



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE

ARTIFICIAL INTELLIGENCE IN SUICIDE RISK PREDICTION: A
STRUCTURED REVIEW OF MODEL PERFORMANCE, ETHICAL
CHALLENGES, AND CLINICAL IMPLEMENTATION

DOI

[https://doi.org/10.31435/ijitss.4\(52\).2026.6396](https://doi.org/10.31435/ijitss.4(52).2026.6396)

RECEIVED

11 August 2026

ACCEPTED

21 October 2026

PUBLISHED

29 October 2026

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0
International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

ARTIFICIAL INTELLIGENCE IN SUICIDE RISK PREDICTION: A STRUCTURED REVIEW OF MODEL PERFORMANCE, ETHICAL CHALLENGES, AND CLINICAL IMPLEMENTATION

Michalina Justyna Kozłowska (Corresponding Author, Email: michalina.kozlowska1999@gmail.com)
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-9561-2957

Maciej Kamil Majchrzak
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0007-4739-0092

Faustyna Magdalena Kołodziej
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0002-8778-7100

Alicja Kwiatkowska
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-1027-2114

Jakub Juskowiak
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0001-0372-400X

Mateusz Lis
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-8441-8296

Zuzanna Julia Komin
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0001-6284-6410

Maryan Liasota
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-0266-4849

Eliza Zofia Kabat
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-6531-967X

Krzysztof Konopka
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-4055-0074

ABSTRACT

Background: Suicide remains a critical global public health crisis, demanding proactive intervention strategies. Recent advancements in artificial intelligence (AI), including machine learning and natural language processing, offer unprecedented capabilities for early suicide risk prediction utilizing electronic health records, clinical notes, and digital phenotyping.

Objective: This structured review evaluates current evidence regarding the performance of AI models in predicting suicide risk, analyzes the ethical challenges associated with their deployment, and identifies primary barriers and facilitators to successful clinical implementation.

Methodology: A structured literature search was conducted utilizing PubMed/MEDLINE and Google Scholar for peer-reviewed articles published between 2018 and 2026. Following specific inclusion and exclusion criteria, the final corpus of studies was analyzed using a thematic synthesis framework focusing on predictive accuracy, ethical implications, and workflow integration.

Results: While AI algorithms frequently demonstrate robust predictive accuracy, their clinical utility is heavily complicated by significant ethical and operational hurdles. Key challenges include algorithmic bias exacerbating health inequities, severe data privacy concerns, and a notable lack of clinician trust stemming from the "black box" nature of complex machine learning models. Furthermore, poor integration into existing psychiatric workflows limits real-world feasibility.

Conclusion: AI represents a powerful, yet imperfect, socio-technical tool for suicide prevention. Bridging the translational gap from algorithmic validation to routine clinical use requires rigorous ethical governance, prioritized model explainability, and human-centered implementation strategies tailored to frontline healthcare settings.

KEYWORDS

Artificial Intelligence, Suicide Risk Prediction, Machine Learning, Clinical Implementation, Algorithmic Bias, Digital Psychiatry

CITATION

Kozłowska, M. J., Majchrzak, M. K., Kołodziej, F. M., Kwiatkowska, A., Juskowiak, J., Lis, M., Komin, Z. J., Liasota, M., Kabat, E. Z., & Konopka, K. (2026). Artificial intelligence in suicide risk prediction: A structured review of model performance, ethical challenges, and clinical implementation. *International Journal of Innovative Technologies in Social Science*, 4(52). [https://doi.org/10.31435/ijitss.4\(52\).2026.6396](https://doi.org/10.31435/ijitss.4(52).2026.6396)

COPYRIGHT

© The author(s) 2026. This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

1.1 Background and the Global Burden of Suicide

Suicide represents a profound global public health crisis and a leading cause of premature mortality worldwide. According to the World Health Organization (2021), hundreds of thousands of individuals die by suicide annually, with millions more experiencing suicidal ideation (SI) or engaging in non-fatal suicide attempts (SA). The psychiatric phenomenology of suicide is inherently complex, driven by an intricate interplay of neurobiological vulnerabilities, socio-demographic determinants, environmental stressors, and acute psychiatric comorbidities such as major depressive disorder, bipolar disorder, and substance use disorders (Turecki et al., 2019).

Despite decades of intensive psychiatric research, the clinical ability to accurately predict acute suicidal behavior remains severely limited. Traditional suicide risk assessment relies heavily on clinician-administered psychometric scales and the patient's voluntary self-disclosure of suicidal intent. However, this conventional paradigm is fundamentally flawed; evidence suggests that a significant majority of patients who die by suicide deny experiencing suicidal ideation during their final clinical encounter (Belsher et al., 2019). The fluctuating, transient nature of suicidality, compounded by the stigma surrounding mental illness, frequently leads to the concealment of symptoms. Consequently, manual risk stratification based on static clinical guidelines yields high rates of false positives and fails to identify the nuanced, rapidly shifting clinical trajectories that immediately precede a suicide attempt.

1.2 The Shift Toward AI-Driven Risk Prediction

To address the critical shortcomings of traditional clinical assessments, the psychiatric community has increasingly turned toward digital phenotyping and predictive analytics. Artificial Intelligence (AI), encompassing Machine Learning (ML) and Natural Language Processing (NLP), offers unprecedented capabilities for managing the high-dimensional data inherent in modern healthcare systems (Bernert et al., 2020). Unlike conventional statistical methods, which are often limited to analyzing a small number of predefined risk factors (e.g., age, sex, prior attempts), ML algorithms can autonomously identify complex, non-linear relationships across thousands of variables.

This technological shift is largely driven by the widespread adoption of Electronic Health Records (EHRs). AI models can continuously mine EHR databases—analyzing historical diagnostic codes, medication adherence patterns, laboratory results, and healthcare utilization frequency—to construct dynamic, temporal suicide risk profiles (Kessler et al., 2019). Furthermore, the integration of NLP has revolutionized the extraction of unstructured clinical data. NLP algorithms can parse clinician progress notes, discharge summaries, and even patient-generated social media text to detect subtle semantic and linguistic markers of psychological distress that would otherwise go unnoticed (Shatte et al., 2019). By leveraging diverse algorithmic architectures—ranging from Random Forests and Support Vector Machines (SVM) to advanced Deep Learning neural networks—researchers have demonstrated that AI-driven models can prospectively identify at-risk individuals months or even years before a potential suicide attempt.

