



# 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 PSYCHIATRIC DIAGNOSIS AND  
CLINICAL DECISION-MAKING: CURRENT EVIDENCE, CLINICAL  
APPLICATIONS, CHALLENGES AND FUTURE DIRECTIONS

---

## DOI

[https://doi.org/10.31435/ijitss.3\(51\).2026.6517](https://doi.org/10.31435/ijitss.3(51).2026.6517)

---

## RECEIVED

02 July 2026

---

## ACCEPTED

17 September 2026

---

## PUBLISHED

28 September 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 PSYCHIATRIC DIAGNOSIS AND CLINICAL DECISION-MAKING: CURRENT EVIDENCE, CLINICAL APPLICATIONS, CHALLENGES AND FUTURE DIRECTIONS

**Aleksandra Palka** (Corresponding Author, Email: [ola.palka001@gmail.com](mailto:ola.palka001@gmail.com))

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0005-5346-3871

**Wojciech Kopyciński**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0005-2538-0534

**Norbert Krawczyk**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0000-7304-2226

**Piotr Wilczkowski**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0004-8118-1188

**Natalia Krzysztofek**

Collegium Medicum, Jan Kochanowski University, Kielce, Poland

ORCID ID: 0009-0001-6545-3962

**Monika Kamela**

Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland

ORCID ID: 0009-0000-2503-2058

**Katarzyna Boszczyk**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0005-9974-3668

**Małgorzata Grzybowska**

Medical University of Lublin, Lublin, Poland

ORCID ID: 0009-0002-6611-3983

**Adrianna Baran**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0002-3986-2559

**Olaf Krawczyk**

Medical University of Łódź, Łódź, Poland

ORCID ID: 0009-0009-1493-5652

**ABSTRACT**

Artificial intelligence (AI) is increasingly investigated as a means of supporting psychiatric diagnosis and clinical decision-making. This narrative review critically evaluates its current clinical applications, limitations, ethical challenges, and readiness for implementation in psychiatric practice. A literature search of PubMed, Scopus, Web of Science, and Google Scholar was conducted for English-language publications issued between January 2020 and July 2026, supplemented by earlier landmark contributions. Following duplicate removal and eligibility screening, 111 publications were included in a clinically oriented narrative synthesis. Machine learning, deep learning, natural language processing, digital biomarkers, and large language models have demonstrated potential applications in diagnostic classification, symptom monitoring, treatment-response prediction, relapse and suicide-risk assessment, clinical documentation, and precision psychiatry. Promising performance in retrospective or internally validated datasets has not yet been matched by robust evidence of improved clinical decisions or patient outcomes. Major limitations include small and unrepresentative samples, methodological heterogeneity, insufficient external and prospective validation, inadequate calibration, limited interpretability, inconsistent safety assessment, and vulnerability to bias and automation-related errors. These concerns are particularly important in high-risk applications involving suicide prediction, treatment allocation, and direct patient interaction. AI should augment rather than replace comprehensive clinical assessment and professional judgment. Translation into routine practice will require prospective multicenter studies, representative longitudinal datasets, transparent reporting, subgroup-specific fairness evaluation, meaningful human oversight, stakeholder co-design, and continuous post-deployment monitoring. The value of psychiatric AI should ultimately be determined by whether it safely and equitably improves clinical decisions and patient-relevant outcomes.

---

**KEYWORDS**

Artificial Intelligence, Psychiatry, Clinical Decision Support, Machine Learning, Digital Biomarkers, Large Language Models

---

**CITATION**

Palka, A., Kopyciński, W., Krawczyk, N., Wilczkowski, P., Krzysztofek, N., Kamela, M., Boszczyk, K., Grzybowska, M., Baran, A., & Krawczyk, O. (2026). Artificial intelligence in psychiatric diagnosis and clinical decision-making: Current evidence, clinical applications, challenges and future directions. *International Journal of Innovative Technologies in Social Science*, 3(51). [https://doi.org/10.31435/ijitss.3\(51\).2026.6517](https://doi.org/10.31435/ijitss.3(51).2026.6517)

---

**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

Psychiatric diagnosis and treatment remain primarily based on clinical interviews, behavioral observation and symptom-based classification. Successive editions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD) have improved diagnostic reliability and facilitated communication across clinical and research settings. Their categories do not fully capture the dimensional, overlapping and dynamic nature of psychopathology. Contemporary psychiatry faces a persistent tension between the practical utility of established diagnostic systems and the need for more individualized approaches to assessment and treatment, suggesting that progress is more likely to arise through the gradual integration of complementary methods than through the replacement of existing clinical frameworks by a single new paradigm (Stein et al., 2022). This challenge is reinforced by substantial heterogeneity within psychiatric diagnoses: patients assigned to the same disorder may differ markedly in symptom profiles, underlying biological mechanisms, longitudinal trajectories and treatment response. Extensive research across genetic, transcriptomic, proteomic, metabolomic and epigenetic domains has not yet found a biomarker that has achieved sufficiently robust and generalizable performance for routine diagnostic or therapeutic use across major psychiatric disorders. The development of clinically meaningful markers may depend on combining multiple sources of information to stratify heterogeneous populations and identify patterns associated with prognosis or treatment response, reducing reliance on trial-and-error management (García-Gutiérrez et al., 2020).

The analysis of such complex and multidimensional information is facilitated by artificial intelligence (AI). Machine learning, deep learning and natural language processing can integrate clinical, sociodemographic, neurobiological and linguistic data, while digital biomarkers derived from smartphones and wearable devices enable behavioral and physiological changes to be captured longitudinally between clinical encounters. Across the mental healthcare pathway, these approaches have been investigated for diagnostic classification, early detection, symptom monitoring, prognostic modeling, treatment-response prediction and digitally delivered interventions. Their potential clinical role ranges from cross-sectional assessment to the evaluation of illness trajectories (Cruz-Gonzalez et al., 2025). By identifying interactions that may be difficult to recognize through conventional analysis, predictive models could estimate individualized risks, distinguish clinically relevant subgroups and support decisions regarding monitoring intensity or preventive intervention (Lieslehto et al., 2025; Perry et al., 2026). Recent models have demonstrated that routinely available clinical data can be used to estimate cardiometabolic outcomes in young people with psychosis-spectrum disorders. This illustrates how psychiatric decision support may also address the substantial physical-health burden associated with severe mental illness. Although the model was developed using primarily penalized logistic regression, it provides an important benchmark for AI-supported tools (Perry et al., 2026). External validation of a mortality-risk model in first-episode bipolar disorder suggests that a machine-learning-based algorithm developed in one diagnostic population may retain prognostic value in another. These findings should be interpreted in terms of the model's calibration, decision-curve performance and transportability rather than discrimination alone (Lieslehto et al., 2025). The expansion of psychiatric AI has occurred more rapidly than the accumulation of evidence supporting its clinical utility. In antidepressant selection, for example, many models predict general treatment outcomes rather than the differential benefit of competing medications. Lack of evaluation using data from independent clinical cohorts, inadequate calibration and insufficient modeling of heterogeneous treatment effects restrict their ability to support the choices clinicians encounter in practice (He et al., 2026). The distinction between technical performance and clinical usefulness is even more consequential in suicide-risk assessment. Results in selected datasets are promising but methodological heterogeneity and the scarcity of externally validated models prevent current algorithms from providing sufficiently reliable risk stratification for autonomous clinical use (Pigoni et al., 2024). High accuracy or area under the curve within a development sample should not be interpreted as evidence that a model improves clinical decisions, patient outcomes or resource allocation (He et al., 2026; Pigoni et al., 2024).

Clinical implementation also raises ethical and organizational challenges extending beyond model performance. Psychiatric AI systems may reproduce demographic and socioeconomic biases embedded in their training data. Responsible integration into clinical practice consequently requires transparent development, representative datasets, continuous evaluation, clearly defined responsibility and meaningful human supervision throughout the AI lifecycle (Goktas & Grzybowski, 2025). These safeguards are particularly important in mental healthcare, where empathy, therapeutic alliance and contextual interpretation are integral to assessment and treatment. AI may enhance access, personalization and continuity of care, but should complement rather than displace human interaction and clinician-led judgment (Babu & Joseph, 2024).

Existing literature has extensively described potential applications of AI in psychiatry, but less attention has been paid to whether reported predictive performance translates into clinically actionable and patient-relevant benefits. The present review critically evaluates current evidence regarding the application of AI in psychiatric diagnosis and clinical decision-making. It examines the methodological foundations and clinical applications of machine learning, deep learning, natural language processing, digital biomarkers and large language models. Particular emphasis is placed on their use in diagnosing major psychiatric disorders, predicting treatment response and relapse, assessing suicide risk and supporting precision psychiatry. The review further considers methodological limitations, ethical challenges and the requirements for translating promising computational performance into safe, equitable and clinically meaningful psychiatric practice.

## 2. Methodology

### 2.1. Review Design

This article was designed as a clinically oriented narrative review informed by a literature search. A narrative approach was chosen given the substantial heterogeneity of AI methodologies, data modalities, and clinical outcomes across psychiatric disorders, which did not support meaningful quantitative synthesis and favored a critical, clinically oriented interpretation of the available evidence. Its purpose was to critically synthesize current evidence regarding the applications of artificial intelligence (AI) in psychiatric diagnosis and clinical decision-making, with particular emphasis on clinical utility, methodological limitations, patient safety, ethical challenges, and readiness for implementation in routine psychiatric practice. The review was intended primarily for physicians. Technical and computational concepts were discussed only to the extent necessary to understand their clinical applications and limitations. The review was not registered, and no formal review protocol was prepared.

### 2.2. Literature Search

A literature search was conducted in PubMed, Scopus, Web of Science and Google Scholar. The final search was performed on July 20, 2026. The search strategy combined terms related to AI and its principal applications in psychiatry. The core search terms included "artificial intelligence AND psychiatry," "machine learning AND psychiatry," "clinical decision support AND psychiatry," and "artificial intelligence AND depression." These were supplemented with analogous combinations referring to psychiatric diagnosis, treatment-response prediction, relapse prediction, suicide and self-harm risk assessment, prognostic modeling, risk stratification, early intervention, precision psychiatry, natural language processing, large language models, digital biomarkers, digital phenotyping and specific psychiatric disorders, including schizophrenia, bipolar disorder, anxiety disorders, attention-deficit/hyperactivity disorder, and autism spectrum disorder. The search was restricted to English-language publications issued between January 2020 and July 2026. Earlier publications were considered only when they represented landmark contributions or were necessary to explain the development of clinically relevant concepts. Reference lists of selected publications were also reviewed to identify potentially relevant sources that may not have been retrieved through the primary database searches. Duplicate records across databases were identified and removed manually.

