reliably detect suicidal thoughts and behaviors across settings, while emphasizing the need for standardized outcome definitions and transparent reporting of thresholds and false-alarm trade-offs.[12] For broader psychiatric informatics, transformer-based extraction of personal and family psychiatric history shows how language models can convert free-text into robust phenotypes for downstream prediction, a pattern likely to generalize to progress notes and crisis-text interactions.[10]

Digital Phenotyping and Relapse Monitoring

Continuous sensing captures behavior change earlier and outside clinic walls. In schizophrenia, longitudinal smartphone-based sensing (mobility entropy, activity rhythms, communication patterns) has been used to anticipate relapse risk, enabling proactive outreach in some cohorts.[13] Synthetic overviews argue that digital phenotyping can deliver clinically meaningful benefits—early warning, personalized trajectories—if signal processing and human-factors hurdles (battery, burden, adherence) are addressed.[14,15] Recent adolescent studies suggest feasibility of long-term sensing and point to age-specific engagement and privacy considerations.[16] Still, external validation remains scarce, pipelines are heterogenous, and linking alerts to effective, acceptable interventions is the next translational step.[14,15]

Social Media Signals for Depression

Systematic reviews of ML on social platforms (e.g., Reddit, Twitter) consistently find moderate-to-strong performance for detecting depression using lexical, semantic, and temporal features, including with transformer embeddings.[1] Recent meta-analyses emphasize heterogeneity in labeling strategies (self-report vs. community membership), dataset shift across platforms/languages, and limited prospective designs. As such, social signals are currently most valuable for population-level surveillance and research enrichment (e.g., identifying individuals who might benefit from outreach or screening), with any operational use requiring explicit consent, harm-mitigation protocols, and oversight.[16]

Fairness, Bias, and Generalizability

Bias arises from data (who is represented, which outcomes are documented) and design (which predictors and thresholds are chosen). Reviews of real-world data (RWD) models highlight gaps in fairness assessments and call for robust subgroup performance reporting and mitigation beyond simple reweighting.[2] In suicide-risk prediction specifically, site/population shifts can degrade calibration and amplify disparities; recent external validations and population-specific evaluations underscore the need for local recalibration and participatory design with impacted communities.[5,9] Rigorous

reporting (TRIPOD+AI) and structured risk-of-bias appraisal (PROBAST+AI) are essential to make these issues visible and correctable across the model lifecycle.[6],[7]

Explainability and Decision Support

Explainability is not a checkbox; it is a design requirement for safe, effective clinical decision support. In mental-health applications, reviews recommend aligning explanation methods (e.g., SHAP/feature attributions, exemplar-based explanations, counterfactuals) with the decisions clinicians must make (safety planning, follow-up cadence) and with the bandwidth of fast clinical encounters. Transparent baselines, calibrated probabilities, and error analyses are often more actionable than complex post-hoc visualizations.[2] Pragmatically, CDS trials show that simple, interruptive prompts anchored to clear thresholds can increase suicide-risk assessments and safety planning; explanation add-ons should serve, not distract from, that goal.[3]

Privacy, Security, and Governance

Mental-health data are among the most sensitive in medicine. Governance must address both model development (data minimization, consent, and de-identification) and deployment (on-device inference, audit trails). Differential-privacy and federated-learning approaches are gaining traction as practical tools to reduce centralization of raw data: reviews and proof-of-concepts in mental-health detection show promising performance while keeping text/sensor data local.[17],[18] A 2021 Delphi study on ethical digital phenotyping stresses informed consent, data protection, and participant control—principles echoed by the World Health Organization's guidance on AI for health (2021) and its 2025 recommendations for large multimodal models.[19–21] Enforcement matters too: the U.S. Federal Trade Commission's actions against a major teletherapy platform for sharing sensitive mental-health data with advertisers (with a $7.8M settlement and ongoing prohibitions) illustrate the regulatory expectations for privacy-preserving design and truthful disclosures.[22]

Reporting Standards and Evaluation: From Aucs to Practice

Transparent reporting and bias appraisal have advanced rapidly. TRIPOD+AI (BMJ 2024) updates reporting guidance for studies developing or validating prediction models using regression or ML, including detailed items on data curation, hyperparameter tuning, and evaluation.[6] PROBAST+AI (BMJ 2025) extends risk-of-bias assessment to AI-enabled prediction studies, supporting more trustworthy evidence synthesis.[7] For implementation, external validation across diverse sites and populations (including populations disproportionately affected by suicide) is the minimum bar, and randomized or quasi-experimental studies should test whether model-guided interventions change processes and outcomes.[3],[5],[8],[9]

CONCLUSION

Predictive modeling for mental health is no longer a proof-of-concept: well-calibrated, parsimonious EHR models—augmented by NLP from clinical text—can support earlier, safer care in targeted use cases like short-term suicide-risk stratification. Digital phenotyping and social-media signals are promising complements, but they require stronger external validation, engagement strategies, and explicit consent to move from pilots to practice. Equity, privacy, and transparency must anchor deployments through local recalibration, participatory design, and adherence to TRIPOD+AI and PROBAST+AI. Future work should prioritize impact evaluations over incremental AUC gains, linking risk scores to effective, acceptable interventions that improve patient-centered outcomes at scale.