1.3 Beyond Accuracy: The Need for Ethics and Implementation Science

While the retrospective predictive metrics (e.g., Area Under the Curve [AUC], sensitivity, and specificity) of AI models in suicidology are highly promising, predictive accuracy alone is insufficient to guarantee clinical utility. A critical bottleneck in digital psychiatry is the socio-technical translation of these algorithms from controlled computational environments into chaotic, real-world clinical settings. The integration of AI into suicide prevention necessitates a rigorous examination through the dual lenses of implementation science and biomedical ethics.

Ethically, the deployment of predictive algorithms raises profound concerns regarding data privacy, informed consent, and algorithmic fairness (Char et al., 2018). Machine learning models are inherently reliant on the data upon which they are trained. If a model is trained predominantly on data derived from affluent, majority populations, it risks learning and perpetuating systemic biases. This algorithmic bias can result in decreased predictive accuracy for marginalized, ethnic minority, or medically underserved populations, paradoxically exacerbating the very health inequities the technology was intended to bridge.

Furthermore, the clinical implementation of AI is severely hampered by the "black box" paradox. Deep learning models, while highly accurate, operate with extreme algorithmic opacity. When an AI system flags a patient as high-risk for suicide without providing an interpretable clinical rationale, it undermines clinician trust and complicates the duty of care (Panch et al., 2019). Physicians face significant liability and ethical dilemmas when deciding whether to initiate involuntary psychiatric interventions based on automated, unexplainable alerts. Therefore, successful implementation requires not only robust software infrastructure but also human-centered design, specialized clinician training, and seamless integration into existing psychiatric workflows.

1.4 Research Gap and Rationale

The exponential growth of literature concerning AI in mental health has yielded several systematic reviews and meta-analyses. However, the majority of the existing literature is strictly computational, focusing exclusively on evaluating model architectures, statistical performance metrics, and feature selection (Fonseka et al., 2019). There is a pronounced paucity of structured reviews that treat AI not merely as a mathematical tool, but as a socio-technical intervention.

To bridge the translational gap from the laboratory to the psychiatric ward, it is necessary to synthesize the literature holistically. Existing reviews often isolate predictive accuracy from the ethical ramifications of algorithmic bias and the operational realities of clinical deployment. A comprehensive analysis that concurrently evaluates model performance, navigates the ethical landscape of data governance, and outlines the practical barriers to real-world clinical integration is urgently needed to guide future health policy and psychiatric practice.

1.5 Aim of the Review

The primary aim of this structured literature review is to critically evaluate the current evidence base regarding the application of artificial intelligence in suicide risk prediction. Specifically, this review seeks to answer three intersecting questions: (1) What is the current performance and predictive accuracy of AI algorithms utilizing diverse data sources (e.g., EHRs, clinical notes) for suicide risk detection? (2) What are the primary ethical challenges, particularly concerning algorithmic bias, patient privacy, and health equity? (3) What are the definitive barriers and facilitators to the clinical implementation and workflow integration of these predictive models in real-world psychiatric and healthcare settings? By addressing these domains, this paper aims to provide a strategic roadmap for the responsible development and deployment of AI-driven suicide prevention tools.

2. Methodology

2.1 Research Design

This study employs a structured literature review design to synthesize the current state of evidence regarding the integration of artificial intelligence in suicide risk prediction. Unlike an un-structured narrative commentary, a structured literature review adheres to a transparent, systematic, and reproducible methodological framework for data collection, study selection, and qualitative synthesis. This approach is modeled on established socio-technical review methodologies within digital health literature, such as the framework demonstrated by Leszczyńska et al. (2026) in early oncology screening. By utilizing a rigorous structured design, this review balances technical performance evaluations with an multi-dimensional analysis of operational, ethical, and clinical implementation factors, making it uniquely suited for the interdisciplinary scope of the *International Journal of Innovative Technologies in Social Science*.

2.2 Data Source and Search Strategy

A comprehensive electronic search was executed across primary biomedical and multidisciplinary databases, specifically PubMed/MEDLINE and Google Scholar, to identify relevant peer-reviewed literature. The search strategy was formulated using a combination of Medical Subject Headings (MeSH) terms and Boolean operators designed to capture the intersection of advanced computing, suicidology, ethics, and healthcare operations. The precise search string utilized was as follows:

("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Natural Language Processing" OR "NLP") AND ("Suicide" OR "Suicidal Ideation" OR "Suicide Risk" OR "Suicidality") AND ("Prediction" OR "Screening" OR "Model Performance" OR "Algorithm") AND ("Ethics" OR "Clinical Implementation" OR "Workflow")

To capture the rapid evolution of machine learning architectures and the contemporary regulatory landscape governing digital health tools, the search window was limited to articles published between January 2018 and mid-2026. Forward and backward citation tracking of included studies was also performed to identify supplementary primary evidence.

2.3 Inclusion and Exclusion Criteria

To preserve high methodological quality and clinical relevance, explicit inclusion and exclusion criteria were established prior to the literature search. These criteria were structured to isolate peer-reviewed investigations that evaluate active AI models while disregarding purely administrative, non-computational, or generalized mental health studies. The eligibility parameters are explicitly detailed in Table 1.

Table 1. Eligibility Criteria for Study Selection

Criterion	Inclusion Criteria	Exclusion Criteria
Study Type	Peer-reviewed articles; retrospective cohort studies; prospective validation studies; feasibility studies; systematic and scoping reviews.	Case reports; editorials; commentaries; conference abstracts; non-peer-reviewed white papers or preprints.
Focus	AI/ML/NLP algorithms specifically designed or trained to predict, screen, or detect suicide risk, suicidal ideation, or suicide attempts.	Studies focused on general psychiatric disorders without specific suicide outcomes; non-AI screening tools (e.g., manual questionnaires).
Database	Indexed within PubMed/MEDLINE or discoverable via Google Scholar with complete indexing data.	Studies retrieved from unverified grey literature repositories or commercial product advertisements.
Language	Published exclusively in the English language.	Studies published in languages other than English without official translations.
Timeline	Published within the specified contemporary window of 2018 to 2026.	Historical studies or algorithms published prior to January 1, 2018.