### 2.3. Eligibility Criteria

Publications were considered eligible if they examined the application of AI to the screening, detection, classification, or diagnosis of psychiatric disorders, or to clinical decision-making in psychiatric care. Relevant applications included treatment selection, prediction of treatment response or relapse, prognostic modeling, early intervention, suicide and self-harm risk assessment, and patient risk stratification. Studies addressing precision psychiatry, digital biomarkers, digital phenotyping, clinically relevant multimodal data analysis, or the use of natural language processing and large language models in psychiatric documentation, patient communication, information retrieval, and decision support were also eligible. In addition, publications concerning the clinical implementation, safety, ethical implications, explainability, privacy, accountability, or regulation of AI-based tools in psychiatric care were included. Eligible sources included original clinical studies, systematic reviews, meta-analyses, clinically oriented narrative reviews, consensus or expert papers, and reports published by recognized international organizations. Priority was given to publications that provided information relevant to physicians, patient management, or healthcare implementation. Publications were excluded if they were not written in English, fell outside the specified publication period without providing essential historical or conceptual context, or lacked a clear relationship to psychiatric diagnosis or clinical decision-making. Studies focused predominantly on algorithm architecture, software development, mathematical optimization, or engineering performance without an identifiable clinical application were not prioritized. Conference abstracts and proceedings papers were eligible when they provided sufficient methodological and clinically relevant information to support the narrative synthesis. Records lacking adequate information for interpretation were excluded. Duplicate records were removed before eligibility screening.

## 2.4. Study Selection

The database searches initially identified 202 records. After the removal of 38 duplicate records, 164 unique records were screened for relevance and eligibility. Titles, abstracts, and, where necessary, full texts were assessed in relation to the review aim and eligibility criteria. Fifty-three records were excluded because they did not meet the eligibility criteria or did not provide sufficient clinically relevant information. Ultimately, 111 publications were included in the narrative synthesis.

## 2.5. Data Extraction and Narrative Synthesis

Data extraction was performed by members of the review team using a predefined set of clinically relevant variables, and uncertain interpretations were discussed until consensus was reached. Information relevant to the objectives of the review was extracted from the included publications and organized according to the principal clinical domains: technological foundations relevant to psychiatric practice, psychiatric diagnosis and screening, treatment-response prediction, relapse and suicide-risk prediction, prognostic modeling and risk stratification, precision psychiatry, clinical decision-support systems, large language models, clinical implementation, and ethical and regulatory considerations.

For original studies, particular attention was paid to the study population, psychiatric condition, source and type of input data, intended clinical task, comparator, validation approach, reported performance, and potential clinical application. Where reported by the authors, internal and external validation were distinguished, and diagnostic or predictive performance was interpreted in the context of study design, sample characteristics, and clinical setting. Findings from original studies were considered separately from conclusions derived from systematic reviews, meta-analyses, narrative reviews, expert commentaries, and organizational guidance.

Given the narrative design and the substantial heterogeneity of the included publications in terms of populations, data modalities, AI methods, clinical outcomes, and validation procedures, no quantitative meta-analysis was undertaken. No formal risk-of-bias or certainty-of-evidence assessment was performed. Instead, the evidence was synthesized descriptively and critically, with emphasis on external and prospective validation, generalizability, comparison with clinician assessment or established standards of care, interpretability, potential sources of bias, patient safety, and evidence of benefit in real-world clinical practice. High performance reported in retrospective or internally validated datasets was not considered sufficient evidence of clinical utility in the absence of external validation, prospective evaluation, or demonstrated improvement in patient-relevant outcomes.

## 2.6. Methodological Limitations

Several methodological limitations should be acknowledged. Google Scholar does not offer the same level of search reproducibility as bibliographic databases, which may have affected the consistency of record retrieval. The review was restricted to English-language publications, introducing a potential language bias. Finally, because this was a narrative rather than a systematic review, no formal risk-of-bias assessment was conducted, and the findings should be interpreted as a clinically focused critical synthesis rather than an exhaustive or quantitatively pooled evaluation of all available evidence.

### **3. Results and Discussion**

#### **3.1. Foundations of artificial intelligence in psychiatry**

##### **3.1.1. Machine learning and deep learning in psychiatry**

###### **3.1.1.1. Machine learning**

Machine learning (ML) comprises computational methods that identify patterns in multidimensional data and use them to generate predictions without requiring all relevant relationships to be specified in advance. In psychiatry, ML can integrate clinical symptoms, sociodemographic characteristics, electronic health records, neuroimaging and digitally captured behavior. It supports diagnostic classification (Iyortsuun et al., 2023). Supervised algorithms learn relationships between input variables and predefined outcomes, such as diagnosis, relapse or therapeutic response, whereas unsupervised methods identify latent structures without relying on existing clinical labels. The latter may be particularly relevant to psychiatry because data-driven clustering can reveal clinical or biological subgroups that remain obscured within heterogeneous DSM- or ICD-defined disorders (Squires et al., 2023). ML extends conventional analyses focused predominantly on average group effects by aiming to generate generalizable predictions for individual patients. It may support decisions concerning diagnosis, prognosis and treatment selection (Dwyer et al., 2018).

###### **3.1.1.2. Deep learning**

Deep learning (DL), a specialized subset of ML, uses multilayer artificial neural networks to learn hierarchical representations of complex data. Whereas conventional ML commonly depends on predefined or manually engineered variables, DL can extract informative features directly from relatively unprocessed neuroimaging, speech, clinical text, facial expressions or physiological signals. This capacity has supported its investigation in psychiatric diagnosis, symptom characterization, identification of patient subgroups and prediction of illness trajectories or treatment response (Koppe et al., 2021). The greater representational capacity of DL is accompanied by greater dependence on sample size and computational resources (Kufel et al., 2023). Separating data into training, validation and independent test sets is essential to determine whether a model has learned reproducible clinical signals rather than memorized sample-specific features, while assessing performance on data from different sites or populations ("out-of-domain" prediction) is required to establish transportability across institutions and patient populations. These requirements are particularly important in psychiatry, where datasets are frequently small relative to their dimensionality and diagnostic labels may themselves be heterogeneous or imprecise (Koppe et al., 2021).

The distinction between conventional ML and DL should not be interpreted as a simple hierarchy of clinical performance. Both approaches are data-intensive, but DL is particularly suited to large, high-dimensional or minimally processed inputs such as images, speech or text — a capacity that often comes at the cost of insight into which data features are driving its outputs. For this reason, clinical value ultimately depends on rigorous validation, including randomized controlled trials and comparison against existing clinical practice, rather than on the sophistication of the algorithm alone (Chakraborty et al., 2024).

###### **3.1.2. Natural language processing**

Natural language processing (NLP) is a branch of artificial intelligence that enables computational analysis of unstructured language to extract information, identify emotional or semantic patterns and infer aspects of an individual's mental state. In psychiatry, NLP has been applied to language-based data, such as social media posts, electronic health records, clinical interview transcripts, psychotherapy sessions and patient diary entries. It supports tasks ranging from symptom detection and relapse prediction to estimation of therapeutic engagement and treatment response. Although these applications demonstrate that clinically relevant signals may be derived from routinely generated text, most studies remain proof-of-concept or feasibility investigations and have not yet established sufficient clinical validity to directly guide patient care decisions (Linardon, 2025). The methodological foundations of psychiatric NLP have evolved from conventional machine-learning classifiers and deep-learning architectures such as convolutional neural networks and long short-term memory networks towards transformer-based architectures. These architectures include BERT and RoBERTa, which can represent bidirectional contextual relationships and detect linguistically subtle indicators of depression. Models adapted to mental health corpora, including MentalBERT and DepRoBERTa, have achieved favorable results in selected datasets. These results indicate that domain-specific pretraining may improve recognition of psychiatric language patterns. By contrast, general-purpose

models such as GPT-3.5, Bard and Alpaca have shown inconsistent performance and require further task-specific refinement. The generalizability and responsible clinical application of NLP-based psychiatric prediction remain restricted by insufficient dataset diversity, class imbalance, language and platform dependence, insufficient interpretability, and unresolved concerns regarding privacy and consent (Tahir et al., 2025). Text-based models have demonstrated a substantial pooled association with standardized depression labels. It should be noted that high heterogeneity and a wide prediction interval indicate that their performance may vary considerably across populations and settings. Embedding-based representations and deep-learning architectures outperformed traditional linguistic features and shallow models, while models trained against clinician-established diagnoses performed better than those using Patient Health Questionnaire-9 (PHQ-9) self-report labels. This evidence demonstrates that the validity of the reference standard materially influences apparent model performance (S. Zhang et al., 2026). A broader systematic review likewise found a pooled accuracy of 0.80 and an AUC of 0.79 for depression detection across 43 independent datasets, although extreme heterogeneity indicates that these estimates should not be interpreted as consistent performance across settings. Structured clinical interviews produced higher accuracy than open-ended responses or patient–therapist interactions, while text source was the only significant predictor in meta-regression. These findings imply that the context and method of language collection may be more influential than model complexity alone (H. Fisher et al., 2026).

Naturalistic speech analysis may also enable dimensional monitoring of psychopathology rather than merely distinguishing patients from healthy controls. In schizophrenia spectrum disorders, AI-derived acoustic, syntactic, semantic and sentiment features were associated with the severity of positive and negative symptoms across independent Dutch and US cohorts. Reduced verbal output and flatter acoustic profiles characterized negative symptoms, whereas longer utterances and altered discourse organization were associated with positive symptoms. The ability to estimate hallucinations, suspiciousness and blunted affect within the agreement margin accepted during clinical-rater training suggests that brief speech samples may eventually support interpretable, low-burden monitoring of symptom dynamics (Ciampelli et al., 2026).

### **3.1.3. Digital biomarkers**

Digital biomarkers can be derived from physiological and behavioral data, including electroencephalography, speech, facial expression, eye movement and gait. They may provide objective information that complements conventional depression screening. Multimodal approaches appear particularly promising because they integrate several manifestations of the disorder and have achieved better pooled discrimination than individual modalities, while deep learning has produced modest improvements over traditional machine-learning methods (L. Wang et al., 2025).

Speech is an especially accessible digital biomarker because it can be collected repeatedly and remotely using widely available devices such as smartphones. Meta-analytic evidence indicates potentially useful performance in psychiatric disorders, but the results vary according to the recording device, language, speech task, extracted features and analytical algorithm. Speech-derived measures should be regarded as scalable candidates for screening and monitoring rather than established diagnostic tests, pending validation in larger, prospective and linguistically diverse cohorts (Yang et al., 2026). Smartphones, wearables and online platforms extend digital phenotyping beyond speech by enabling longitudinal measurement of mobility, geolocation, sleep, device interaction and other behavioral or physiological signals associated with mood disorders. Personalized models may be more informative than universal algorithms because they detect deviations from an individual baseline. These technologies are better positioned as adjunctive monitoring and early-warning tools than replacements for clinical assessment (Garzón-Partida et al., 2025). Voice monitoring may be particularly relevant to bipolar disorder because changes in mood state can influence speech rate, prosody and other acoustic characteristics. Available studies indicate better discrimination of manic than depressive states. It is noteworthy that high accuracy has primarily been obtained under controlled conditions, whereas performance is less consistent in naturalistic settings. Along with other limitations, this means that vocal biomarkers are not yet suitable for standalone diagnosis or prediction of bipolar episodes (Briganti & Lechien, 2025). Evidence from neurodegenerative disorders further illustrates the general methodological value of multimodal integration. Combining finger-tapping, gait and speech measures improved discrimination of Parkinson's disease compared with the corresponding unimodal models. These findings suggest that complementary digital signals may capture heterogeneous clinical manifestations more effectively than any single modality. The limited number of multimodal studies and continued reliance on cohorts with pre-established diagnoses demonstrate the need for blinded testing and prospective validation before such biomarkers can be regarded as independent diagnostic tools (Song et al., 2025).