REFERENCES

1. 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. doi:10.2196/27244. Available from: https://mental.jmir.org/2022/3/e27244/
2. Shortreed SM, Simon GE, Coleman KJ, et al. Complex modeling with detailed temporal predictors does not improve health records-based suicide risk prediction. npj Digit Med. 2023;6(1):47. doi:10.1038/s41746-023-00772-4. Available from: https://www.nature.com/articles/s41746-023-00772-4
3. Walsh CG, Lundahl A, Poznanovic J, et al. Risk Model–Guided Clinical Decision Support for Suicide Prevention in Outpatient Mental Health: A Randomized Clinical Trial. JAMA Netw Open. 2024;7(12):e2452371. doi:10.1001/jamanetworkopen.2024.52371. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813472
4. Edgcomb JB, Zima BT, Cowan J, et al. Electronic Health Record Phenotyping of Pediatric Suicide-Related Behaviors in Emergency Departments. JAMA Netw Open. 2024;7(10):e2442091. doi:10.1001/jamanetworkopen.2024.42091. Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2820284
5. Haroz EE, O'Keefe VM, Satterfield K, et al. Performance of Machine Learning Suicide Risk Models in a Majority American Indian Population. JAMA Netw Open. 2024;7(5):e241201. doi:10.1001/jamanetworkopen.2024.1201.

Available from: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2819145
6. Collins GS, Dhiman P, Andaur-Navarro CL, et al. TRIPOD+AI: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378. Available from: https://www.bmj.com/content/385/bmj-2023-078378
7. Moons KGM, Wolff RF, Riley RD, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505. Available from: https://www.bmj.com/content/388/bmj-2024-082505
8. Papini S, Walsh CG, Klinker MW, et al. Validation of a Multivariable Model to Predict Suicide Attempt in a Mental Health Intake Sample. JAMA Netw Open. 2024 Jul;7(7). Available from: https://pubmed.ncbi.nlm.nih.gov/38536187/
9. Yang J-H, Kim J-H, Kim J, et al. Development and external validation of logistic and penalized logistic models to predict suicide attempts: a multicenter study. J Psychiatr Res. 2024;170:236-244. doi:10.1016/j.jpsychires.2024.06.019. Available from: https://www.sciencedirect.com/science/article/pii/S0022395624002488
10. Adekkanattu P, Xu J, Luo Y, et al. Deep learning for identifying personal and family history of psychiatric disorders in clinical notes. npj Digit Med. 2024;7:86. doi:10.1038/s41746-024-01266-7. Available from: https://www.nature.com/articles/s41746-024-01266-7
11. Levis M, Westgate CL, Gui J, Watts BV, Shiner B. Using natural language processing to evaluate temporal changes in suicidal ideation documentation and its predictive value. Psychol Med. 2024;Epub ahead of print. doi:10.1017/S0033291724001375. Available from: https://pubmed.ncbi.nlm.nih.gov/38784261/
12. Young J, Bishop S, Humphrey C, Pavlacic JM. A review of natural language processing in the identification of suicidal behavior. J Affect Disord Rep. 2023;12:100507. doi:10.1016/j.jadr.2023.100507. Available from: https://www.sciencedirect.com/science/article/pii/S2666915323000458
13. Zhou J, Xu R, Wang X, et al. Predicting Psychotic Relapse in Schizophrenia With Mobile Sensing and Unsupervised Behavioral Representations. JMIR Mhealth Uhealth. 2022;10(4):e31006. doi:10.2196/31006. Available from: https://mhealth.jmir.org/2022/4/e31006/
14. Oudin A, Le Sommer J, De Angel V, et al. Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. J Med Internet Res. 2023;25:e44502. doi:10.2196/44502. Available from: https://www.jmir.org/2023/1/e44502/
15. Zhang Y, Gao M, Hu J, et al. The comprehensive clinical benefits of digital phenotyping. npj Digit Med. 2025;8. doi:10.1038/s41746-025-01602-5. Available from: https://www.nature.com/articles/s41746-025-01602-5
16. Huang D, Emedom-Nnamdi P, Onnela J-P, Van Meter A. Development and initial evaluation of a digital phenotype collection tool for adolescent mental health. JMIR Form Res. 2024;8(1):e59623. doi:10.2196/59623. Available from: https://formative.jmir.org/2024/1/e59623
17. Khalil SS, Arora A, Ahmad A, et al. Federated learning for privacy-preserving depression detection with multilingual language models. Patterns. 2024;5(7):100990. doi:10.1016/j.patter.2024.100990. Available from: https://www.sciencedirect.com/science/article/pii/S2666389924001627
18. Grataloup A, Frossard P, Jayaraman D. A systematic survey on the application of federated learning in mental health and human activity recognition. Front Digit Health. 2024;6:1495999. doi:10.3389/fdgth.2024.1495999. Available from: https://www.frontiersin.org/articles/10.3389/fdgth.2024.1495999/full
19. Martinez-Martin N, Dasgupta I, Carter A, et al. Ethical Development of Digital Phenotyping Tools for Mental Health: Delphi Study. JMIR Mhealth Uhealth. 2021;9(7):e27343. doi:10.2196/27343. Available from: https://mhealth.jmir.org/2021/7/e27343/
20. World Health Organization. Ethics and governance of artificial intelligence for health. Geneva: WHO; 2021. Available from: https://www.who.int/publications/i/item/9789240029200
21. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multimodal models. Geneva: WHO; 2025. Available from: https://www.who.int/publications/i/item/9789240084759
22. U.S. Federal Trade Commission. FTC Gives Final Approval to Order Banning BetterHelp from Sharing Sensitive Health Data for Advertising, Requiring It to Pay $7.8 Million. Press release; July 14, 2023. Available from: https://www.ftc.gov/news-events/news/press-releases/2023/07/ftc-gives-final-approval-order-banning-betterhelp-sharing-sensitive-health-data-advertising