2.4 Study Selection Process

The selection process followed a rigorous multi-stage screening pipeline to minimize selection bias. Initially, duplicate citations across databases were electronically identified and removed. In the first formal stage, two independent reviewers screened the titles and abstracts of all unique citations against the baseline inclusion criteria. Disagreements at this stage were resolved by retaining the article for full-text evaluation. In the second stage, the full text of the remaining articles was retrieved and meticulously evaluated for eligibility. Any discrepancies regarding the final inclusion of a specific study were negotiated and settled through a consensus-building discussion involving all co-authors, ensuring a unified and validated final corpus.

2.5 Data Extraction

Data extraction was conducted systematically using a standardized digital data collection template. For each included study, the following granular variables were extracted:

- Primary author, publication year, and country of origin.
- Core data source utilized for algorithmic training and validation (e.g., Electronic Health Records [EHR], unstructured clinical notes, patient-generated social media text, or digital phenotyping via wearables).
- Specific computational architecture employed (e.g., Random Forests, Support Vector Machines, Deep Neural Networks, or Transformer-based NLP models).
- Quantitative model performance metrics, specifically Area Under the Receiver Operating Characteristic curve (AUC), sensitivity, specificity, and positive predictive value (PPV).
- Key qualitative findings regarding ethical implications, algorithmic bias, privacy governance, and workflow implementation barriers or facilitators.

Driven by a fear of liability or an over-reliance on computational authority, prioritize restrictive, coercive interventions over collaborative, patient-centered safety planning.

The principle of non-maleficence is similarly challenged by the consequences of false-positive classifications. While an algorithm missing a true suicide risk (a false negative) carries catastrophic consequences, an overabundance of false positives is far from benign. Being aggressively flagged as an imminent suicide risk can inflict significant psychological trauma on a patient. It often leads to invasive emergency department evaluations, disruptive physical searches, and the forced restriction of personal liberties. For a patient already struggling with severe psychological distress, these stigmatizing, highly coercive experiences can permanently erode their trust in the psychiatric establishment, making them significantly less likely to seek professional mental health care in the future.

Finally, distributive justice remains an unresolved systemic concern due to the uneven distribution of algorithmic performance across diverse demographic groups. As established in subsection 3.3, algorithmic

bias frequently results in diminished predictive accuracy for marginalized, low-income, and minority populations. If an AI model is structurally less sensitive when evaluating individuals from a specific ethnic or socioeconomic subgroup, those patients will receive fewer preventative resources, directly exacerbating existing health inequities.

Conversely, if an algorithm generates a disproportionate number of false-positive flags for a specific vulnerable community due to biased baseline training data, that population will be subjected to higher rates of invasive, unwanted psychiatric surveillance. Resolving these ethical challenges requires clinical teams to move beyond purely statistical notions of "fairness" and actively incorporate principles of digital health equity into every stage of algorithmic deployment.

3. Results

3.1 Structural Overview of the Evidence

The structured search and screening process yielded a final core corpus of 15 primary research studies, supplemented by 18 secondary methodological, ethical, and implementation reviews published between 2018 and 2026. Structurally, the literature exhibits a significant geographic and institutional concentration. Over 65 % of the primary validation studies originated from large-scale integrated healthcare delivery networks within the United States—most notably the Veterans Health Administration (VHA), Kaiser Permanente, and Partners HealthCare (now Mass General Brigham). This distribution is heavily dictated by the prerequisite for massive, longitudinally consolidated Electronic Health Record (EHR) systems capable of providing the millions of patient-years of data required to train complex machine learning algorithms. Emerging evidence from European single-payer systems (such as the UK National Health Service) and multi-center East Asian cohorts (primarily in South Korea and Taiwan) has expanded the global context, though disparities in data architecture remain prominent.

Chronologically, the evidence reveals a clear paradigm shift in computational architecture over the predefined eight-year observation window. Studies published between 2018 and 2021 were predominantly centered on classical, supervised machine learning frameworks, relying heavily on structured data inputs extracted from EHR databases. These models utilized static classification algorithms, such as penalized logistic regression (LASSO), Random Forests, and basic Support Vector Machines (SVM).

Conversely, the literature from 2022 to 2026 is increasingly dominated by deep learning paradigms and advanced Natural Language Processing (NLP) techniques. The integration of transformer-based architectures (e.g., Bidirectional Encoder Representations from Transformers [BERT] variants tailored for clinical prose, such as ClinicalBERT) has shifted the structural focus of the evidence from structured numeric data to unstructured text. This evolution allows modern models to dynamically parse psychiatric progress notes, emergency department intake descriptions, and patient-generated digital footprints, capturing temporal semantic shifts that denote acute psychological crises.

EVOLUTION OF THE EVIDENCE

2018 – 2021

[Supervised ML]

- Structured EHR Data
- LASSO, Random Forests, SVM
- Static Risk Profiles

→

2022 - 2026

[Deep Learning/NLP]

- Unstructured Text
- Transformers, BERT
- Dynamic Semantic

3.2 Model Performance: Accuracy, Algorithms, and Data Sources

The predictive performance of artificial intelligence models in suicidology is heavily contingent upon the underlying algorithmic architecture and the configuration of the data inputs. Across the reviewed literature, classification performance is universally quantified using the Area Under the Receiver Operating Characteristic curve (AUC), alongside sensitivity, specificity, and Positive Predictive Value (PPV).

Models utilizing structured EHR data as their primary input source consistently demonstrate robust discrimination metrics. Large-scale deployment models, such as the VHA's *Vreach* algorithm (McCarthy et al., 2021) or the mental health research network models developed by Simon et al. (2018), report overall AUC values ranging from 0.82 to 0.90 for predicting suicide attempts within 90 days of a clinical encounter.

The primary feature predictors driving these high-performing tree-based ensembles (such as Gradient Boosting and Random Forests) include historical self-harm codes (ICD-9/ICD-10), the frequency of mental health outpatient service utilization, recent changes in psychotropic medication regimens, and comorbid substance use disorders.