### **3.2. AI in clinical decision-making**

Artificial intelligence may support psychiatric decision-making by integrating questionnaires, electronic health records, neuroimaging, and digital phenotyping data to predict symptom trajectories, relapse risk, and treatment response. Current evidence supports AI as a complement to clinical expertise rather than an autonomous substitute for clinician-led assessment and treatment planning (Jalali et al., 2025).

#### **3.2.1. Treatment prediction**

Machine learning may support treatment selection by estimating individual benefit rather than relying exclusively on average intervention effects. In an individual-participant meta-analysis of cognitive training for anxiety and depression, baseline symptom severity and training dose were the most influential predictors of improvement. At the same time, active training, particularly attention and interpretation bias modification, was predicted to be the optimal condition for most participants. These findings illustrate how interpretable models combining random forests with SHAP analysis could identify patient- and intervention-level characteristics relevant to personalized treatment assignment (Richter et al., 2025). In the studies included in the meta-analysis, neuroimaging and biological predictors generally outperformed behavioral or combined feature sets. Considering their cost, methodological heterogeneity and limited external generalizability, particularly in depression, standardized imbalance correction, rigorous validation and carefully designed multimodal integration are required before these models can guide real-world treatment selection (Curtiss & DiPietro, 2025).

The predictive value of different data modalities has also been investigated in schizophrenia, where EEG-based models demonstrated greater pooled sensitivity and specificity for treatment response than models based on clinical characteristics or neuroimaging. EEG may be particularly informative because it captures fine-grained temporal patterns of neural activity and is relatively inexpensive and noninvasive. These estimates were derived from only three studies with 144 participants and require further validation. Multimodal models performed less favorably than EEG or imaging alone, potentially because harmonizing clinical, neuroimaging and genetic information introduces preprocessing complexity and additional noise, but only two multimodal studies were available, precluding firm conclusions regarding the value of data integration (Saboori Amleshi et al., 2025).

In bipolar disorder, treatment prediction is complicated by the limited representativeness of randomized trials, which frequently exclude patients with severe illness, suicidality, rapid cycling, multimorbidity, treatment resistance, or complex polypharmacy. Early neuroimaging studies suggest that pretreatment brain morphology and changes observed during the first treatment week may predict response to quetiapine or lithium. AI-supported treatment selection should integrate algorithmic predictions with therapeutic drug monitoring, pharmacogenetic information, clinical characteristics, and shared decision-making rather than reduce treatment choice to an automated model output (Fornaro et al., 2025).

#### **3.2.2. Relapse prediction and digital phenotyping**

Machine-learning models have been developed to predict relapse, remission, and functional outcomes across psychotic-spectrum disorders, but their performance varies substantially according to the predicted outcome and the data available at the time of assessment. Electronic health records have produced particularly favorable results for relapse prediction, whereas models based on facial, physiological, or audio features have performed less consistently. These results suggest that data modality and temporal proximity may be as influential as algorithm selection (Tay et al., 2025). In a small heterogeneous cohort, an unsupervised model combining convolutional autoencoders with clustering performed best when trained on sleep-period activity and heart-rate variability signals, suggesting that sleep-related data may be particularly informative for distinguishing relapse-associated patterns (Yan et al., 2025). These findings are exploratory rather than clinically generalizable. Prognostic modeling has also been applied to substance use disorders. It is able to estimate outcomes such as treatment adherence, dropout, readmission, relapse, and substance-use severity. Current models may support risk stratification but cannot yet independently guide clinical management (de Mattos et al., 2026). Digital phenotyping extends psychiatric assessment beyond episodic consultations by translating longitudinal patterns into clinically meaningful signals. Mobility, activity, sleep, communication, and other smartphone- or wearable-derived signals are turned into indicators of changes in mental state. Digital phenotyping may complement the monitoring of severe mental illness, such as schizophrenia and bipolar disorder (Benoit et al., 2020). Its clinical value may lie not only in supporting symptom and relapse monitoring but also in helping patients recognize changes in their functioning and participate more actively in managing their mental health. This may be achieved by integrating actively reported data, such as ecological momentary

assessments and symptom questionnaires, with passively collected measures of mobility, sleep, heart rate, voice, social interaction, and language use (Oudin et al., 2023). Combining active and passive data may provide a more comprehensive representation of mental state than either source alone. In the LifeRhythm study, smartphone-derived features predicted self-reported and clinician-assessed depression, while their integration with repeated self-reports improved predictive performance. The related DepWatch framework was designed to combine longitudinal sensing data with clinical information to predict antidepressant response and support treatment decisions (Kamath et al., 2022).

Evidence from adjacent psychiatric prediction tasks indicates that limited generalizability remains a broader methodological obstacle. In PTSD research, tree-based models frequently achieved favorable internal performance across different trauma populations, but their performance deteriorated during the external validation conducted in only two studies. Almost half of the included studies were judged to have a high risk of bias. Apparently strong predictive accuracy may reflect overfitting, reinforcing the need for larger and more diverse cohorts, standardized reporting, and independent validation before such models are used to guide clinical decisions (Vali et al., 2025).

### **3.2.3. Suicide risk prediction**

Machine-learning models have been developed to support suicide-risk assessment by detecting complex patterns across clinical, demographic and behavioral data (Linthicum et al., 2019). Apparently favorable global discrimination metrics have not translated into clinically adequate predictive performance. Meta-analytic evidence demonstrates only modest sensitivity and very low positive predictive values. This means that many individuals who subsequently die by suicide or repeat self-harm are classified as low risk, while most of those assigned to high-risk categories do not experience the predicted outcome. Because these models perform no better than traditional risk-assessment scales, they should not be used to determine access to treatment. Patients presenting with self-harm should instead receive needs-based psychosocial assessment, identification of modifiable risk factors and evidence-based aftercare irrespective of their predicted risk category (Spittal et al., 2025). Adolescent-specific models have shown stronger discrimination for suicide attempts than for suicidal ideation or nonsuicidal self-injury, with ensemble approaches such as XGBoost producing some of the most favorable results. Current adolescent suicide-prediction models also require prospective evaluation in independent and diverse clinical populations before they can be incorporated into clinical decision-making (L. Liu et al., 2025).

### **3.2.4. Precision psychiatry**

Precision psychiatry aims to move beyond decisions based solely on population-level diagnostic categories by identifying clinically meaningful variation between individuals who share the same diagnosis. Unsupervised methods, including K-means, hierarchical clustering and Bayesian mixture models, have been used to derive subtypes within schizophrenia, major depressive disorder and autism spectrum disorder, while frameworks such as Dimensional Neuroimaging Endophenotypes link neuroimaging-derived variation with genetic characteristics and clinical outcomes. Causal machine-learning approaches may extend this stratification by estimating heterogeneous treatment effects and identifying which patients are more likely to benefit from specific interventions (Sun et al., 2025). This effort to characterize variation within and across diagnostic categories is conceptually aligned with the Research Domain Criteria (RDoC). RDoC provides a complementary framework by organizing psychopathology around dimensional brain-behavior constructs rather than restricting research to DSM or ICD categories. Its matrix examines domains such as negative and positive valence, cognition, social processes and arousal across multiple levels of analysis, including genes, molecules, cells, neural circuits, physiology, behavior and self-report. By enabling recruitment across conventional diagnostic boundaries, RDoC is intended to facilitate the identification of dimensions and subgroups that may be obscured within heterogeneous diagnostic categories, but it remains a research framework rather than an AI method or a clinically validated replacement for existing diagnostic systems (Cuthbert, 2022; Insel et al., 2010). Within the broader dimensional framework, genomic information represents one potential source of clinically meaningful individual variation. The growing use of whole-exome and whole-genome sequencing in prenatal and neonatal care is also shifting psychiatric genetics from retrospective explanation toward the identification of neurodevelopmental and psychiatric vulnerability before symptoms become apparent. Integrating rare and common genetic variation could ultimately improve individual risk stratification and inform preventive or precision-based interventions (Vorstman et al., 2025). Translating these developments into routine decision-making will require coordinated evidence generation rather than isolated predictive models. The proposed Precision Psychiatry Roadmap emphasizes global

harmonization of research procedures, consensus regarding the predictive validity of candidate measures and gradual implementation of an evolving biology-informed diagnostic framework supported by longitudinal, multimodal resources. Because current candidate biomarkers and predictive models remain vulnerable to overfitting, fragmented datasets and limited external validity, independent validation, transparent reporting, prospective intervention trials and regulatory evaluation are necessary before computational stratification can reliably guide diagnosis or treatment selection (Kas et al., 2025).

AI-based risk stratification represents a key component of precision psychiatry by integrating heterogeneous individual-level data to identify clinically relevant differences in vulnerability before a formal diagnosis is established (Jalali et al., 2025). In ADHD, interpretable machine-learning models applied to routinely collected behavioral, questionnaire and demographic data may generate individualized risk estimates and support earlier prioritization for specialist assessment. Such predictions should complement clinical judgment and require external validation before informing routine care (Ameri et al., 2025).

### **3.3. AI in psychiatric diagnosis**

Artificial intelligence has been investigated as a means of complementing symptom-based psychiatric assessment with patterns extracted from clinical, biological and behavioral data (G. Liu et al., 2020). A meta-analysis of 14 studies reported a pooled diagnostic accuracy of 85% across heterogeneous diagnostic tasks, although differences between datasets, model architectures, and target disorders limit the extent to which this estimate can be generalized to routine practice. AI should be understood as a heterogeneous group of task-specific methods rather than a single diagnostic technology with uniform performance across psychiatry (Rony et al., 2025). The existing evidence also remains unevenly distributed across clinical applications. A systematic review of 129 publications found that AI had been studied more frequently for treatment evaluation, patient stratification, outcome prediction, and prognosis than for direct diagnosis, with conventional regression, decision-tree, and support-vector-machine methods predominating over deep learning (Tornero-Costa et al., 2023). Diagnostic models increasingly integrate clinical assessments and electronic health records with neuroimaging, electrophysiological signals, speech, language, and digital phenotyping. This multimodal approach may capture clinically relevant patterns that are not apparent from any single source, but most reported models have been developed in small or selected samples and remain vulnerable to bias, poor transferability, and limited external validation (Baydili et al., 2025). Explainable AI may facilitate clinical scrutiny by indicating which linguistic, behavioral, physiological, or imaging features contribute to a prediction. It is important to note that explainable outputs remain dependent on models that may inherit demographic or cultural biases from training data, which limits how far interpretability alone can substitute for independent clinical evaluation (Močnik et al., 2025). Translation into clinical practice remains limited. A scoping review identified only 12 empirical studies evaluating mental-health AI systems at relatively advanced levels of technological readiness, with insufficient attention to organizational integration and adaptation of clinical workflows (Auf et al., 2025). AI-informed learning healthcare systems could use continuously collected data to support assessment and periodically recalibrate models, but their predictions should remain transparent, prospectively evaluated, and subordinate to clinician-led diagnostic reasoning (Raket et al., 2020).

