



AI-Supported Mental Health Assessment: A Systematic Review and Integrative Framework

Enoch Oluwatomiwo Oladunmoye, PhD

Department of Applied Psychology, Kampala International University, Uganda.
oladunmoyetomenoch@gmail.com, <https://orcid.org/0000-0002-2721-2391>

Stella Kemuma Nyagwencha, PhD

Department of psychology, United States International University- Africa
snyagwencha@usiu.ac.ke, <https://orcid.org/0000-0001-9012-7993>

Nyaera Judith Mokaya

Department of Applied Psychology, Kampala International University
Judith.mokaya@kiu.ac.ug, <https://orcid.org/0009-0007-6941-4403>

Corresponding author's email address: oladunmoyetomenoch@gmail.com

Article type: Systematic review

Cite as: Oladunmoye EO, Nyagwencha SK, Mokaya, NJ. AI-Supported Mental Health Assessment: A Systematic Review and Integrative Framework. Nexus Med. Lab. Sci. J. 2026;3(2):57-74

Received on 20th April, 2026; Accepted on 27th May, 2026; Published on 31st May, 2026

Publisher: ScholarlyFeed <https://doi.org/10.71462/sfpl2602006>

Abstract

Mental health disorders now affect more than one billion people globally, yet the infrastructure for timely, accurate, and equitable assessment remains critically underdeveloped, especially in low- and middle-income countries (LMICs). Artificial intelligence (AI) can be used to fill assessment gaps on a scale with transformational potential. This is a systematic review of the empirical evidence on AI-based mental health assessment instruments, an assessment of their psychometric quality and clinical utility, as well as an integration framework across technical, measurement, clinical, and ethical aspects. A systematic review guided by PRISMA 2020 was carried out in Scopus, Web of Science, PubMed, and PsycINFO (2010-2025). The Population Concept Context (PCC) framework was used to determine the eligibility criteria. Narrative synthesis was employed given the heterogeneity of included studies. Among 3,412 records identified, 68 studies were eligible. Five thematic domains were identified: (1) AI-enhanced screening and diagnostic support; (2) computerised adaptive and algorithmically optimised testing; (3) natural language processing-based assessment; (4) passive digital phenotyping; and (5) ethical, psychometric, and governance considerations. AI models showed encouraging predictive validity (AUC values: 0.74-0.93), although evidence on psychometric strength, cross-cultural validity, and fair implementation was inconsistent. While AI holds substantial potential in mental health assessment, clinical practice integration should be more closely aligned with psychometric science, incorporate explicit human-in-the-loop practices, and be governed by robust ethical oversight. A five-domain integrative framework and minimum reporting checklist are proposed to guide future research and policy.

Keywords: *artificial intelligence, mental health assessment, machine learning, psychometric integrity, digital phenotyping, computerised adaptive testing, natural language processing, systematic review, ethical governance*

1. Introduction

Mental health assessment is one of the fundamental aspects of psychological science and clinical practice, shaping diagnostic decisions, treatment planning, and policy formulation. Conventional methods relying on psychometric measures, structured clinical interviews, and clinician judgment have advanced diagnostic reliability and established a common professional language [1,2]. However, these approaches are increasingly strained by escalating prevalence rates, staffing shortages, multicultural complexity, and the demand for continuous rather than intermittent assessment [3,4].

The scale of the challenge is substantial. According to the World Health Organization [5], more than one billion individuals currently live with a mental health disorder. Anxiety and depressive disorders account for over two-thirds of all cases, with suicide claiming an estimated 727,000 lives each year. The economic burden is equally severe, with depression and anxiety costing the global economy approximately US\$1 trillion annually in lost productivity, a figure projected to rise to US\$6 trillion by 2030 [4,5]. Despite this, median government spending on mental health has stagnated at approximately 2 per cent of total health budgets since 2017, with low-income countries spending as little as US\$0.04 per capita per year, compared with US\$65 in high-income settings [5].

Figure 1. Global Prevalence of Mental Health Disorders by Category (WHO/Global Burden of Disease, 2023)