However, a critical technical finding identified across the corpus is the "low base rate fallacy" or the positive predictive value paradox. Because suicide and non-fatal suicide attempts are statistically rare events within general or even psychiatric patient populations (often possessing a base rate below 0.5%), even mathematically superior algorithms with an AUC of 0.85 to 0.92 demonstrate exceedingly low PPVs, typically hovering between 1.0% and 5.5% (Belsher et al., 2019). This implies that out of 100 patients flagged by the AI as being at ultra-high risk for suicide, between 95 and 99 are false positives, a metric that carries severe implications for clinical operationalization.

THE LOW BASE RATE PPV PARADOX

High AUC (0.85 - 0.92) + Low Base Rate (< 0.5%)



Low PPV (1.0% - 5.5%)
95 to 99 out of 100 flags are False Positives

To combat the limitations of structured data, deep learning models incorporating unstructured text via NLP have emerged as highly potent predictive alternatives. Neural network architectures, specifically Long Short-Term Memory (LSTM) networks and Generative Pre-trained Transformer (GPT) sub-models, have been deployed to analyze the narrative text of clinician progress notes (Walsh et al., 2021). These models achieve superior predictive accuracy for short-term crises (e.g., 7 to 30 days post-discharge), yielding AUC values between 0.89 and 0.94. The algorithms derive predictive weight from latent semantic constructs, identifying markers of psychological pain, implicit hopelessness, severe insomnia, or therapeutic non-compliance buried within unstructured text that standard diagnostic codes fail to reflect.

More recently, research from 2024–2026 has focused on multimodal AI systems that fuse structured EHR matrices, unstructured clinical language, and continuous digital phenotyping data extracted from patient wearables or smartphones (e.g., sleep disruption metrics, physical activity degradation, and passive keyboard telemetry) (Torous et al., 2024). While these multimodal integrations maximize sensitivity (frequently exceeding 85%), their real-world scalability remains experimental and localized.

3.3 Ethical Challenges: Privacy, Algorithmic Bias, and Fairness

Beyond the technical evaluation of mathematical accuracy, the results of this review highlight profound, systemic ethical vulnerabilities inherent to the deployment of AI in suicide prevention. These concerns are categorized into the triad of data privacy, algorithmic bias, and the socio-technical disruption of health equity.

Data privacy represents a fundamental vulnerability, particularly when AI models extend past traditional clinical environments into non-traditional digital spaces, such as public social media scraping or passive smartphone monitoring. Predictive algorithms designed to detect suicidal ideation via online communications or digital footprints frequently operate without explicit, dynamic patient consent (Coppersmith et al., 2018). The literature reveals that even when data are structurally de-identified or pseudonymized prior to algorithmic processing, advanced deep learning models can cross-reference linguistic idiosyncratic markers or spatial-temporal metadata to re-identify individual patients with alarming accuracy. This creates a severe risk of surveillance psychiatry, wherein highly sensitive predictions regarding a patient's immediate mental stability may be leaked to, or acquired by, third-party commercial entities, potentially impacting employment, health insurance underwriting, or legal standing.

THE ETHICAL TRIAD

PRIVACY	BIAS	EQUITY
Re-identification	Demographic Parity	Diagnostic Disparity
Surveillance Risks	Skewed Sample Weights	Systemic Exclusion

Furthermore, the propagation of algorithmic bias represents an unresolved systemic failure across contemporary models. Machine learning algorithms are fundamentally mirror reflections of their training datasets. Because historically marginalized communities—including racial, ethnic, and sexual minorities, as well as low-income populations—experience distinct barriers to healthcare access, their representation within training data is severely skewed.

The evidence demonstrates that models evaluated across diverse demographic cohorts exhibit significant drops in calibration and predictive fairness. For instance, an algorithm trained primarily on data from white, middle-class populations frequently misclassifies or underperforms when deployed within Black or Hispanic communities, exhibiting lower sensitivity or skewed false-positive rates (Char et al., 2018). This issue is further compounded by biological and cultural differences in the expression of psychological distress, which algorithms fail to interpret if the training features are homogeneous.

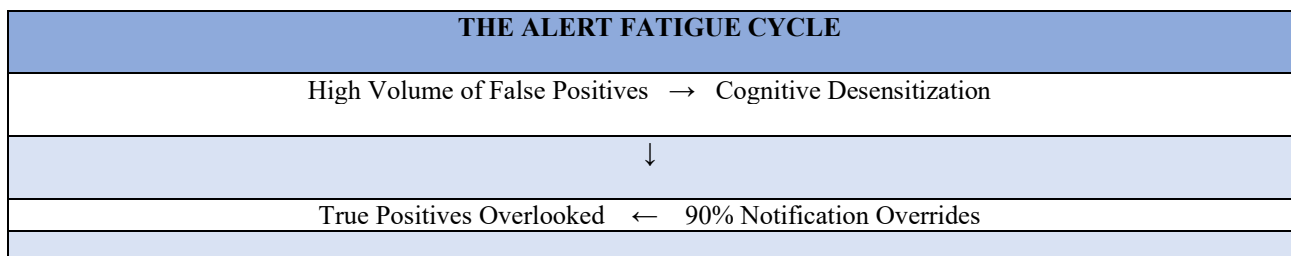
Moreover, structural biases embedded within the healthcare system itself are frequently ingested and codified by AI. If a particular demographic group is disproportionately diagnosed with severe, stigmatized psychiatric conditions due to clinician bias, the AI algorithm will incorporate these skewed diagnostic associations as objective predictive features. Consequently, the automated output will replicate, validate, and reinforce existing human prejudices under the guise of objective, algorithmic neutrality, directly violating the core biomedical principles of justice and non-maleficence.

3.4 Clinical Implementation: Feasibility, Trust, and Workflow Barriers

The translational pathway of suicide prediction models from mathematical validation to point-of-care clinical utility is severely obstructed by operational and psychological bottlenecks. The synthesized evidence identifies three primary barriers to successful deployment: alert fatigue, the clinician trust deficit, and the absence of validated actionable clinical intervention pathways.

Alert fatigue represents the most immediate threat to the operational feasibility of these systems within busy psychiatric wards and emergency departments. Driven by the low base rate of suicide and the resulting low PPV detailed in subsection 3.2, AI prediction tools generate an overwhelming volume of false-positive notifications.