#### **3.3.1. Depression**

AI-supported assessment of depression has drawn on EEG, speech, linguistic patterns, and electronic records (Baydili et al., 2025). Electrophysiological models have identified patterns associated with treatment-resistant and treatment-responsive depression, while speech-based systems can detect acoustic and linguistic features related to depressive symptoms (Metin et al., 2025; Watts et al., 2022). Speech and language analysis may broaden assessment beyond information obtained during scheduled consultations. Changes in pitch, rhythm, lexical content, sentiment, and patterns of digital interaction may reveal behavioral correlates of depression. These signals are sensitive to language, culture, context, and recording conditions, so their most plausible clinical role is to prompt further assessment or support longitudinal evaluation rather than determine diagnosis independently (Olawade et al., 2024; Arias de la Torre et al., 2026).

#### **3.3.2. Schizophrenia**

Neuroimaging represents one of the most extensively investigated sources of diagnostic information in schizophrenia. In task-based fMRI studies, machine-learning models achieved pooled sensitivity and specificity of approximately 0.83, with selective-attention paradigms producing greater specificity than working-memory tasks (Idowu et al., 2025). Structural MRI models have shown similarly promising but provisional performance. A meta-analysis of 16 studies found pooled sensitivity of 0.76 and specificity of 0.78,

with deep-learning and whole-brain approaches generally outperforming support-vector-machine and grey-matter-restricted models (AUC 0.89 vs. 0.78–0.82), suggesting that whole-brain inputs may preserve diagnostically relevant information omitted by more restricted approaches (Kotochinsky et al., 2025). Longitudinal electronic health records may extend AI applications from cross-sectional classification towards earlier recognition of psychosis. A recurrent neural network trained on records from more than 100,000 individuals achieved an AUC of 0.80 in an external cohort, illustrating the potential of routinely collected clinical data for identifying patients who may require further assessment. Such systems predict risk rather than establish schizophrenia, and their outputs should therefore trigger proportionate clinical evaluation rather than function as diagnostic labels (Raket et al., 2020).

### **3.3.3. Bipolar disorder**

Machine-learning models may be particularly valuable in bipolar disorder because of the clinical difficulty of distinguishing bipolar depression from unipolar depression. Across 18 studies, models differentiated bipolar disorder from healthy controls with an AUC of 0.94 and from depression with an AUC of 0.89, while blood-based models produced the strongest subgroup performance (Pan et al., 2025). Another systematic review of 35 studies examining the use of classical AI approaches for treatment optimization across the bipolar disorder spectrum found pooled AUCs of 0.68 for predicting acute symptomatic response, 0.80 for long-term maintenance response, and 0.71 for relapse or readmission. The authors assessed models using PROBAST+AI which revealed a high risk of bias in most studies, while only 2 of the 35 studies evaluated their models in an independent cohort (De Francesco et al., 2026). Milic et al. reviewed the potential applications of AI in the management of bipolar disorder and concluded that AI-based tools require further clinical validation and appropriate safeguards concerning privacy, algorithmic bias, and clinician oversight before they can be routinely implemented in psychiatric care (Milic et al., 2025).

### **3.3.4. Anxiety disorders**

Available reviews provide less disorder-specific evidence for AI-supported diagnosis of anxiety disorders than for depression or psychotic disorders. Vocal, linguistic, facial, physiological, and behavioral features may identify patterns associated with anxiety, but these signals are not necessarily disorder-specific and may overlap with depression, stress, trauma-related disorders, and transient emotional states (Degante-Aguilar et al., 2025; Olawade et al., 2024). A systematic review of 119 studies reported that AI-based approaches frequently achieved high classification accuracy and could complement conventional assessment methods. Substantial variation in populations, data sources, algorithms, and validation procedures limit direct comparisons and prevents firm conclusions regarding their clinical superiority (Degante-Aguilar et al., 2025). AI-based chatbots are also being investigated as a means of providing short-term mental health support (C. Chen et al., 2025), and AI-assisted psychotherapy is being evaluated as an adjunct to pharmacotherapy for anxiety disorders (Su et al., 2022).

### **3.3.5. Attention-deficit/hyperactivity disorder**

AI-supported diagnosis of ADHD has incorporated structural and functional neuroimaging, EEG, behavioral and cognitive measures, and natural language processing. Across 54 reviewed studies, reported performance varied widely, with convolutional neural networks and support vector machines showing particular promise for imaging and electrophysiological data (Zaheer & Akhtar, 2025). The multimodal nature of ADHD may favor models that combine neurophysiological and behavioral information rather than rely on a single marker. Ensemble approaches integrating EEG and behavioral measures have reported particularly high classification performance, but such estimates remain vulnerable to overfitting when derived from small or locally collected datasets (Fallahpour et al., 2026; Zhao et al., 2025).

### **3.3.6. Autism spectrum disorder**

AI applications in autism spectrum disorder (ASD) have focused more strongly on multimodal assessment and individualized intervention than on autonomous diagnosis. Virtual-reality environments combined with eye-tracking, EEG, and adaptive features may help characterize social communication, emotional recognition, cognitive performance, and engagement. These technologies may enrich functional assessment and intervention planning (Mahmud et al., 2022; M. Zhang et al., 2022). The researchers emphasize that these systems remain at an early stage of development and require greater explainability, data integration, and clinical evaluation (Joudar et al., 2023; Solek et al., 2025). The potential role of AI in ASD extends beyond diagnostic assessment to the delivery, personalization, and monitoring of therapeutic interventions. A 12-month observational study of 43 children evaluated an AI-based platform used alongside standard therapy for

ASD. Children who consistently used the platform showed improvements in autism severity, social and developmental functioning, and receptive and expressive language scores. Because group assignment was based on adherence rather than randomization and the intervention included concurrent standard therapy and caregiver training, the observed improvements cannot be attributed conclusively to the AI platform itself (Atturu et al., 2025).

### **3.4. Large language models in psychiatry**

Large language models (LLMs) differ from earlier rule-based chatbots by generating responses from learned linguistic patterns rather than following predefined scripts or decision trees (Singh & Namin, 2025). Their capacity to sustain context-sensitive conversations has prompted growing interest in applications such as mental health assessment, psychoeducation and emotional support. Most LLM-based systems have been evaluated in simulated scenarios rather than clinical practice (Hua et al., 2025). Psychiatric evidence remains dominated by studies of ChatGPT, with comparatively limited evaluation of other generative models. Most available studies are prompt experiments, supplemented by a smaller number of surveys, pilot studies and case reports, indicating that the literature primarily demonstrates technical feasibility rather than effectiveness under routine clinical conditions (Kolding et al., 2025). Preliminary evidence suggests that LLMs could support several stages of mental health care, including personalized psychoeducation, recognition of relapse or crisis indicators, diagnostic assessment and treatment optimization based on patient-specific information (Torous et al., 2025). Evidence from other medical specialties suggests that their clinician-facing use may be particularly relevant to clinical documentation, where automated generation and organization of clinical encounter notes could substantially reduce the time devoted to documentation (Razaghi et al., 2026). Beyond generating consultation notes, LLMs can process unstructured clinical and nursing notes, medical records, patient records and patient-generated text, converting selected information into structured formats suitable for integration into electronic health records. Their performance is strongly task-dependent: results are generally more favorable for information extraction, entity recognition and standardized documentation than for diagnostic tasks requiring complex clinical judgment. Direct comparisons further show that superiority over conventional algorithms does not establish equivalence to human expertise, as clinicians outperformed LLMs more frequently than LLMs outperformed clinicians in the studies reviewed (Alkalbani et al., 2025). The distinction between technical performance and clinical utility is especially important in high-risk applications such as suicide prevention. Most studies in this area have focused on identifying or classifying suicide risk, whereas applications involving direct support, education and training remain less developed (Holmes et al., 2025). Evidence from the broader conversational-agent literature provides relevant context, although these findings should not be interpreted as direct evidence of the effectiveness of generative LLM-based systems. Conversational agents may provide scalable psychological support for adults experiencing depressive or anxiety symptoms. A meta-analysis of 18 randomized controlled trials found small short-term improvements in both outcomes, with benefits emerging after approximately eight weeks but not clearly persisting at three-month follow-up (Zhong et al., 2024).

Current psychiatric research is limited by heterogeneous study designs, insufficient methodological transparency and the predominance of early feasibility studies over prospective clinical evaluations (Avula & Amalakanti, 2024). Future research should employ clinically relevant experimental designs, improve methodological transparency, adhere to relevant reporting standards such as TRIPOD-LLM and involve patients or users in the design process. Until such evidence becomes available, LLMs should be regarded as promising but insufficiently validated tools whose clinical use requires careful evaluation of safety, ethics and context-specific performance (Kolding et al., 2025).

### **3.5. Ethical challenges**

The integration of artificial intelligence into psychiatric screening, diagnosis, treatment personalization, and longitudinal monitoring introduces ethical concerns that extend beyond technical performance. Risks related to privacy, data ownership, algorithmic bias, limited transparency, and the depersonalization of care indicate that clinical implementation should be multidisciplinary and subject to rigorous validation, with AI supporting rather than replacing clinician-led care (Wilhelmy et al., 2023). Even apparently routine applications, such as automated documentation, screening, and information synthesis, are not ethically neutral, as they may influence reflective practice, clinical reasoning, professional competence, and the allocation of care. Clinicians should critically review AI-generated outputs and retain responsibility for clinical decisions rather than treating algorithmic recommendations as an independent source of authority (C. E. Fisher, 2025).

Ethical governance should be embedded throughout the AI lifecycle rather than addressed retrospectively after deployment. The Integrated Ethical Approach for Computational Psychiatry organizes evaluation into identification, analysis, decision-making, implementation, and review, guided by beneficence and non-maleficence, autonomy and informed consent, justice and equity, privacy and confidentiality, transparency and explainability, and scientific integrity and validity (Putica et al., 2025). Ethical evaluation must also extend beyond immediate interactions between patients, clinicians, and algorithms. Healthcare AI depends on global infrastructures involving substantial energy and water consumption, mineral extraction, labor, and electronic waste. The environmental costs and health burdens may fall disproportionately on communities that may receive limited benefits from these technologies. A comprehensive assessment should consider individual-relational, organizational, societal, global, and historical consequences, linking local requirements for privacy and fairness with environmental sustainability, global health equity, and intergenerational justice (Vandemeulebroucke, 2025). Large multimodal models further illustrate why satisfactory performance at initial evaluation cannot be regarded as evidence of sustained clinical safety, as their outputs may be unpredictable, vary over time, or be applied in contexts not anticipated by developers. WHO guidance consequently emphasizes autonomy, well-being, safety, transparency, accountability, inclusiveness, and sustainability, supported by impact assessments, transparent documentation, independent audits, and continuous post-deployment monitoring. Responsibility should be distributed across developers, technology providers, healthcare institutions, clinicians, and governments, while preserving meaningful human oversight and accessible mechanisms of redress for patients harmed by AI-supported care (World Health Organization, 2025).