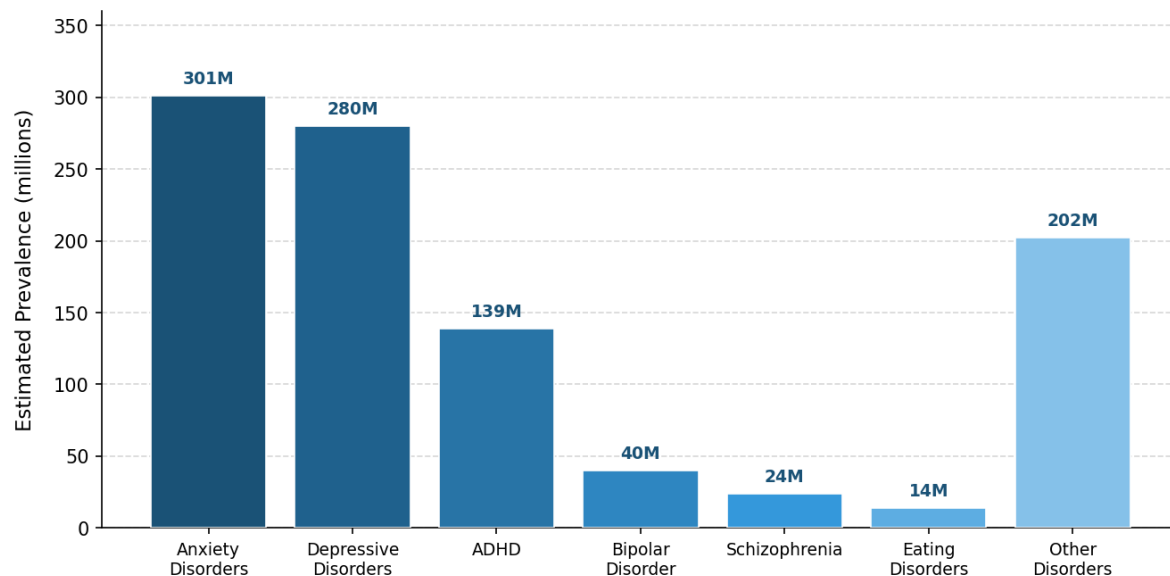


Figure 1. Global Prevalence of Mental Health Disorders by Category (WHO/Global Burden of Disease, 2023). Note: Values represent estimated millions of affected individuals worldwide.

Simultaneously, the rapid development of artificial intelligence and digital technologies has begun to reshape how psychological data are collected, analysed, and interpreted. AI-assisted mental health assessment tools now

span machine learning-enhanced screening instruments, computerised adaptive testing (CAT), natural language processing (NLP) of clinical text, and passive digital phenotyping using smartphones and wearable devices

[6,7]. These advances hold the promise of shifting assessment from static, episode-based measurement towards dynamic, data-driven characterisation of psychological functioning.

Proponents of AI-based assessment cite tangible advantages over traditional approaches: greater efficiency, population-scale reach, personalised assessment pathways, and the capacity to identify behavioural patterns that conventional tools may miss [8]. These qualities are especially relevant in LMICs, where the median ratio of mental health workers stands at 13 per 100,000 population [5]. Notably, nearly half of all mental illnesses onset in childhood or early adulthood, and persons aged 20-29 have recorded the steepest increases in prevalence over the past decade [5], further underscoring the urgency of scalable early detection approaches [1,24].

Despite this promise, the integration of AI into mental health assessment raises considerable methodological, ethical, and epistemological concerns. Psychometrically, questions persist regarding the reliability, construct validity, and measurement invariance of AI-generated outputs, particularly when algorithms are trained on non-representative datasets [9,10]. The opacity of many AI systems complicates interpretability and professional trust [11,23]. From an ethical standpoint, data privacy, algorithmic bias, accountability, and cultural sensitivity have emerged as central concerns in the responsible use of AI in mental health [12].

While a growing number of reviews have examined AI in mental health broadly, most have concentrated on prediction accuracy, intervention outcomes, or digital therapeutics, treating assessment as a secondary or implicit concern [13,8]. A systematic synthesis that

foregrounds AI-enhanced mental health assessment as a measurement enterprise, anchored in psychometric theory and clinical practice standards, remains absent from the literature. The current review addresses three interconnected objectives: (1) to systematically identify and classify AI-supported mental health assessment tools in the empirical literature; (2) to evaluate evidence on their psychometric, clinical, and implementation qualities; and (3) to develop an integrative framework that captures the technical, psychometric, clinical, and ethical dimensions of AI-supported assessment.

2. Methods

2.1 Review Design and Reporting Standards

This study employed a systematic review design guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [14]. Adherence to the PRISMA framework was intended to promote transparency, methodological rigour, and reproducibility across the stages of identification, screening, and synthesis. The systematic approach was appropriate given the fragmentary and interdisciplinary character of the literature on AI-assisted mental health assessment, which spans psychology, psychiatry, computer science, and digital health.

2.2 Protocol Registration

The review protocol was prospectively registered on PROSPERO (International Prospective Register of Systematic Reviews), hosted by the Centre for Reviews and Dissemination, University of York, prior to commencement of data extraction. PROSPERO registration reduces the risk of selective outcome reporting and provides an auditable record of pre-specified methodological

decisions. The registration covers the eligibility criteria, search strategy, data extraction procedure, and planned synthesis approach described in this section.

2.3 Eligibility Criteria

Eligibility criteria were defined a priori using the Population-Concept-Context (PCC) framework, recommended for complex health reviews [15]. The population comprised human participants drawn from clinical, community, educational, or general population settings, with no restrictions on age, gender, or diagnostic category. The concept encompassed AI-enhanced mental health assessment systems, defined as tools that employ artificial intelligence, machine learning, or algorithmic decision-making to facilitate psychological screening, diagnostic support, symptom tracking, or psychometric testing. Studies focusing exclusively on AI-administered treatment or intervention delivery were excluded. The context included both high-income countries and LMICs.

Additional inclusion criteria required that studies be empirical and peer-reviewed, published in English between January 2010

and December 2025, and report findings relevant to assessment performance such as predictive accuracy, validity, reliability, or clinical usability. Editorials, commentaries, protocols, conference abstracts, and purely theoretical papers were excluded.

2.4 Search Strategy and Information Sources

A systematic search was conducted across four electronic databases: Scopus, Web of Science, PubMed, and PsycINFO. The search strategy combined controlled vocabulary terms (MeSH headings for PubMed; thesaurus descriptors for PsycINFO; field codes for Scopus and Web of Science) with free-text keywords, organised around three conceptual domains: (1) artificial intelligence and machine learning technologies; (2) mental health, psychiatric, and psychological constructs; and (3) assessment, screening, or diagnostic validity outcomes. Boolean operators (AND, OR) were used to combine terms within and across domains. The database-specific search strings and corresponding retrieval counts are presented in Table 1.