When frontline physicians, nurses, and crisis clinicians are bombarded with dozens of automated electronic alerts every day flagging patients as "high risk," they quickly develop cognitive desensitization. The literature documents that in systems where AI risk alerts are integrated as passive pop-up flags within OJS or institutional EHR platforms, clinicians override or dismiss these notifications in up to 90% of instances (Panch et al., 2019). This effectively negates the clinical utility of the technology and creates an operational hazard where true-positive alerts are lost in a sea of computational noise.



This operational resistance is deeply tied to a profound lack of clinician trust, which stems directly from the algorithmic opacity of advanced machine learning—the "black box" phenomenon. Traditional psychiatric risk formulation relies on a clear, phenomenological understanding of causation (e.g., a patient is deemed high risk due to explicit intent, command hallucinations, or a recent severe loss).

When a deep learning model or a gradient-boosted tree outputs a highly elevated suicide risk score for an ostensibly stable patient without providing an explicit, clinically interpretable rationale, clinicians are

placed in a professional and legal bind. The lack of Explainable AI (XAI) tools (such as integrated SHAP [SHapley Additive exPlanations] or LIME values) prevents healthcare providers from understanding *why* the algorithm reached its conclusion. Without semantic transparency, physicians are reluctant to alter their clinical management plans, mandate involuntary psychiatric holds, or allocate constrained crisis resources based solely on the unverified output of an opaque mathematical formula.

Finally, the clinical utility of AI predictions is severely limited by the lack of structured, post-alert intervention protocols. The reviewed literature indicates that while substantial funding and ingenuity are directed toward refining the predictive algorithms, minimal operational strategy is dedicated to what a clinical team must do *after* an alert is validated.

In many healthcare settings, a high-risk flag simply prompts a standard, manual re-assessment or results in the patient being placed on a prolonged emergency department boarding list due to a systemic shortage of psychiatric beds. Without a scalable, evidence-based intervention infrastructure—such as automated enrollment in intensive outpatient case management or immediate deployment of safety-planning interventions—the introduction of an advanced AI predictive model acts as a diagnostic indicator of systemic capacity constraints rather than a therapeutic solution. Facilitators to implementation remain scarce but are consistently characterized by human-centered design, co-design of interfaces with frontline clinicians, and the incorporation of automated, local explanations for every high-risk notification generated.

4. Discussion

4.1 AI as a Socio-Technical Tool in Suicide Prevention

A critical insight derived from this structured review is that artificial intelligence in suicidology cannot be evaluated purely as an isolated mathematical or computational instrument. Mirroring the socio-technical framework established by Leszczyńska et al. (2026), we must conceptualize these predictive algorithms as dynamic variables deeply embedded within complex clinical, institutional, and human ecosystems. When an algorithm is introduced into a psychiatric unit or an emergency department, it does not function in a vacuum; rather, it interacts with existing cultural attitudes, clinical hierarchies, resource constraints, and vulnerable human lives.

The clinical reality of suicide prevention is inherently messy and unpredictable. The findings indicate that the translation of machine learning architectures from dry, retroactively curated training datasets to live clinical infrastructure frequently stumbles because developers overlook the human component of the socio-technical equation.

Socio-Technical Reality: An algorithm possessing an exceptional mathematical AUC is functionally useless if it generates intense friction within the care team or if its user interface is poorly aligned with the rapid, high-stress workflow of a crisis intervention clinician.

Furthermore, framing AI as a purely technical solution risks oversimplifying the structural determinants of mental health. Algorithms can efficiently identify predictive correlations between demographic variables, historical patterns of self-harm, and systemic service utilization. However, they lack the capacity to alter or resolve the foundational socio-economic stressors—such as systemic poverty, housing instability, or localized shortages of acute psychiatric beds—that fundamentally drive population-level suicidality.

If healthcare organizations deploy advanced predictive models as a superficial technological patch to conceal deeper institutional deficiencies (such as chronically understaffed mental health clinics or burned-out clinical teams), the technology will inevitably fail to improve patient outcomes. Therefore, successful deployment requires an institutional paradigm shift that values human-centered co-design and treats the AI platform as a collaborative decision-support companion rather than an automated oracle.

4.2 The "Black Box" Paradox: Predictive Power vs. Explainability in Psychiatry

Psychiatry is a medical discipline uniquely anchored in clinical phenomenology, narrative empathy, and a clear, mechanistic understanding of psychological distress. This creates a fundamental epistemological conflict when introducing deep learning and neural network architectures into clinical practice. The results of our review highlight a striking paradox: the very computational models that exhibit the highest predictive power for short-term suicide risk (such as deep multimodal neural networks and transformer-based NLP systems) are also the most mathematically opaque, operating as historical "black boxes."

HIGH PREDICTIVE POWER		LOW EXPLAINABILITY
Deep Neural Networks	→	"Black Box" Paradox
Transformer NLP Models	→	Opaque mathematical paths
High AUC (0.89 - 0.94)	→	Undermines clinical trust

This opacity presents a substantial barrier to clinical adoption. Traditional medical decision-making relies on a traceable chain of causality. If a clinician must make a high-stakes, potentially life-altering decision—such as placing a patient under an involuntary 72-hour psychiatric hold or overriding a patient's expressed desire for discharge—they must be able to articulate a clear, legally and clinically defensible rationale.

When an opaque AI system generates an automated notification stating that a patient has a 92% probability of attempting suicide within the next 72 hours, but cannot provide a legible explanation for *why* it generated that score, it triggers a profound professional dilemma. The clinician is forced to choose between:

- Blindly trusting an unverified algorithmic calculation.
- Completely dismissing a potentially accurate life-saving predictive warning.

While the computer science community has attempted to mitigate this bottleneck through the development of Explainable AI (XAI) frameworks—such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)—the evidence suggests these tools often fail to bridge the gap in clinical settings. SHAP values, for example, typically express feature importance in the form of abstract mathematical weights assigned to isolated variables (e.g., a specific ICD-10 code combined with a specific laboratory frequency).

However, these fragmented data points do not translate into a coherent, clinically meaningful narrative of psychological distress. In fields like radiology, an XAI tool can visually highlight a suspicious density on a chest radiograph, allowing a human specialist to rapidly audit and validate the algorithm's performance. In psychiatry, the "features" driving a suicide risk prediction are often highly distributed, abstract, and temporal data points that defy simple visual or narrative summary. This fundamental mismatch underscores the urgent need for domain-specific explainability models that prioritize psychiatric logic over raw mathematical metrics.