### **3.6. Limitations**

#### **3.6.1. Clinical effectiveness and patient safety**

Evidence for the clinical effectiveness of generative AI mental health chatbots remains preliminary and highly context-dependent. Although these systems have produced small-to-moderate improvements in mental health outcomes, effects have varied substantially across populations and settings, with the clearest benefit observed for depressive symptoms and socially oriented chatbots emphasizing interaction and emotional support. Limited trial numbers, potential publication bias, and insufficient evidence in adolescents, older adults, and people with severe psychiatric disorders support their use as adjuncts rather than established standalone treatments (Q. Zhang et al., 2025). Similar limitations emerge from the broader chatbot literature, in which pooled reductions in depression and anxiety have been modest and accompanied by substantial heterogeneity and a high risk of bias. Inconsistent monitoring and reporting of adverse events indicate that evidence of symptomatic benefit has developed faster than knowledge about the safety of these interventions (Sohn et al., 2026). Patient-facing use of agentic large language models may pose particular risks for individuals vulnerable to psychosis, as these systems may mirror, validate, or amplify delusional or grandiose content (Hudon & Stip, 2025). Current evidence does not establish whether such interactions can precipitate *de novo* psychosis in individuals without pre-existing vulnerability, so causal claims should remain cautious. Proposed safeguards include personalized instruction protocols, reflective check-ins, digital advance statements, and escalation pathways, but these measures require co-design with service users and clinicians and prospective evaluation before they can be regarded as effective clinical protections (Morrin et al., 2026).

#### **3.6.2. Generalizability and clinical validation**

High performance during model development also frequently fails to translate across institutions and patient populations. Multicenter prospective evaluation and continuous human oversight are necessary before promising accuracy estimates can justify scalable implementation (Bobkov et al., 2025). Similar weaknesses are evident in digital-biomarker research on Alzheimer's disease and mild cognitive impairment, where favorable discrimination has largely been demonstrated using retrospective or cross-sectional datasets. The scarcity of prospective studies, external validation, calibration assessment, adherence to reporting standards, and publicly available code means that strong classification performance cannot yet be interpreted as reliable diagnostic probability or real-world generalizability across populations, devices, languages, and clinical environments (Qi et al., 2025).

### **3.6.3. Reporting standards and methodological transparency**

Transparent reporting is essential for determining whether findings reflect genuine clinical effectiveness or performance obtained under curated development conditions. The SPIRIT-AI and CONSORT-AI extensions require investigators to describe the intended role of an AI system, its inputs and outputs, integration into clinical workflows, human–AI interaction, and procedures for identifying and managing errors. Adherence to these guidelines cannot compensate for small samples, inadequate validation, or the absence of meaningful clinical outcomes (McGenity & Treanor, 2021). CONSORT-AI improves the interpretability and reproducibility of trial reports by adding 14 AI-specific items, although it remains a reporting framework rather than a guarantee of representative sampling, appropriate comparators, external validity, or patient benefit (X. Liu et al., 2020). Correspondingly, the 15 additional SPIRIT-AI items promote prospective specification of the system and its clinical use, but cannot ensure methodological quality and do not adequately address continuously learning or dynamically updated models (Cruz Rivera et al., 2020). Ribeiro Junior et al. (2025) argued that CONSORT 2025 and SPIRIT 2025 do not explicitly require authors to report how AI is used in participant selection, randomization, outcome adjudication, or statistical analysis. They recommended updating the AI-specific extensions to include the reporting of training data, validation procedures, model performance, fairness, preregistration, and, where feasible, source code. The authors further emphasized that effective implementation would require active endorsement and enforcement by journals, research ethics committees, and funding bodies rather than reliance on voluntary reporting alone.

### **3.6.4. Clinical actionability, human oversight, and explainability**

A further limitation is that predictive accuracy does not necessarily translate into clinical actionability. Most psychiatric models identify associations without determining which modifiable factors cause an outcome or which intervention would benefit an individual patient (Krishnadas et al., 2025). Changes in populations, treatments, and clinical practice may progressively undermine model calibration and contribute to over- or undertreatment (Guan et al., 2026). Human interaction with AI introduces an additional source of error through automation bias. Reliance on algorithmic recommendations is influenced by different factors, such as professional expertise, trust, cognitive load, time pressure, verification demands, and the order in which advice is presented. Explanations may reinforce inappropriate trust when interpreted as endorsements of the model's output (Romeo & Conti, 2026). Explainable AI likewise cannot be assumed to resolve concerns about reliability or accountability. Methods such as SHAP and LIME may improve the accessibility of model outputs, but their explanations can be unstable, computationally burdensome or difficult to standardize (Kalasampath et al., 2025).

### **3.7. Future directions**

Future research should move beyond demonstrating that artificial intelligence can identify statistical patterns in psychiatric data and determine whether these systems produce clinically meaningful benefits. Model evaluation should include prospective and multicenter validation, clinically relevant comparators, calibration, decision-curve analysis and outcomes reflecting changes in diagnosis, treatment selection, relapse prevention or patient functioning (Sun et al., 2025). The development of a predeployment PTSD model illustrates the value of temporal and geographic validation and clinically calibrated risk estimates, while also demonstrating that predictive performance does not establish whether risk-informed interventions improve outcomes (Papini et al., 2023). Similarly, models developed for predicting depression risk in underrepresented populations should undergo external validation across healthcare systems and demographic groups, with continued assessment of calibration and subgroup performance before their use for clinical triage (Y. Wang & Pan, 2026). Future studies should prospectively define their intended clinical use, performance measures and criteria for clinically meaningful benefit. Transparent reporting and, where ethically and legally permissible, the sharing of data-processing procedures and model code will be necessary to improve reproducibility and enable independent evaluation (Z. S. Chen et al., 2026). Performance should also be reported separately across relevant demographic and clinical subgroups because acceptable average accuracy may conceal systematic errors affecting populations underrepresented in the development data (Y. Wang & Pan, 2026). Multimodal integration of clinical information, neuroimaging, speech, language, physiological measurements and digitally captured behavior may provide a more comprehensive representation of mental states. Studies should establish whether each additional modality produces clinically meaningful incremental value rather than assuming that greater data complexity necessarily improves prediction (Narimani & Naeim, 2025). Federated learning may facilitate collaboration between institutions by allowing models to be trained without transferring raw patient

data; however, its psychiatric application requires further investigation of heterogeneity between participating sites, communication and computational demands, privacy leakage and the fairness of the resulting models (Khalil et al., 2024). The goals of psychiatric modeling should also extend beyond prediction to clinical actionability. Models that estimate the probability of relapse or treatment response do not necessarily indicate which intervention would alter that outcome. Precision psychiatry will require closer integration of predictive modeling with causal inference, particularly when algorithms are expected to compare therapeutic alternatives or estimate individual treatment effects (Bernstorff & Jefsen, 2025).

Clinical implementation should be evaluated as an interaction between patients, clinicians and AI rather than as the isolated performance of an algorithm. A human-assisted framework in which models receive structured feedback and adapt to local contexts may improve reliability and trust, but it will require culturally responsive evaluation and AI education for mental health professionals (Z. S. Chen et al., 2026). The concept of the augmented clinician similarly positions AI as a means of extending pattern recognition, documentation and monitoring capacity while preserving professional responsibility, contextual reasoning and therapeutic engagement. Research should compare clinicians working with and without AI, investigate when recommendations should be presented, and measure automation bias, workload, diagnostic reasoning and effects on the therapeutic relationship (Ruan et al., 2026). Evaluation must continue after implementation because performance established during development may deteriorate as patient populations, diagnostic practices, treatments and data-collection systems change. Future clinical AI infrastructure should incorporate continuous monitoring of discrimination and calibration, detection of data and model drift, subgroup-specific performance assessment and predefined procedures for recalibration or controlled updating (El Arab et al., 2026). Lifecycle governance should additionally specify action thresholds, incident-reporting pathways, responsibility for corrective intervention and conditions under which a model should be suspended or rolled back. These operational requirements must be prospectively evaluated, as recommendations for monitoring and governance have developed more rapidly than evidence showing how they can be implemented reliably in routine healthcare (Guan et al., 2026).

Patients, clinicians, and other stakeholders should be involved in identifying clinical needs and designing AI-assisted mental healthcare. Patient and public involvement, guided by co-design and design-justice principles, may help ensure that these technologies reflect diverse user needs and reduce inequalities in participation, representation, and access (Zidaru et al., 2021). In the more specific context of an intervention combining virtual reality with generative AI, a co-creation approach has also been proposed as a means of strengthening trust and therapeutic engagement (Della Greca et al., 2024). Taken together, these findings support moving beyond one-time consultation towards more active collaboration with intended users during the development and evaluation of AI-assisted mental health interventions.

Progress in psychiatric AI will depend less on increasingly complex algorithms than on a shift towards clinically driven, reproducible and continuously evaluated research. Future studies should prioritize representative longitudinal data, independent validation, causal and comparative treatment modeling, transparent human–AI interaction, subgroup-specific fairness assessment and lifecycle-based governance. These requirements would position AI not as an autonomous replacement for psychiatric expertise, but as an accountable component of collaborative care whose value is determined by demonstrable improvements in clinical decisions, patient outcomes and equity.

## 4. Conclusions

Artificial intelligence has the potential to support psychiatric diagnosis and clinical decision-making by integrating complex clinical, biological, linguistic and behavioral data that cannot be readily synthesized through conventional assessment alone. Machine learning, deep learning, natural language processing, digital biomarkers and large language models have demonstrated promising applications in diagnostic classification, symptom monitoring, relapse and suicide-risk prediction, treatment selection and clinical documentation. However, performance remains strongly dependent on the clinical task, data source and evaluation setting, and favorable results obtained during model development should not be interpreted as evidence of clinical effectiveness.

The major barrier to implementation is not the absence of technically capable algorithms but the limited evidence that their use improves decisions or patient outcomes across diverse real-world populations. Small or unrepresentative samples, methodological heterogeneity, insufficient external and prospective validation, inadequate calibration, limited interpretability and inconsistent assessment of safety continue to restrict

generalizability. These concerns are particularly important in high-risk applications, including suicide-risk assessment, direct patient interaction and treatment allocation, where inaccurate predictions or automation bias may cause substantial harm. AI should consequently augment rather than replace comprehensive clinical assessment, contextual interpretation and professional accountability.

Translation into psychiatric practice will require prospective multicenter studies, clinically meaningful comparators, transparent reporting and continuous evaluation throughout the AI lifecycle. Greater emphasis should be placed on representative longitudinal datasets, causal and comparative treatment modeling, subgroup-specific fairness, patient and clinician co-design, and post-deployment monitoring for performance deterioration. When embedded within clearly governed human–AI collaboration, these technologies may contribute to more personalized, continuous and evidence-informed psychiatric care. Their ultimate value, however, should be judged not by algorithmic complexity or accuracy alone, but by whether they safely and equitably improve clinical decisions, therapeutic relationships and outcomes that matter to patients.

**Funding:** This research received no external funding.

**Data Availability Statement:** No new datasets were generated in this narrative review. All information analyzed in the review is available in the cited publications.