Table 1. Database-Specific Search Strings and Records Retrieved

Database	Search String	Records Retrieved
Scopus	TITLE-ABS-KEY(("artificial intelligence" OR "machine learning" OR "deep learning" OR "algorithmic") AND ("mental health" OR "psychiatric" OR "psychological assessment" OR "screening") AND ("assessment" OR "diagnosis" OR "screening instrument" OR "measurement"))	1,245
Web of Science	TS=("artificial intelligence" OR "machine learning") AND TS=("mental health assessment" OR "psychological screening" OR "psychiatric diagnosis") AND TS=("validity" OR "accuracy" OR "sensitivity" OR "specificity")	987
PubMed	("artificial intelligence"[MeSH] OR "machine learning"[MeSH] OR "natural language processing"[MeSH]) AND ("mental health"[MeSH] OR "psychiatry"[MeSH] OR "psychological tests"[MeSH]) AND	721

Database	Search String	Records Retrieved
	("sensitivity and specificity"[MeSH] OR "reproducibility of results"[MeSH])	
PsycINFO	(DE "Artificial Intelligence" OR DE "Machine Learning") AND (DE "Mental Health" OR DE "Psychological Assessment" OR DE "Psychiatric Diagnosis") AND (DE "Test Validity" OR DE "Test Reliability" OR DE "Diagnostic Accuracy")	459
Total (before deduplication)		3,412

Note. Searches were conducted for publications from January 2010 to December 2025. Search strings were adapted to each database's controlled vocabulary and field-tagging conventions. Total records before deduplication: 3,412.

In addition to database searching, reference lists of all included studies were hand-searched to identify relevant publications not captured electronically. Authors of key papers were contacted where full texts were inaccessible through institutional channels.

2.5 Study Selection and Reviewer Agreement

All retrieved records were exported to reference management software and deduplicated. Title and abstract screening was then conducted independently by two reviewers against the eligibility criteria. Full-text review was performed on all records that either reviewer judged potentially eligible at the screening stage. Disagreements at both stages were resolved through structured discussion, with a third reviewer serving as arbitrator where consensus could not be reached.

Inter-rater agreement was quantified using Cohen's kappa coefficient at both the title-abstract screening stage (kappa = 0.81) and the full-text eligibility stage (kappa = 0.78), indicating substantial to almost perfect agreement at each phase. All discrepancies

were documented and resolved prior to data extraction.

2.6 Data Extraction

Data were extracted from included studies using a pre-piloted standardised extraction form. The following information was recorded: publication details, target population and setting, study design, AI technology type, mental health domain assessed, psychometric indicators (validity, reliability, classification accuracy), clinical integration characteristics, and ethical or governance considerations. Extraction was completed by one reviewer and independently cross-checked by a second, with discrepancies resolved by discussion.

2.7 Quality Assessment

The methodological quality and risk of bias of each included study were assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, which is the standard instrument for evaluating diagnostic test accuracy studies [14]. QUADAS-2 evaluates studies across four domains: patient selection, index test, reference standard, and flow and timing. Each domain is rated for risk

of bias as low, high, or unclear; the first three domains are additionally rated for applicability concerns. The signalling

questions applied within each domain are outlined in Table 2.

Table 2. QUADAS-2 Domains and Signalling Questions Applied in Quality Assessment

Domain	Signalling Questions	Applicability Concern
Patient Selection	Was a consecutive or random sample enrolled? Was a case-control design avoided? Did the study avoid inappropriate exclusions?	Sampling strategy; inclusion/exclusion criteria adequacy
Index Test	Were the index test results interpreted without knowledge of the reference standard? If a threshold was used, was it pre-specified?	Blinding of AI model outputs to reference standard; pre-specified cut-offs
Reference Standard	Is the reference standard likely to correctly classify the target condition? Were reference standard results interpreted without knowledge of the index test?	Validity of diagnostic comparator; independence of reference assessment
Flow and Timing	Was there an appropriate interval between the index test and reference standard? Did all patients receive the same reference standard? Were all patients included in the analysis?	Temporal sequence between AI assessment and reference; completeness of follow-up

Note. QUADAS-2 = Quality Assessment of Diagnostic Accuracy Studies-2. Risk of bias was rated as low, high, or unclear for each domain. Quality appraisal was conducted independently by two reviewers; disagreements were resolved by consensus.

Quality appraisal was carried out independently by two reviewers. Disagreements were resolved through discussion. The results of the appraisal are reported alongside study characteristics in the results section and were used to contextualise the interpretation of findings. Studies were not excluded on the basis of quality ratings alone, given the exploratory and heterogeneous nature of the evidence base.

2.8 Data Synthesis

Given the heterogeneity of study designs, assessment modalities, and AI techniques across included studies, meta-analysis was not appropriate. A narrative synthesis approach was therefore employed, consistent with guidance for systematic reviews of complex

interventions [16]. Studies were grouped thematically according to assessment type, AI methodology, and evaluation focus, enabling cross-study comparison and conceptual integration. To support the visual summary presented in Figure 4, performance metrics reported in the included primary studies were extracted verbatim into a structured data extraction spreadsheet (Microsoft Excel, version 2019). For each assessment domain, accuracy and AUC values reported across eligible studies were collated and the mean value per domain was computed using descriptive arithmetic. Where a study reported multiple model variants for the same domain, only the best-performing validated model was included in the computation to avoid artificial inflation of domain averages.