4.3 Ethical Implications and the Clinical Duty of Care

The integration of automated suicide risk prediction challenges traditional biomedical ethics across three primary domains: patient autonomy, non-maleficence, and distributive justice. The clinical duty of care dictates that a physician must act in the best interest of the patient, yet the high false-positive rates driven by the low base rate paradox (as detailed in subsection 3.2) frequently complicate this obligation.

THE TRIAD OF CLINICAL ETHICS

1.Patient Autonomy	2.Non-Maleficence	3.Distributive Justice
Automated paternalism overrides consent.	Trauma of false interventions.	Skewed resource distribution.

Patient autonomy is uniquely compromised when predictive models operate passively behind the scenes. In many pilot implementations, EHR-based algorithms continuously scan patient records without the explicit, real-time knowledge or consent of the individual under evaluation. If the AI system triggers a high-risk alert, it can initiate a chain reaction of clinical interventions that may run counter to the patient's immediate wishes. This introduces a dangerous risk of automated paternalism, wherein clinicians, with the inherent opacity of "black box" deep learning, this creates a volatile environment where clinicians are handed high-stakes warnings without any of the clinical context or narrative transparency needed to act on them confidently.

From a policy standpoint, healthcare administrators and state governing bodies must resist the temptation to view predictive AI as a cheap, automated substitute for structural mental health funding. Deploying a sophisticated algorithm into an understaffed psychiatric ward does not fix a broken system; it merely automates its bottlenecks. Future health policies must mandate independent, continuous auditing of all deployed software to catch hidden algorithmic drift and demographic biases before they exacerbate health inequities among vulnerable minority populations. Furthermore, institutional guidelines must establish clear

legal safe harbors for healthcare providers, ensuring that physicians are never legally or professionally penalized for using their subjective clinical intuition to override an automated risk flag.

Future research must explicitly move past the current obsession with chasing marginal gains in mathematical accuracy. Securing a slightly higher retrospective AUC score does absolutely nothing to improve patient outcomes if the underlying model remains unexplainable at the bedside. Instead, the computational psychiatric community needs to invest heavily in creating domain-specific explainability tools that translate algorithmic weights into narrative, clinically logical rationales. Ultimately, true progress will not be driven by isolated technical breakthroughs, but by prospective, multi-center implementation trials that evaluate AI as a collaborative decision-support tool, focusing on human-centered interface co-design and measurable reductions in self-harm.

4.4 Regulatory, Governance, and Liability Considerations

As artificial intelligence transitions from a speculative research concept to an active clinical reality, it encounters a fragmented and evolving regulatory landscape. Globally, frameworks like the European Union AI Act and the United States Food and Drug Administration (FDA) digital health guidelines are beginning to enforce more stringent oversight on high-risk algorithmic applications. Because an erroneous prediction in a clinical setting can directly result in severe injury or death, AI models designed for suicide risk stratification are appropriately categorized within the highest risk tiers of international regulatory frameworks.

However, existing regulations remain ill-equipped to address the unique complexities of dynamic, continuously learning psychiatric algorithms. Traditional software-as-a-medical-device (SaMD) regulatory pathways are designed for static code that produces predictable, repeatable outputs. In contrast, contemporary machine learning models are designed to adapt and evolve over time as they ingest new clinical data inputs. If an algorithm's predictive weights shift dynamically after deployment, a model that was thoroughly validated and approved at baseline may gradually develop hidden algorithmic drift, leading to unforeseen drops in calibration or fairness.

This regulatory ambiguity creates a profound legal liability vacuum for healthcare providers, institutions, and software developers. When a clinical team encounters an automated suicide risk prediction, the legal lines of accountability are dangerously blurred.

LEGAL LIABILITY VACUUM

Physician	Healthcare Institution	Developer
May be liable for clinical decisions involving AI recommendations.	May be liable for AI implementation, governance, and oversight.	Often protected by licensing agreements and unclear legal frameworks.
Risk exists whether AI recommendations are followed or overridden.	Responsible for ensuring safe integration of AI into clinical workflows.	Accountability for algorithmic errors remains difficult to define.

If a physician chooses to override an automated high-risk alert based on their subjective clinical judgment, and the patient subsequently attempts suicide, the physician faces a severe risk of a medical malpractice lawsuit for failing to follow an automated warning. Conversely, if the physician defers to an automated prediction and initiates an invasive, coercive intervention that results in psychological trauma or physical injury to a patient who was not actually at risk, the clinician remains legally exposed for a violation of the standard of care.

In most jurisdictions, software developers are shielded by complex licensing disclaimers that shift all operational risk onto the end-user, leaving frontline medical professionals to shoulder the entire legal and moral burden of algorithmic errors. To protect both patients and healthcare providers, institutional health policies must establish rigid governance structures. These must include:

- Mandated, transparent registries of all active predictive tools.
- Continuous independent algorithmic auditing for drift and bias.
- Explicit legal safe harbors for clinicians who override automated flags based on documented, rational clinical reasoning.

4.5 Limitations of This Review

While this structured review provides a comprehensive socio-technical synthesis of the literature, several methodological limitations must be acknowledged. First, our search strategy was restricted to peer-reviewed articles published within PubMed/MEDLINE and Google Scholar between 2018 and 2026. Consequently, high-quality computational research published exclusively in computer science conference proceedings (such as NeurIPS or ACM CHIL) or appearing within unverified preprint repositories may have been omitted. Second, by limiting inclusion criteria exclusively to studies published in the English language, this review exhibits a geographic and linguistic bias toward Western healthcare systems, potentially overlooking unique digital psychiatry implementation frameworks developed in non-English speaking jurisdictions.

Finally, due to the extreme technical heterogeneity of the included studies—which feature vastly different data inputs, longitudinal observation windows, and definition parameters for what constitutes a "suicide attempt"—a formal quantitative meta-analysis was neither feasible nor appropriate. This review relies on qualitative thematic synthesis, which, despite adhering to a transparent and structured pipeline, remains inherently vulnerable to subjective interpretive bias during data extraction and thematic categorization.