**Conflicts of Interest:** No conflicts of interest to declare.

## REFERENCES

1. Alkalbani, A. M., Alrawahi, A. S., Salah, A., Haghighi, V., Zhang, Y., Alkindi, S., & Sheng, Q. Z. (2025). A systematic review of large language models in medical specialties: Applications, challenges and future directions. *Information*, 16(6), 489. <https://doi.org/10.3390/info16060489>
2. Ameri, A., Salmanizadeh, F., Samzadeh Kermani, H., & Ghaemi, M. M. (2025). Predicting ADHD in children and adolescents with artificial intelligence: A scoping review of common models. *Health Science Reports*, 8(12), e71679. <https://doi.org/10.1002/hsr2.71679>
3. Arias de la Torre, J., Dahl, R., Alonso, J., Bakolis, I., Botella-Juan, L., Gonzalez-Diez, A., Juruena, M. F., Martín, V., Martínez-Alés, G., Munblit, D., Vilagut, G., Dregan, A., & Valderas, J. M. (2026). Use and performance of language-based artificial intelligence (AI) models to screen for depressive disorders. *Current Psychiatry Reports*, 28(1), 27. <https://doi.org/10.1007/s11920-026-01676-2>
4. Atturu, H., Naraganti, S., & Rao, B. R. (2025). Effectiveness of artificial intelligence–based platform in administering therapies for children with autism spectrum disorder: 12-month observational study. *JMIR Neurotechnology*, 4, e70589. <https://doi.org/10.2196/70589>
5. Auf, H., Svedberg, P., Nygren, J., Nair, M., & Lundgren, L. E. (2025). The use of AI in mental health services to support decision-making: Scoping review. *Journal of Medical Internet Research*, 27, e63548. <https://doi.org/10.2196/63548>
6. Avula, V. C. R., & Amalakanti, S. (2024). Artificial intelligence in psychiatry, present trends, and challenges: An updated review. *Archives of Mental Health*, 25(1), 85–90. [https://doi.org/10.4103/amh.amh\\_167\\_23](https://doi.org/10.4103/amh.amh_167_23)
7. Babu, A., & Joseph, A. P. (2024). Artificial intelligence in mental healthcare: Transformative potential vs. the necessity of human interaction. *Frontiers in Psychology*, 15, 1378904. <https://doi.org/10.3389/fpsyg.2024.1378904>
8. Baydili, İ., Tasci, B., & Tasci, G. (2025). Artificial intelligence in psychiatry: A review of biological and behavioral data analyses. *Diagnostics*, 15(4), 434. <https://doi.org/10.3390/diagnostics15040434>
9. Benoit, J., Onyeaka, H., Keshavan, M., & Torous, J. (2020). Systematic review of digital phenotyping and machine learning in psychosis spectrum illnesses. *Harvard Review of Psychiatry*, 28(5), 296–304. <https://doi.org/10.1097/hrp.0000000000000268>
10. Bernstorff, M., & Jepsen, O. H. (2025). Precision psychiatry needs causal inference. *Acta Neuropsychiatrica*, 37, e32. <https://doi.org/10.1017/neu.2024.29>
11. Bobkov, A., Cheng, F., Xu, J., Bobkova, T., Deng, F., He, J., Jiang, X., Khuzin, D., & Kang, Z. (2025). Telepsychiatry and artificial intelligence: A structured review of emerging approaches to accessible psychiatric care. *Healthcare*, 13(11), 1348. <https://doi.org/10.3390/healthcare13111348>

12. Briganti, G., & Lechien, J. R. (2025). Voice quality as digital biomarker in bipolar disorder: A systematic review. *Journal of Voice*. Advance online publication. <https://doi.org/10.1016/j.jvoice.2025.01.002>
13. Chakraborty, C., Bhattacharya, M., Pal, S., & Lee, S. S. (2024). From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Current Research in Biotechnology*, 7, 100164. <https://doi.org/10.1016/j.crbiot.2023.100164>
14. Chen, C., Lam, K. T., Yip, K. M., So, H. K., Lum, T. Y. S., Wong, I. C. K., Yam, J. C., Chui, C. S. L., & Ip, P. (2025). Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: Pilot randomized controlled trial. *JMIR Human Factors*, 12, e65785. <https://doi.org/10.2196/65785>
15. Chen, Z. S., Schultebrasucks, K., & Wu, W. (2026). A cautionary tale for AI and machine learning in psychiatry. *Translational Psychiatry*, 16(1), 136. <https://doi.org/10.1038/s41398-026-03930-w>
16. Ciampelli, S., de Boer, J. N., Koops, S., Troelstra, E., Jebens, A., Marsman, J. B. C., Smit, A. C., Hossein Nikzad, A., Partlan, R., Homan, P., Hinzen, W., Tang, S. X., & Sommer, I. E. C. (2026). Automated speech-based modeling of item-level symptom severity in schizophrenia. *JAMA Network Open*, 9(6), e2620239. <https://doi.org/10.1001/jamanetworkopen.2026.20239>
17. Cruz-Gonzalez, P., He, A. W. J., Lam, E. P., Ng, I. M. C., Li, M. W., Hou, R., Chan, J. N. M., Sahni, Y., Vinas Guasch, N., Miller, T., Lau, B. W. M., & Sánchez Vidana, D. I. (2025). Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine*, 55, e18. <https://doi.org/10.1017/s0033291724003295>
18. Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, Darzi, A., Holmes, C., Yau, C., Moher, D., Ashrafian, H., Deeks, J. J., Ferrante di Ruffano, L., Faes, L., Keane, P. A., Vollmer, S. J., SPIRIT-AI and CONSORT-AI Consensus Group, Lee, A. Y., ..., Rowley, S. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine*, 26(9), 1351–1363. <https://doi.org/10.1038/s41591-020-1037-7>
19. Curtiss, J., & DiPietro, C. (2025). Machine learning in the prediction of treatment response for emotional disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, 120, 102593. <https://doi.org/10.1016/j.cpr.2025.102593>
20. Cuthbert, B. N. (2022). Research domain criteria (RDoC): Progress and potential. *Current Directions in Psychological Science*, 31(2), 107–114. <https://doi.org/10.1177/09637214211051363>
21. De Francesco, S., Archetti, D., Baronio, C. M., Demaria, C., Boccali, A., Crema, C., Tura, G. B., & Redolfi, A. (2026). Optimizing treatment strategies in the bipolar disorder spectrum with classical AI approaches: Systematic review of performance, bias, and clinical applicability. *JMIR Mental Health*, 13, e93307.
22. de Mattos, B. P., Mattjie, C., Ravazio, R., Barros, R. C., & Grassi-Oliveira, R. (2026). Craving for a robust methodology: A systematic review of machine learning algorithms on substance-use disorders treatment outcomes. *International Journal of Mental Health and Addiction*, 24(2), 1090–1117. <https://doi.org/10.1007/s11469-024-01403-z>
23. Degante-Aguilar, E., Melendez-Armenta, R. A., Luna-Chontal, G., & Fernandez-Dominguez, F. J. (2025). Artificial intelligence techniques applied to anxiety disorders recognition: A systematic review. *Frontiers in Digital Health*, 7, 1646724.
24. Della Greca, A., Amaro, I., Barra, P., Rosapepe, E., & Tortora, G. (2024). Enhancing therapeutic engagement in mental health through virtual reality and generative AI: A co-creation approach to trust building. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 6805–6811). IEEE. <https://doi.org/10.1109/bibm62325.2024.10822177>
25. Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1), 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
26. El Arab, R. A., Mustafa, M. H., Almagharbeh, W. T., Saleem, N. H., Al Abdulmohsen, S., Boathab, R., & Bu Washl, M. (2026). Beyond model development in healthcare AI: post-development robustness, post-deployment monitoring, and lifecycle governance—a scoping review of reviews. *Healthcare*, 14(11), 1459. <https://doi.org/10.3390/healthcare14111459>
27. Fallahpour, B., Dastjerdi, G., Akbarian, E., Emarati, A., Sadr, Z., Dastgheib, S. A., Shahbazi, A., Bahrami, R., Golshan-Tafti, M., Shiri, A., Nematzadeh, F., & Nematzadeh, H. (2026). Artificial intelligence in ADHD diagnosis: A comprehensive review of machine learning applications, clinical validation challenges, and implementation barriers in precision medicine. *Egyptian Pediatric Association Gazette*, 74(1), 36. <https://doi.org/10.1186/s43054-026-00530-7>

28. Fisher, C. E. (2025). The real ethical issues with AI for clinical psychiatry. *International Review of Psychiatry*, 37(1), 14–20. <https://doi.org/10.1080/09540261.2024.2376575>
29. Fisher, H., Jaffe, N. M., Pidvirny, K., Tierney, A. O., Vaidean, M. S., Dongre, P., & Webb, C. A. (2026). Language-based detection of depression with machine learning: Systematic review and meta-analysis. *npj Digital Medicine*, 9(1), 273. <https://doi.org/10.1038/s41746-026-02448-1>
30. Fornaro, M., Miola, A., De Berardis, D., Squassina, A., Manchia, M., & Solmi, M. (2025). Perspectives on precision psychiatry using antipsychotics in the management of bipolar disorder. *Brain Sciences*, 15(5), 430. <https://doi.org/10.3390/brainsci15050430>
31. García-Gutiérrez, M. S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A., & Manzanares, J. (2020). Biomarkers in psychiatry: Concept, definition, types and relevance to the clinical reality. *Frontiers in Psychiatry*, 11, 432. <https://doi.org/10.3389/fpsy.2020.00432>
32. Garzón-Partida, A. P., Padilla-Gómez, C. B., Martínez-Fernández, D. E., García-Estrada, J., Luquin, S., & Fernández-Quezada, D. (2025). The implementation of digital biomarkers in the diagnosis, treatment and monitoring of mood disorders: A narrative review. *Frontiers in Digital Health*, 7, 1595243. <https://doi.org/10.3389/fdgth.2025.1595243>
33. Goktas, P., & Grzybowski, A. (2025). Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine*, 14(5), 1605. <https://doi.org/10.3390/jcm14051605>
34. Guan, H., Bates, D., & Zhou, L. (2026). Keeping medical AI healthy and trustworthy: A review of detection and correction methods for system degradation. *IEEE Transactions on Biomedical Engineering*. Advance online publication. <https://doi.org/10.1109/tbme.2025.3642706>
35. He, F., Huang, S., Wang, R., Chang, A., Phillips, J. L., & Sun, C. (2026). Machine learning for comparative antidepressant selection in major depressive disorder: Systematic review. *JMIR Mental Health*, 13, e89352. <https://doi.org/10.2196/89352>
36. Holmes, G., Tang, B., Gupta, S., Venkatesh, S., Christensen, H., & Whitton, A. (2025). Applications of large language models in the field of suicide prevention: Scoping review. *Journal of Medical Internet Research*, 27, e63126. <https://doi.org/10.2196/63126>
37. Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, 24(3), 383–394. <https://doi.org/10.1002/wps.21352>
38. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or “AI psychosis”. *JMIR Mental Health*, 12, e85799. <https://doi.org/10.2196/85799>
39. Idowu, O., Aderinto, N., Olatunji, G., & Kokori, E. (2025). Machine learning in schizophrenia: A systematic review and meta-analysis of diagnostic and predictive models. *BJPsych Open*, 11(S1), S44. <https://doi.org/10.1192/bjo.2025.10148>
40. Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7), 748–751. <https://doi.org/10.1176/appi.ajp.2010.09091379>
41. Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, 11(3), 285. <https://doi.org/10.3390/healthcare11030285>
42. Jalali, S., You, Q., Xu, V., Chen, L., Zini, J., Zamorano, T., Luo, Y., Friesen, T., Fontana, G. J., & Benrimoh, D. (2025). The use of artificial intelligence for personalized treatment in psychiatry. *Current Psychiatry Reports*, 28(1), 7. <https://doi.org/10.1007/s11920-025-01656-y>
43. Joudar, S. S., Albahri, A. S., Hamid, R. A., Zahid, I. A., Alqaysi, M. E., Albahri, O. S., & Alamoodi, A. H. (2023). Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: A systematic review of current trends and open issues. *Artificial Intelligence Review*, 56(S1), 53–117. <https://doi.org/10.1007/s10462-023-10536-x>
44. Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Maruthamuthu, A. (2025). A literature review on applications of explainable artificial intelligence (XAI). *IEEE Access*, 13, 41111–41140. <https://doi.org/10.1109/access.2025.3546681>