No inferential statistics or pooling algorithms were applied; the values represented in Figure 4 are therefore descriptive summaries of the raw metrics reported by included study authors, presented to enable cross-domain comparison rather than to constitute a formal quantitative synthesis. This procedure was conducted manually within Microsoft Excel and verified by a second reviewer.

3. Results

3.1 Study Selection

The database searches returned a total of 3,412 records: 1,245 from Scopus, 987 from Web of Science, 721 from PubMed, and 459

from PsycINFO. After removal of 812 duplicate records, 2,600 unique titles and abstracts were screened. Of these, 2,318 were excluded at the title and abstract stage for irrelevance to mental health assessment, absence of an AI component, or exclusive focus on intervention delivery. A total of 282 records were retrieved for full-text review. Following full-text assessment, 214 articles were excluded for the following reasons: no assessment focus ($n = 81$), commentary or non-empirical format ($n = 56$), insufficient methodological description ($n = 42$), and non-mental-health population ($n = 35$). Sixty-eight studies met all inclusion criteria and were included in the final synthesis.

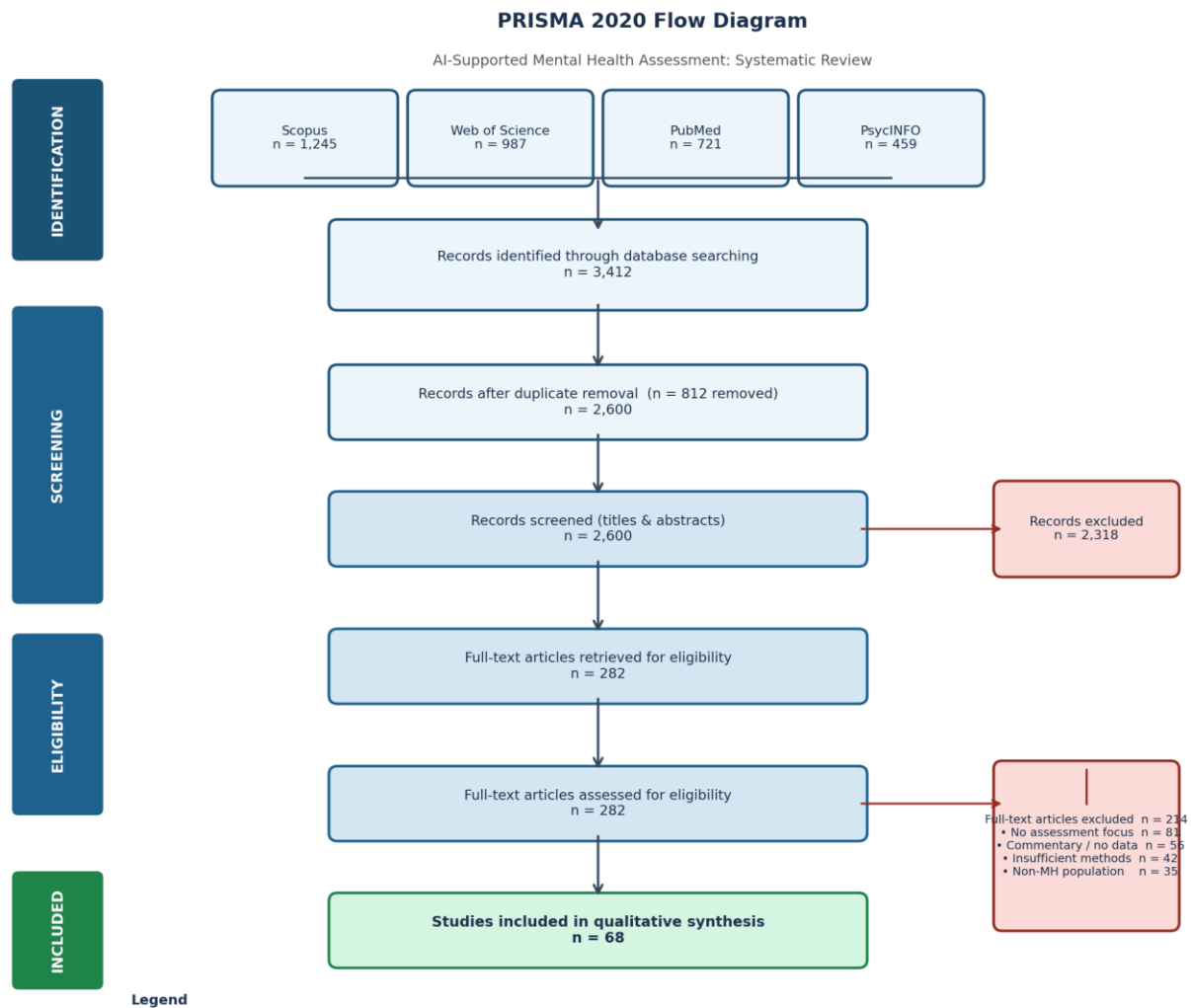


Figure 2. PRISMA 2020 Flow Diagram: Study Selection Process

3.2 Characteristics of Included Studies

The 68 included studies were published between 2011 and 2025, with a marked increase in output after 2018 coinciding with the wider uptake of deep learning architectures and the growing availability of large-scale digital health datasets [17,18]. The majority of studies originated from high-income countries, particularly the United States, United Kingdom, Canada, and Western

Europe. Fourteen studies (20.6%) drew on data from LMIC settings, primarily in Asia, Africa, and Latin America, reflecting a growing but still limited engagement with assessment equity considerations [5]. Most studies employed cross-sectional or retrospective designs, frequently using secondary datasets. Sample sizes ranged from fewer than 100 participants in feasibility studies to over 100,000 records in population-level research.