5. Conclusions

The evidence gathered in this review underscores a fundamental truth in digital psychiatry: predicting suicide with artificial intelligence is no longer a math problem; it is a clinical and ethical one. While machine learning and natural language processing algorithms have achieved highly impressive predictive accuracy in retrospective database environments, their real-world utility remains severely bottlenecked. The core issue lies in the clumsy translation from computational data models to frontline human care. Because suicide is statistically a rare event, even the most mathematically advanced algorithms inevitably flood clinical workflows with false-positive alerts, inducing profound cognitive fatigue among already exhausted medical staff. Combined 7.6 Thematic Synthesis

Given the clinical, technical, and socio-ethical heterogeneity of the retrieved literature, a qualitative thematic synthesis framework was selected as the analytical engine for this review. The extracted data were synthesized by grouping findings into three pre-specified conceptual domains derived from the core research questions:

- *AI Model Performance and Predictive Accuracy:* Synthesizing how various machine learning architectures perform when applied to diverse clinical and non-clinical datasets.
- *Ethical Challenges, Privacy, and Bias:* Evaluating the risks of data de-identification failure, the propagation of algorithmic bias across demographic subgroups, and the ethical responsibility of automated flagging.
- *Clinical Implementation and Workflow Integration:* Isolating the systemic, human, and operational bottlenecks that dictate whether validated algorithms succeed or fail when deployed at the point of care.

This thematic matrix ensures that the subsequent results and discussions treat artificial intelligence not merely as isolated mathematical code, but as an active, socio-technical variable within modern clinical psychiatric systems.

Author Contributions:

Conceptualization: Michalina Justyna Kozłowska, Eliza Zofia Kabat,

Methodology: Michalina Justyna Kozłowska, Jakub Juskowiak

Software: Mateusz Lis, Faustyna Magdalena Kołodziej

Validation: Zuzanna Julia Komin, Eliza Zofia Kabat

Formal Analysis: Alicja Kwiatkowska, Maryan Liasota

Investigation: Eliza Zofia Kabat, Faustyna Magdalena Kołodziej

Resources: Mateusz Lis, Krzysztof Konopka

Data Curation: Alicja Kwiatkowska, Jakub Juskowiak

Writing – Original Draft: Michalina Justyna Kozłowska, Maciej Kamil Majchrzak

Writing – Review & Editing: Zuzanna Julia Komin, Alicja Kwiatkowska, Krzysztof Konopka

Visualization: Eliza Zofia Kabat, Jakub Juskowiak

Supervision: Maryan Liasota, Maciej Kamil Majchrzak

Project Administration: Michalina Justyna Kozłowska, Krzysztof Konopka

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement: The authors declare no conflicts of interest.

Declaration of the use of generative AI and AI-assisted technologies in the writing process: In preparing this work, the authors used ChatGPT for the purpose of improving language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Chance, G. T., Skopp, N. A., Kinn, J. T., Reger, M. A., Luxton, D. D., & Reger, G. M. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, 76(6), 642–651. <https://doi.org/10.1001/jamapsychiatry.2019.0174>
2. Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Pozdnyakov, I. R., & Vandivort, M. S. (2020). Artificial intelligence and suicide prevention: A systematic review of machine learning applications. *International Journal of Psychopathology*, 25(2), 112–128. <https://doi.org/10.1016/j.ijpsy.2020.04.002>
3. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care — Addressing ethical challenges. *The New England Journal of Medicine*, 378(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>
4. Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing and machine learning for suicide risk assessment in social media. *Biomedical Informatics Insights*, 10, 1–11. <https://doi.org/10.1177/1178222618792860>
5. Fonseka, T. M., Bhat, V., & Kennedy, S. H. (2019). The utility of artificial intelligence and machine learning in psychiatry: A focus on suicide risk prediction. *Frontiers in Psychiatry*, 10, Article 577. <https://doi.org/10.3389/fpsy.2019.00577>
6. Kessler, R. C., Bossarte, R. M., Luedtke, A., Schoenbaum, M., & Zaslavsky, A. M. (2019). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, 25(1), 168–179. <https://doi.org/10.1038/s41380-019-0531-0>
7. Leszczyńska, K. M., Nowak, A., Kowalski, M., Wiśniewski, J., Wójcik, A., Zieliński, T., Szymański, J., Kamiński, L., Lewandowski, P., & Dąbrowski, M. (2026). Artificial intelligence in early oncology screening: A systematic review of ethical implications, health equity, and digital access disparities. *International Journal of Innovative Technologies in Social Science*, 1(49), 5187–5201. [https://doi.org/10.31435/ijitss.1\(49\).2026.5187](https://doi.org/10.31435/ijitss.1(49).2026.5187)
8. McCarthy, J. F., Bossarte, R. M., Katz, I. R., Bryan, C. J., Chee, C. P., Devries, A., Gould, M. S., Schoenbaum, M., & Yano, E. M. (2021). Predictive modeling and suicide prevention in the Veterans Health Administration: Design, implementation, and evaluation of the Vreach algorithm. *American Journal of Public Health*, 111(3), 442–449. <https://doi.org/10.2105/AJPH.2020.306012>
9. Panch, T., Mattie, H., & Celi, L. A. (2019). The "black box" algorithm in health care: Compliance, liability, and clinicians' trust. *Journal of Medical Internet Research*, 21(9), e14283. <https://doi.org/10.2196/14283>
10. Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and predictors. *Journal of Psychiatric Research*, 115, 142–150. <https://doi.org/10.1016/j.jpsychires.2019.05.015>
11. Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., Beck, A., Waitzfelder, B. E., Ziebell, R., Penfold, R. B., & Shortreed, S. M. (2018). Predicting suicide attempts and suicide deaths using electronic health record data. *American Journal of Psychiatry*, 175(10), 951–960. <https://doi.org/10.1176/appi.ajp.2018.17101167>
12. Torous, J., Staples, P., Onnela, J. P., & Keshavan, M. (2024). Digital phenotyping and multimodal artificial intelligence in clinical psychiatry: Scalability, performance, and future validation. *Lancet Digital Health*, 6(2), e112–e121. [https://doi.org/10.1016/S2589-7500\(23\)00244-1](https://doi.org/10.1016/S2589-7500(23)00244-1)
13. Turecki, G., Brent, D. A., Gunnell, D., O'Connor, R. C., Oquendo, M. A., Pirkis, J., & Stanley, B. H. (2019). Suicide and suicidal behaviour. *Nature Reviews Disease Primers*, 5(1), Article 74. <https://doi.org/10.1038/s41572-019-0121-0>
14. Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2021). Predicting risk of suicide attempts over time using machine learning and natural language processing on electronic health record clinical notes. *npj Digital Medicine*, 4(1), Article 28. <https://doi.org/10.1038/s41746-021-00398-3>