45. Kamath, J., Barriera, R. L., Jain, N., Keisari, E., & Wang, B. (2022). Digital phenotyping in depression diagnostics: Integrating psychiatric and engineering perspectives. *World Journal of Psychiatry*, 12(3), 393–409. <https://doi.org/10.5498/wjp.v12.i3.393>
46. Kas, M. J. H., Penninx, B. W. J. H., Knudsen, G. M., Cuthbert, B., Falkai, P., Sachs, G. S., Ressler, K. J., Białkowicz-Iskra, E., Butlen-Ducuing, F., Leboyer, M., Marston, H., Luthman, J., & Mantua, V. (2025). Precision psychiatry roadmap: Towards a biology-informed framework for mental disorders. *Molecular Psychiatry*, 30(8), 3846–3855. <https://doi.org/10.1038/s41380-025-03070-5>
47. Khalil, S. S., Tawfik, N. S., & Spruit, M. (2024). Exploring the potential of federated learning in mental health research: A systematic literature review. *Applied Intelligence*, 54(2), 1619–1636. <https://doi.org/10.1007/s10489-023-05095-1>
48. Kolding, S., Lundin, R. M., Hansen, L., & Østergaard, S. D. (2025). Use of generative artificial intelligence (AI) in psychiatry and mental health care: A systematic review. *Acta Neuropsychiatrica*, 37, e37. <https://doi.org/10.1017/neu.2024.50>
49. Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46(1), 176–190. <https://doi.org/10.1038/s41386-020-0767-z>
50. Kotochinsky, M., Fonseca, P. E. O., Lopera, V. R., Mora, L., Amador, W. F. O., Sirena, E. C. T., de Melo Guimarães, F. B., Herlyn, D. L., Sherpa, N. N., Gonzalez Lezana, A., & Fagundes, T. P. (2025). Comparative diagnostic performance of artificial intelligence models in structural MRI for schizophrenia: A systematic review and meta-analysis. *Asian Journal of Psychiatry*, 114, 104759. <https://doi.org/10.1016/j.ajp.2025.104759>
51. Krishnadas, R., Leighton, S. P., & Jones, P. B. (2025). Precision psychiatry: Thinking beyond simple prediction models – enhancing causal predictions. *The British Journal of Psychiatry*, 226(3), 184–188. <https://doi.org/10.1192/bjp.2024.258>
52. Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M. A., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics*, 13(15), 2582. <https://doi.org/10.3390/diagnostics13152582>
53. Lieslehto, J., Tihiainen, J., Lähteenvuo, M., Kautzky, A., Akhtar, A., Ármannsdóttir, B., Leucht, S., Correll, C. U., Mittendorfer-Rutz, E., Tanskanen, A., & Taipale, H. (2025). Machine learning-based mortality risk assessment in first-episode bipolar disorder: A transdiagnostic external validation study. *eClinicalMedicine*, 81, 103108. <https://doi.org/10.1016/j.eclinm.2025.103108>
54. Linardon, J. (2025). Navigating the future of psychiatry: A review of research on opportunities, applications, and challenges of artificial intelligence. *Current Treatment Options in Psychiatry*, 12(1), 8. <https://doi.org/10.1007/s40501-025-00344-1>
55. Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law*, 37(3), 214–222. <https://doi.org/10.1002/bsl.2392>
56. Liu, G. D., Li, Y. C., Zhang, W., & Zhang, L. (2020). A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering*, 6(4), 462–467. <https://doi.org/10.1016/j.eng.2019.06.008>
57. Liu, L., Li, Z., Hu, Y., Li, C., He, S., Zhang, S., Gao, J., Zhu, H., & Huang, G. (2025). Predictive performance of machine learning for suicide in adolescents: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 27, e73052. <https://doi.org/10.2196/73052>
58. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, Chan, A. W., Darzi, A., Holmes, C., Yau, C., Ashrafian, H., Deeks, J. J., Ferrante di Ruffano, L., Faes, L., Keane, P. A., Vollmer, S. J., SPIRIT-AI and CONSORT-AI Consensus Group, Lee, A. Y., ..., Rowley, S. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine*, 26(9), 1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
59. Mahmud, M., Kaiser, M. S., Rahman, M. A., Wadhera, T., Brown, D. J., Shopland, N., Burton, A., Hughes-Roberts, T., Mamun, S. A., Ieracitano, C., Tania, M. H., Moni, M. A., Islam, M. S., Ray, K., & Hossain, M. S. (2022). Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. In M. Antona & C. Stephanidis (Eds.), *Universal access in human-computer interaction: User and context diversity* (Lecture Notes in Computer Science, Vol. 13309, pp. 356–370). Springer. [https://doi.org/10.1007/978-3-031-05039-8\\_26](https://doi.org/10.1007/978-3-031-05039-8_26)

60. McGenity, C., & Treanor, D. (2021). Guidelines for clinical trials using artificial intelligence – SPIRIT-AI and CONSORT-AI. *The Journal of Pathology*, 253(1), 14–16. <https://doi.org/10.1002/path.5565>
61. Metin, S. Z., Uyulan, Ç. L., Farhad, S., Ergüzel, T. T., Türk, Ö., Metin, B., Çerezci, Ö., & Tarhan, N. (2025). Deep learning-based artificial intelligence can differentiate treatment-resistant and responsive depression cases with high accuracy. *Clinical EEG and Neuroscience*, 56(2), 119–130. <https://doi.org/10.1177/15500594241273181>
62. Milic, J., Zrnica, I., Grego, E., Jovic, D., Stankovic, V., Djurdjevic, S., & Sapic, R. (2025). The role of artificial intelligence in managing bipolar disorder: A new frontier in patient care. *Journal of Clinical Medicine*, 14(7), 2515. <https://doi.org/10.3390/jcm14072515>
63. Močnik, G., Rehberger, A., Smogavc, A., Mlakar, I., Smrke, U. K., & Močnik, S. (2025). Multimodal observable cues in mood, anxiety, and borderline personality disorders: A review of reviews to inform explainable AI in mental health. *Frontiers in Artificial Intelligence*, 8, 1696448. <https://doi.org/10.3389/frai.2025.1696448>
64. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530. [https://doi.org/10.1016/s2215-0366\(25\)00396-7](https://doi.org/10.1016/s2215-0366(25)00396-7)
65. Narimani, M., & Naeim, M. (2025). Artificial intelligence in mental health: Integrating opportunities and challenges of multimodal deep learning for mental disorder prevention and treatment. *Annals of Medicine & Surgery*, 87(9), 5757–5761. <https://doi.org/10.1097/ms9.0000000000003624>
66. Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, 3, 100099. <https://doi.org/10.1016/j.jglmedi.2024.100099>
67. Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S., & Adrien, V. (2023). Digital phenotyping: Data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*, 25, e44502. <https://doi.org/10.2196/44502>
68. Pan, Y., Wang, P., Xue, B., Liu, Y., Shen, X., Wang, S., & Wang, X. (2025). Machine learning for the diagnosis accuracy of bipolar disorder: A systematic review and meta-analysis. *Frontiers in Psychiatry*, 15, 1515549. <https://doi.org/10.3389/fpsy.2024.1515549>
69. Papini, S., Norman, S. B., Campbell-Sills, L., Sun, X., He, F., Kessler, R. C., Ursano, R. J., Jain, S., & Stein, M. B. (2023). Development and validation of a machine learning prediction model of posttraumatic stress disorder after military deployment. *JAMA Network Open*, 6(6), e2321273. <https://doi.org/10.1001/jamanetworkopen.2023.21273>
70. Perry, B. I., Osimo, E. F., Si, S., Hitchins, K. V. B., Lewis, C., Laws, B., Griffin, S. J., Khandaker, G. M., Murray, G. K., Shiers, D., Chew-Graham, C. A., Jones, P. B., Denniston, A. K., Bardus, M., Jowett, S., Walsh, A. E. L., Arshad, S., Formanek, T., Pillinger, T., ..., Zavala, G. A. (2026). Cardiometabolic prediction models for young people with psychosis spectrum disorders in the UK (psymetric 2.0): A retrospective, multicohort clinical prediction model study. *The Lancet Psychiatry*, 13(4), 291–303. [https://doi.org/10.1016/s2215-0366\(25\)00398-0](https://doi.org/10.1016/s2215-0366(25)00398-0)
71. Pignoni, A., Delvecchio, G., Turtulici, N., Madonna, D., Pietrini, P., Cecchetti, L., & Brambilla, P. (2024). Machine learning and the prediction of suicide in psychiatric populations: A systematic review. *Translational Psychiatry*, 14(1), 140. <https://doi.org/10.1038/s41398-024-02852-9>
72. Putica, A., Khanna, R., Bosl, W., Saraf, S., & Edgcomb, J. (2025). Ethical decision-making for AI in mental health: The integrated ethical approach for computational psychiatry (IEACP) framework. *Psychological Medicine*, 55, e213. <https://doi.org/10.1017/s0033291725101311>
73. Qi, W., Zhu, X., Wang, B., Shi, Y., Dong, C., Shen, S., Li, J., Zhang, K., He, Y., Zhao, M., Yao, S., Dong, Y., Shen, H., Kang, J., Lu, X., Jiang, G., Boots, L. M. M., Fu, H., Pan, L., ..., Cao, S. (2025). Alzheimer's disease digital biomarkers multidimensional landscape and AI model scoping review. *npj Digital Medicine*, 8(1), 366. <https://doi.org/10.1038/s41746-025-01640-z>
74. Raket, L. L., Jaskolowski, J., Kinon, B. J., Brasen, J. C., Jönsson, L., Wehnert, A., & Fusar-Poli, P. (2020). Dynamic electronic health record detection (DETECT) of individuals at risk of a first episode of psychosis: A case-control development and validation study. *The Lancet Digital Health*, 2(5), e229–e239. [https://doi.org/10.1016/s2589-7500\(20\)30024-8](https://doi.org/10.1016/s2589-7500(20)30024-8)
75. Razaghi, M., Hafez, A., Farina, J. M., Scalia, I. G., Pereyra, M., Abdelfattah, F. E., Sheashaa, H., Awad, K., Lester, S. J., Ayoub, C., & Arsanjani, R. (2026). Transforming clinical documentation with ambient artificial intelligence (AI) scribes: A narrative review of technology, impact, and implementation. *Cardiovascular Diagnosis and Therapy*, 16(1), 11. <https://doi.org/10.21037/cdt-2025-454>