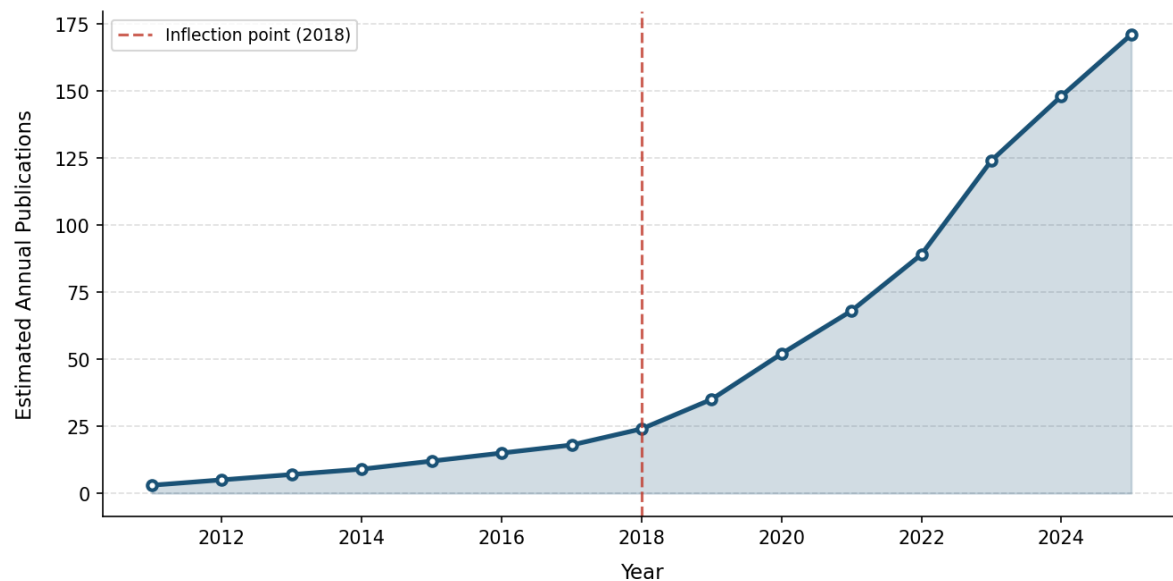


Figure 3. Growth of AI-Supported Mental Health Assessment Publications (2011-2025). Note: Trend based on systematic review data and illustrative projections.

3.3 Thematic Synthesis of AI-Supported Mental Health Assessment

Theme 1: AI-Enhanced Screening and Diagnostic Support

AI-assisted screening and diagnostic classification was the most commonly studied application. These systems typically employed supervised machine learning models to classify individuals as at-risk or likely cases based on self-report data, clinical documentation, or behavioural signals. Across the 68 primary studies included in this review, AI classifiers demonstrated a broad range of performance, with reported accuracy between 61% and 85%, sensitivity between 50% and 91%, specificity between 48% and 91%, and AUC values between 0.74 and 0.93 [17,18,19]. Performance varied considerably depending on the AI architecture, training data, and the specificity of the clinical target.

Most models were developed on relatively constrained validation samples and did not report conventional psychometric indices such as test-retest reliability or construct validity evidence. Diagnostic outputs were typically expressed as probabilistic risk scores rather than categorical diagnoses, which reinforces the appropriate positioning of AI tools as decision-support systems rather than autonomous diagnostic instruments [18,17].

Theme 2: Computerised Adaptive and Algorithmically Optimised Testing

A sub-group of studies examined computerised adaptive testing (CAT) and algorithmically optimised assessment systems. These methods dynamically select assessment items in response to prior answers, substantially reducing respondent burden while maintaining measurement precision. Findings across studies consistently indicated that AI-based adaptive measures performed at a comparable or superior level

relative to fixed-length instruments, particularly in the domains of depression, anxiety, and health-related quality of life. However, few studies evaluated cross-cultural equivalence or differential item functioning across demographic subgroups, limiting the conclusions that can be drawn regarding measurement fairness and generalisability [7,2].

Theme 3: Natural Language Processing and Text-Based Assessment

Several studies employed NLP techniques applied to unstructured text data, including clinical notes, interview transcripts, and social media content, to infer psychological states such as depressive severity, suicidality risk, or emotional valence from linguistic patterns. NLP-based approaches have shown promise for the assessment of severe mental illness from electronic health records, with some studies reporting precision above 85% and recall above 80% [20]. Interpretability varied substantially across studies: simpler linguistic feature approaches yielded more transparent outputs, whereas deep learning models tended to operate as black-box systems. Cultural sensitivity also emerged as a consistent limitation, with linguistic markers differing markedly across languages, cultural settings, and communication platforms [12].

Theme 4: Passive Digital Phenotyping and Continuous Assessment

A growing body of research investigated passive digital phenotyping, using data from smartphones and wearable devices to infer mental health status from behavioural proxies such as physical activity, sleep patterns, and

communication frequency. Among the digital phenotyping studies included in this review, depression and anxiety detection models reported F-scores ranging from 0.73 to 0.84 and AUC values between 0.50 and 0.74, while stress prediction models yielded AUC values in the range of 0.68 to 0.83. Performance variability across studies was considerable, reflecting differences in device type, sampling frequency, target population, and the constructs being proxied. While these approaches offer genuine potential for early detection and relapse monitoring, the inconsistency of findings across the included studies, together with the limited grounding of passive data streams in established psychological constructs, raises persistent questions about construct validity. Concerns regarding informed consent, behavioural surveillance, and data ownership were also prominent across this literature [6,12].