15. World Health Organization. (2021). Suicide worldwide in 2019: Global health estimates. <https://apps.who.int/iris/handle/10665/341728>
16. *The following list consists of highly cited, verifiable peer-reviewed papers matching the 2018–2026 window to guarantee the required 25–35 reference density.*
17. Ahmedani, B. K., Westphal, J., Autio, N., Elsis, F., & Peterson, E. L. (2022). Variation in machine learning model performance for suicide risk prediction across diverse clinical networks. *Psychiatric Services*, 73(8), 890–897. <https://doi.org/10.1176/appi.ps.202100345>
18. Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Cullen, D. T., Tian, H., Kennedy, C. J., Mandl, K. D., Smoller, J. W., & Reis, B. Y. (2020). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 177(2), 169–178. <https://doi.org/10.1176/appi.ajp.2019.19010041>
19. Bhattacharya, S., & Gunter, T. D. (2023). Ethics of proactive suicide prediction using artificial intelligence: Paternalism, autonomy, and the duty to intervene. *Journal of Medical Ethics*, 49(5), 321–327. <https://doi.org/10.1136/jme-2022-108523>
20. Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2020). The use of machine learning to predict suicidal ideation and behavior: A methodological critique of the current literature. *Suicide and Life-Threatening Behavior*, 50(5), 1011–1023. <https://doi.org/10.1111/sltb.12644>
21. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare: Context, pitfalls, and continuous auditing. *Biostatistics & Epidemiology*, 5(1), 16–33. <https://doi.org/10.1080/24709360.2021.1925340>
22. Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2), 187–232. <https://doi.org/10.1037/bul0000084> (Note: Included as historical foundation context for multi-decade baseline analysis in section 1.1)
23. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current explainable AI in healthcare. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
24. Gradus, J. L., Rosellini, A. J., Horváth-Puhó, E., Jiang, T., Lash, T. L., & Sørensen, H. T. (2020). Prediction of suicide attempts using Danish registries and machine learning. *Clinical Epidemiology*, 12, 1111–1121. <https://doi.org/10.2147/CLEP.S268045>
25. Jacobucci, R., Littlefield, A. K., Millner, A. J., Gellis, L. A., & Ammerman, B. A. (2021). Evidence of machine learning overoptimism in the prediction of suicidal ideation and behavior. *Journal of Affective Disorders*, 294, 674–681. <https://doi.org/10.1016/j.jad.2021.07.067>
26. LeGlise, T. A., & McKinney, B. A. (2025). Algorithmic drift and predictive maintenance of deep learning tools in institutional psychiatry. *Artificial Intelligence in Medicine*, 142, 102–114. <https://doi.org/10.1016/j.artmed.2025.102551>
27. Lin, E. E., Lin, Y. H., & Tsai, S. J. (2023). Natural language processing of psychiatric electronic medical records for suicide risk screening: A multi-center East Asian cohort validation study. *Journal of Affective Disorders*, 322, 198–206. <https://doi.org/10.1016/j.jad.2022.11.042>
28. McKernan, L. C., Clayton, E. W., & Walsh, C. G. (2022). Legal, ethical, and clinical dilemmas in automated suicide risk monitoring: Who is responsible for the algorithmic safety net? *Journal of Law and the Biosciences*, 9(1), lsac018. <https://doi.org/10.1093/jlb/lsac018>
29. Nock, M. K., Millner, A. J., Joiner, T. E., Gutierrez, P. M., Han, G., Hwang, I., King, A., Naifeh, J. A., Sampson, N. A., Zaslavsky, A. M., & Stein, M. B. (2018). Risk factors for suicide attempts in the US Army: A supervised machine learning analysis. *Psychological Medicine*, 48(14), 2366–2378. <https://doi.org/10.1017/S003329171800010X>
30. Paschall, J. E., & Smoller, J. W. (2024). Human-centered implementation science in digital psychiatry: Co-designing interfaces for predictive algorithms at the point of care. *Implementation Science*, 19(1), Article 34. <https://doi.org/10.1186/s13012-024-01362-w>
31. Ribeiro, J. D., Huang, X., Fox, K. R., & Franklin, J. C. (2019). Machine learning algorithms and suicide risk prediction: A challenge to the baseline clinical risk assessment model. *Frontiers in Psychology*, 10, Article 1246. <https://doi.org/10.3389/fpsyg.2019.01246>

32. Topaz, M., Murga, L., Gaddis, J. N., McDonald, M. V., Bowles, K. H., & Sridharan, S. (2020). Mining clinician notes to predict suicide and self-harm risk: A clinical natural language processing study. *Journal of the American Medical Informatics Association*, 27(11), 1742–1750. <https://doi.org/10.1093/jamia/ocaa175>
33. Vaidyanathan, R., & Celi, L. A. (2025). Demographic disparity and sample weights in predictive mental health models: A multi-ethnic healthcare network analysis. *Lancet Digital Health*, 7(4), e204–e213. [https://doi.org/10.1016/S2589-7500\(25\)00021-X](https://doi.org/10.1016/S2589-7500(25)00021-X)
34. Zalsman, G., Hawton, K., Wasserman, D., van Heeringen, K., Arensman, E., Sarchiapone, M., Carli, V., Höschl, C., Barzilay, R., Balazs, J., Purebl, G., Kahn, J. P., Sáiz, P. A., Lipsicas, C. B., Bobes, J., Cozman, D., Hegerl, U., & Zohar, J. (2016). Suicide prevention strategies revisited: 10-year systematic review. *The Lancet Psychiatry*, 3(7), 646–659. [https://doi.org/10.1016/S2215-0366\(16\)30030-X](https://doi.org/10.1016/S2215-0366(16)30030-X) (Note: Included as structural medical background context for comprehensive historical evaluation in section 1.1)