76. Ribeiro Junior, H. L., Nepomuceno, F. W. A. B., & Pessoa, C. D. Ó. (2025). AI in clinical trials is missing from CONSORT and SPIRIT 2025 guidelines. *The Lancet*, 406(10498), 25. [https://doi.org/10.1016/s0140-6736\(25\)01268-1](https://doi.org/10.1016/s0140-6736(25)01268-1)
77. Richter, T., Shani, R., Tal, S., Derakshan, N., Cohen, N., Enock, P. M., McNally, R. J., Mor, N., Daches, S., Williams, A. D., Yiend, J., Carlbring, P., Kuckertz, J. M., Yang, W., Reinecke, A., Beevers, C. G., Bunnell, B. E., Koster, E. H. W., Zilcha-Mano, S., & Okon-Singer, H. (2025). Machine learning meta-analysis identifies individual characteristics moderating cognitive intervention efficacy for anxiety and depression symptoms. *npj Digital Medicine*, 8(1), 65. <https://doi.org/10.1038/s41746-025-01449-w>
78. Romeo, G., & Conti, D. (2026). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. *AI & Society*, 41(1), 259–278. <https://doi.org/10.1007/s00146-025-02422-7>
79. Rony, M. K. K., Das, D. C., Khatun, M. T., Ferdousi, S., Akter, M. R., Khatun, M. A., Begum, M. H., Khalil, M. I., Parvin, M. R., Alrazeeni, D. M., & Akter, F. (2025). Artificial intelligence in psychiatry: A systematic review and meta-analysis of diagnostic and therapeutic efficacy. *DIGITAL HEALTH*, 11, 20552076251330528. <https://doi.org/10.1177/20552076251330528>
80. Ruan, Q. N., Hu, S. Q., ShangGuan, Z. H., & Zhou, S. M. (2026). The augmented clinician as a framework for human-AI collaboration in mental healthcare. *Frontiers in Psychiatry*, 17, 1729175. <https://doi.org/10.3389/fpsyt.2026.1729175>
81. Saboori Amleshi, R., Ilaghi, M., Rezaei, M., Zangiabadian, M., Rezazadeh, H., Wegener, G., & Arjmand, S. (2025). Predictive utility of artificial intelligence on schizophrenia treatment outcomes: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 169, 105968. <https://doi.org/10.1016/j.neubiorev.2024.105968>
82. Singh, S. U., & Namin, A. S. (2025). A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal*, 10, 100128. <https://doi.org/10.1016/j.nlp.2025.100128>
83. Sohn, J. S., Ha, B. G., Park, S., Kim, J. J., Lee, E., Oh, H., Lee, S., & Kim, E. (2026). Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *npj Digital Medicine*, 9(1), 377. <https://doi.org/10.1038/s41746-026-02566-w>
84. Solek, P., Nurfitri, E., Sahril, I., Prasetya, T., Rizqiamuti, A. F., Burhan, B., Rachmawati, I., Gamayani, U., Rusmil, K., Chandra, L. A., Afriandi, I., & Gunawan, K. (2025). The role of artificial intelligence for early diagnostic tools of autism spectrum disorder: A systematic review. *Turkish Archives of Pediatrics*, 60(2), 126–140. <https://doi.org/10.5152/TurkArchPediatri.2025.24183>
85. Song, J., Cho, E., Lee, H., Lee, S., Kim, S., & Kim, J. (2025). Development of neurodegenerative disease diagnosis and monitoring from traditional to digital biomarkers. *Biosensors*, 15(2), 102. <https://doi.org/10.3390/bios15020102>
86. Spittal, M. J., Guo, X. A., Kang, L., Kirtley, O. J., Clapperton, A., Hawton, K., Kapur, N., Pirkis, J., & Carter, G. (2025). Machine learning algorithms and their predictive accuracy for suicide and self-harm: Systematic review and meta-analysis. *PLOS Medicine*, 22(9), e1004581. <https://doi.org/10.1371/journal.pmed.1004581>
87. Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, 10(1), 10. <https://doi.org/10.1186/s40708-023-00188-6>
88. Stein, D. J., Shoptaw, S. J., Vigo, D. V., Lund, C., Cuijpers, P., Bantjes, J., Sartorius, N., & Maj, M. (2022). Psychiatric diagnosis and treatment in the 21st century: Paradigm shifts versus incremental integration. *World Psychiatry*, 21(3), 393–414.
89. Su, S., Wang, Y., Jiang, W., Zhao, W., Gao, R., Wu, Y., Tao, J., Su, Y., Zhang, J., Li, K., Zhang, Z., Zhao, M., Wang, Z., Luo, Y., Huang, X., Wang, L., Wang, X., Li, Y., Jia, Q., ..., Xu, Y. (2022). Efficacy of artificial intelligence-assisted psychotherapy in patients with anxiety disorders: A prospective, national multicenter randomized controlled trial protocol. *Frontiers in Psychiatry*, 12, 799917. <https://doi.org/10.3389/fpsyt.2021.799917>
90. Sun, J., Lu, T., Shao, X., Han, Y., Xia, Y., Zheng, Y., Wang, Y., Li, X., Ravindran, A., Fan, L., Fang, Y., Zhang, X., Ravindran, N., Wang, Y., Liu, X., & Lu, L. (2025). Practical AI application in psychiatry: Historical review and future directions. *Molecular Psychiatry*, 30(9), 4399–4408. <https://doi.org/10.1038/s41380-025-03072-3>
91. Tahir, W. B., Khalid, S., Almutairi, S., Abohashrh, M., Memon, S. A., & Khan, J. (2025). Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. *IEEE Access*, 13, 12789–12818. <https://doi.org/10.1109/access.2025.3530862>
92. Tay, J. L., Ang, Y. L., Tam, W. W. S., & Sim, K. (2025). Accuracy of machine learning methods in predicting prognosis of patients with psychotic spectrum disorders: A systematic review. *BMJ Open*, 15(2), e084463. <https://doi.org/10.1136/bmjopen-2024-084463>

93. Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazzeri, L., Traver, V., & Novillo-Ortiz, D. (2023). Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review. *JMIR Mental Health*, 10, e42045. <https://doi.org/10.2196/42045>
94. Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A., & Firth, J. (2025). The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*, 24(2), 156–174. <https://doi.org/10.1002/wps.21299>
95. Vali, M., Nezhad, H. M., Kovacs, L., & Gandomi, A. H. (2025). Machine learning algorithms for predicting PTSD: A systematic review and meta-analysis. *BMC Medical Informatics and Decision Making*, 25(1), 34. <https://doi.org/10.1186/s12911-024-02754-2>
96. Vandemeulebroucke, T. (2025). The ethics of artificial intelligence systems in healthcare and medicine: From a local to a global perspective, and back. *Pflügers Archiv - European Journal of Physiology*, 477(4), 591–601. <https://doi.org/10.1007/s00424-024-02984-3>
97. Vorstman, J., Sebat, J., Bourque, V. R., & Jacquemont, S. (2025). Integrative genetic analysis: Cornerstone of precision psychiatry. *Molecular Psychiatry*, 30(1), 229–236. <https://doi.org/10.1038/s41380-024-02706-2>
98. Wang, L., Wang, C., Li, C., Murai, T., Bai, Y., Song, Z., Zhang, S., Zhang, Q., Huang, Y., Bi, X., & Jiang, J. (2025). AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis. *npj Digital Medicine*, 8(1), 523. <https://doi.org/10.1038/s41746-025-01933-3>
99. Wang, Y., & Pan, Y. (2026). Development and validation of a machine learning model for predicting depression risk in rural Chinese older adults: Evidence from the CHARLS cohort. *DIGITAL HEALTH*, 12, 20552076251411601. <https://doi.org/10.1177/20552076251411601>
100. Watts, D., Pulice, R. F., Reilly, J., Brunoni, A. R., Kapczinski, F., & Passos, I. C. (2022). Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational Psychiatry*, 12(1), 332. <https://doi.org/10.1038/s41398-022-02064-z>
101. Wilhelmy, S., Giupponi, G., Groß, D., Eisendle, K., & Conca, A. (2023). A shift in psychiatry through AI? Ethical challenges. *Annals of General Psychiatry*, 22(1), 43. <https://doi.org/10.1186/s12991-023-00476-9>
102. World Health Organization. (2025). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. <https://www.who.int/publications/i/item/9789240084759>
103. Yan, A. Y., Speed, T. J., & Taylor, C. O. (2025). Relapse prediction using wearable data through convolutional autoencoders and clustering for patients with psychotic disorders. *Scientific Reports*, 15(1), 18806.
104. Yang, Y., Zhao, X., Zhao, P., Ying, D., Wang, J., Jiang, Y., & Wan, Q. (2026). AI-driven speech biomarkers for disease diagnosis and monitoring: A systematic review and meta-analysis. *BMJ Evidence-Based Medicine*, 31(1), 46–56. <https://doi.org/10.1136/bmjebm-2025-113759>
105. Zaheer, A., & Akhtar, A. (2025). Artificial intelligence as a support to diagnose ADHD: An insight of unorthodox approaches: A scoping review. *Child Neuropsychology*, 31(8), 1324–1358. <https://doi.org/10.1080/09297049.2025.2468411>
106. Zhang, M., Ding, H., Naumceska, M., & Zhang, Y. (2022). Virtual reality technology as an educational and intervention tool for children with autism spectrum disorder: Current perspectives and future directions. *Behavioral Sciences*, 12(5), 138. <https://doi.org/10.3390/bs12050138>
107. Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F. H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*, 27, e78238. <https://doi.org/10.2196/78238>
108. Zhang, S., Zhang, C., & Zhang, J. (2026). Text-based depression estimation using machine learning with standard labels: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 28, e82686. <https://doi.org/10.2196/82686>
109. Zhao, C., Xu, Y., Li, R., Li, H., & Zhang, M. (2025). Artificial intelligence in ADHD assessment: A comprehensive review of research progress from early screening to precise differential diagnosis. *Frontiers in Artificial Intelligence*, 8, 1624485. <https://doi.org/10.3389/frai.2025.1624485>
110. Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders*, 356, 459–469. <https://doi.org/10.1016/j.jad.2024.04.057>
111. Zidaru, T., Morrow, E. M., & Stockley, R. (2021). Ensuring patient and public involvement in the transition to AI-assisted mental health care: A systematic scoping review and agenda for design justice. *Health Expectations*, 24(4), 1072–1124. <https://doi.org/10.1111/hex.13299>