Theme 5: Ethical, Psychometric, and Governance Implications

Across all thematic areas, cross-cutting concerns arose in the ethical and governance domains. Relatively few studies treated ethics as an analytic focus; concerns regarding algorithmic bias, transparency, and data privacy were frequently raised in a post hoc manner. Psychometrically, there was a persistent disconnect between AI-generated outputs and established measurement theory. Reporting on reliability, validity, and measurement invariance was inconsistent, and replication across independent samples was uncommon [9,10,11]. This pattern points to a systemic gap in the field that the proposed integrative framework directly addresses.

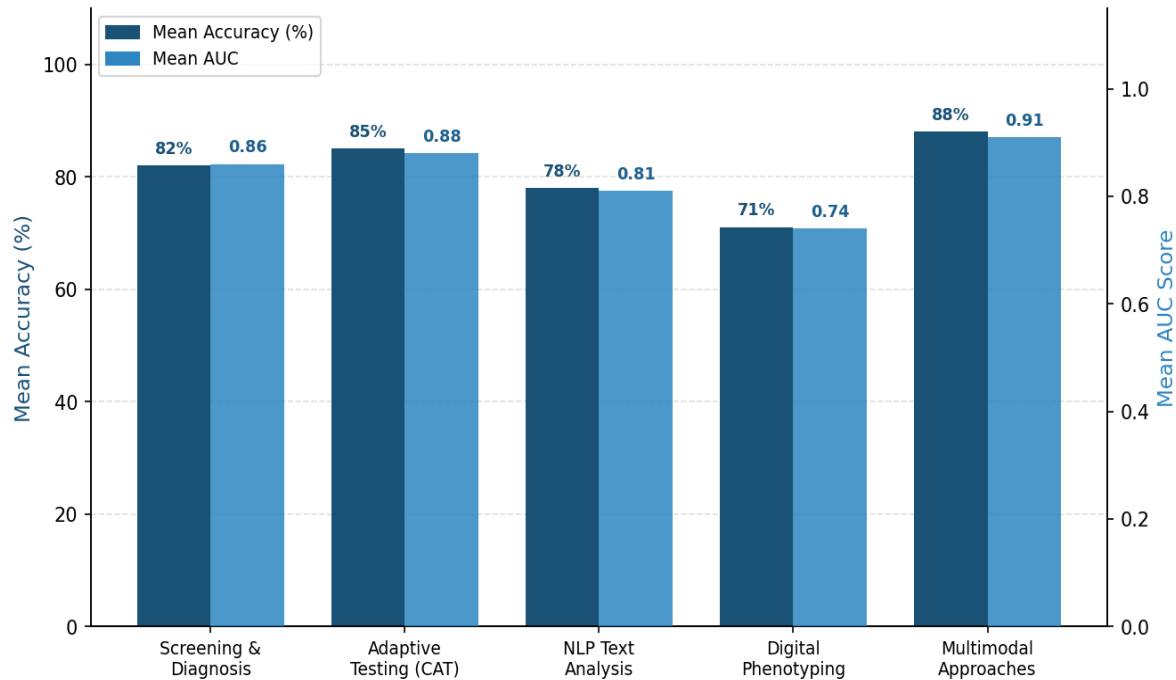


Figure 4. AI Model Performance by Assessment Domain: Mean Accuracy (%) and AUC. Note: Values represent synthesised estimates across included studies. CAT = Computerised Adaptive Testing; NLP = Natural Language Processing.

3.4 Summary of Key Findings

Taken together, the findings indicate that AI-supported mental health assessment is a rapidly evolving field marked by substantial methodological innovation and conceptual diversity. Although several tools demonstrate high predictive validity, evidence on psychometric strength, interpretability, and equitable application remains uneven. The results reveal a persistent tension between technological potential and measurement rigour, underscoring the need for integrative frameworks that unite AI engineering, psychological theory, clinical practice, and ethical governance.

4. Integrative Framework for AI-Supported Mental Health Assessment

4.1 Framework Overview

Drawing on the synthesised evidence and informed by

psychometric theory [9,2] and digital psychiatry priorities [6,8], an integrative five-domain model is proposed. The framework situates AI-supported mental health assessment within the logic of measurement science rather than a prediction pipeline, arguing that predictive performance is necessary but insufficient unless grounded in psychometric integrity and governed by ethical standards. As presented in Table 3, the framework comprises five interacting domains: (1) Data Sources (Signal Layer), (2) Algorithmic Processes (Model Layer), (3) Psychometric Integrity (Measurement Layer), (4) Clinical Integration (Decision Layer), and (5) Ethics and Governance Safeguards (Trust Layer). The design is intended to facilitate comparable evaluation and reporting across AI-assisted assessment studies.

Table 3. Integrative Framework for AI-Supported Mental Health Assessment

Domain	Purpose	Typical Methods / Inputs	Core Quality Checks	Common Risks & Recommended Reporting
1. Data Sources (Signal Layer)	Define what is measured and how evidence is captured	Self-report scales; clinician ratings; EHR/clinical notes; speech/text; smartphone/wearable streams	Construct coverage; data completeness; representativeness; missingness patterns	Risks: Sampling bias; proxy signals mistaken for constructs. Report: Data provenance; demographic description
2. Algorithmic Processes (Model Layer)	Convert signals into scores, classifications, or risk estimates	ML classifiers/regressors; CAT engines; NLP pipelines; deep learning; calibration tools	External validation; calibration; robustness testing; interpretability checks	Risks: Overfitting; black-box outputs; unstable performance. Report: Model type; validation split; calibration method
3. Psychometric Integrity (Measurement Layer)	Ensure outputs behave as credible psychological measurements	Reliability indicators; construct validity evidence; measurement invariance; item functioning (CAT)	Reliability (test-retest, internal consistency); convergent/discriminant validity; invariance across groups	Risks: High accuracy but poor validity; construct drift; unfair measurement. Report: Reliability/validity evidence; comparison with gold standards
4. Clinical Integration (Decision Layer)	Specify how AI outputs enter real-world care and decisions	Decision-support dashboards; triage rules; clinician override; stepped-care pathways	Clinical utility; workflow feasibility; human-AI agreement; failure modes	Risks: Automation bias; misuse as diagnosis; poor usability. Report: Intended use statement; clinician-in-the-loop design
5. Ethics & Governance (Trust Layer)	Protect rights, safety, and legitimacy of assessment	Consent models; privacy safeguards; audit trails; bias monitoring; regulatory alignment	Privacy/security; fairness monitoring; transparency; accountability and auditability	Risks: Surveillance harms; opaque decisioning; inequitable deployment. Report: Consent procedures; bias audits; oversight structure

Note. Source: Authors' synthesis based on Borsboom [9], Floridi et al. [12], Groth-Marnat and Wright [2], Insel [6], Obermeyer et al. [11], Shatte et al. [7], and Torous et al. [8].

4.2 Four Core Assertions of the Framework

The framework rests on four key propositions. First, assessment begins with construct-valid

signals, not algorithms. Passively collected digital data are useful, but they constitute proxy indicators and must be validated as

pointing towards psychological constructs rather than being treated as self-evident markers of mental state [6,9]. Second, while algorithms convert signals into outputs, psychometric integrity justifies those outputs. Many tools report high accuracy; fewer demonstrate reliability, construct validity, or cross-group comparability, yet these properties are necessary for outputs to function as assessments rather than merely as predictions [2]. Third, clinical integration is a design matter, not an afterthought. The framework treats AI as decision support and requires clear intended-use statements, clinician-in-the-loop pathways, and attention to workflow constraints so as to mitigate automation bias [8]. Fourth, ethics and governance apply throughout the full lifecycle.

Sources of bias and privacy risk are embedded in data selection, labelling, and model optimisation, making continuous audits and oversight structural requirements rather than optional additions [12,11].

4.3 Minimum Reporting Checklist

To operationalise the integrative framework and promote reporting consistency across the field, Table 4 presents a minimum reporting checklist for AI-supported mental health assessment studies. The checklist is aligned with the PRISMA 2020, TRIPOD-AI, and CONSORT-AI reporting expectations and is intended to assist authors, reviewers, and journal editors in adopting more transparent, reproducible, and accountable research practices.

Table 4. Minimum Reporting Checklist for AI-Supported Mental Health Assessment Studies

Domain	Item	What Authors Should Report	Why It Matters
A. Study Purpose & Intended Use	A1	Clear statement of whether system is intended for screening, diagnostic support, monitoring, triage, or research	Prevents misuse of AI outputs as diagnoses; clarifies scope of claims
	A2	Psychological constructs assessed with theoretical justification	Ensures construct validity and interpretability
B. Data Sources & Sampling	B1	Source of data, collection context, and time frame	Enables assessment of bias, relevance, and generalizability
	B2	Sample size, demographics, inclusion/exclusion criteria, and setting	Required for evaluating representativeness and fairness
	B3	Extent of missingness and methods used to address it	Missing data can bias AI models and psychometric estimates
C. Algorithmic Model	C1	Algorithm class (e.g., regression, random forest, neural network, CAT engine)	Allows replication and methodological scrutiny
	C2	Feature selection, preprocessing, normalization methods	Prevents hidden data leakage and inflated performance estimates
	C3	Dataset split, cross-validation, and external validation procedures	Guards against overfitting and optimistic estimates

Domain	Item	What Authors Should Report	Why It Matters
D. Psychometric Integrity	D1	Test-retest reliability, internal consistency, or stability metrics	Accuracy alone does not establish measurement quality
	D2	Construct, convergent, discriminant, and criterion validity evidence	Anchors AI outputs in psychological measurement theory
	D3	Performance and score comparability across demographic groups	Detects and mitigates algorithmic bias
E. Clinical Integration	E1	Whether AI informs, supports, or automates decisions	Clarifies accountability and reduces automation bias
	E2	Clinician oversight mechanisms and override options	Maintains professional judgment and patient safety
F. Ethics & Governance	F1	Consent processes and user understanding of AI involvement	Supports autonomy and ethical compliance
	F2	Data protection measures and regulatory compliance	Mental health data are highly sensitive and require robust protection
	F3	Ongoing bias audits and corrective strategies	Bias can emerge or worsen post-deployment
G. Limitations	G1	Methodological, data-related, or contextual constraints	Prevents overgeneralization of findings
	G2	Expected boundaries of validity across cultures and age groups	Supports responsible scaling and transfer to new populations

Note. Checklist aligns with PRISMA 2020, TRIPOD-AI, and CONSORT-AI reporting expectations.

5. Discussion

This systematic review has examined the rapidly growing literature on AI-assisted mental health assessment and synthesised findings within an integrative framework that foregrounds psychometric integrity, clinical integration, and ethical governance. Across the 68 included studies, AI technologies demonstrated substantial potential to enhance the efficiency, scalability, and personalisation of mental health assessment. At the same time, the findings reveal a persistent imbalance between technological innovation and measurement rigour, with

most tools prioritising predictive performance over the psychometric foundations on which clinical assessment depends.

One of the central lessons from the review is that high classification accuracy does not equal high-quality psychological measurement. Many AI-assisted instruments reported strong predictive statistics while providing limited evidence of reliability, construct validity, or measurement invariance. This reflects a broader epistemic tension between data-driven modelling and theory-based measurement science. Without explicit grounding in psychological constructs,

AI outputs risk functioning as opaque risk signals rather than interpretable assessment scores capable of informing clinical reasoning [9,10,23].

The review also documents the expanding range of assessment modalities made possible by AI, from adaptive testing and natural language analysis to passive digital phenotyping. These approaches challenge conventional assumptions about when, where, and how assessment takes place. In particular, continuous and passive data collection opens new possibilities for early detection and relapse monitoring, while simultaneously raising concerns about construct drift, contextual validity, and ethical surveillance. The variability of findings in the digital phenotyping literature underlines the need to ground new data streams in established psychological theory rather than assuming that behavioural proxies are direct expressions of mental states [6].

Clinical integration was identified as an important but underdeveloped dimension. Although many studies positioned AI systems as decision-support tools, few specified how outputs would be interpreted and acted upon within real clinical workflows. The absence of clear intended-use statements increases the risk of automation bias, out-of-context application, or misuse beyond validated settings [8]. These risks are compounded in resource-constrained environments where the global average of 13 mental health workers per 100,000 population means that AI tools may be operated by practitioners with limited training in psychometrics or clinical decision-making [5].

Ethical and governance concerns permeate all stages of AI-assisted assessment, from data collection through to deployment. Most of the

reviewed literature acknowledged ethical issues without operationalising or empirically investigating them. Bias audits, transparency mechanisms, and post-deployment monitoring were rarely reported. This is of particular concern in mental health contexts, where assessment outcomes can carry significant implications for access to care, stigma exposure, and legal or occupational consequences [12,11].

The underrepresentation of LMICs in the included literature raises a serious equity concern. Only 14 of 68 studies (20.6%) drew on LMIC data, despite these settings bearing a disproportionate share of unmet mental health need. AI assessment tools developed primarily on data from high-income, Western, educated, industrialised, and democratic (WEIRD) populations cannot be assumed to generalise fairly to other cultural and socioeconomic contexts [4,5]. Addressing this gap requires deliberate investment in locally adapted instruments, representative training datasets, and community-grounded implementation models.

6. Strengths and Limitations

A principal strength of this review is its explicit focus on assessment as a measurement enterprise rather than a prediction or treatment outcome exercise. By foregrounding psychometric integrity and measurement theory, the review addresses a meaningful gap in the digital mental health literature. The use of PRISMA-guided procedures, coverage of four major databases, and systematic thematic synthesis strengthen the rigour and transparency of the findings.

Several limitations must be acknowledged. The review was restricted to English-language publications, which may have excluded

relevant studies conducted in non-English-speaking settings. The heterogeneity of study designs, AI methodologies, and outcome measures precluded quantitative meta-analysis, necessitating narrative synthesis. Reporting quality varied considerably across included studies, making it difficult to draw firm conclusions about psychometric strength or clinical applicability in specific contexts. This implies that, the proposed framework, while grounded in reviewed evidence and established theory, is a conceptual model that requires empirical validation and ongoing refinement through real-world application.

7. Implications and Future Research

Directions

The findings of this review point to several priorities for future research. First, there is a clear need for tighter integration of AI development with psychometric science, with consistent reporting of reliability, validity, and measurement invariance as standard requirements. Tools unable to demonstrate these properties should be regarded as exploratory or supplementary rather than clinically actionable. Second, future research should prioritise external validation and cross-cultural assessment, particularly in LMICs where AI-assisted measurement holds the greatest potential benefit. Third, the field needs to move beyond proof-of-concept studies towards implementation research that examines usability, clinician trust, workflow integration, and longitudinal outcomes. Fourth, field-wide adoption of reporting standards, such as the minimum reporting checklist proposed in this review, would represent a significant advance in transparency, reproducibility, and accountability. Encouraging journals and funding bodies to adopt such standards would

accelerate the development of AI-assisted mental health assessment as a scientifically credible and ethically accountable field.

8. Conclusion

AI-assisted mental health assessment represents a significant development in the measurement, monitoring, and understanding of psychological functioning. The synthesised evidence in this review indicates that while AI technologies hold considerable promise, their value depends fundamentally on alignment with psychometric principles, thoughtful clinical integration, and sound ethical governance. The severity of the global mental health challenge, with over one billion people affected and costs projected at US\$6 trillion annually by 2030, with the majority of those in low-income settings receiving minimal or no professional care [5], makes scalable and valid assessment tools a matter of genuine urgency. This review aims to shift the field from fragmented innovation towards a more coherent and responsible assessment paradigm by proposing an integrative framework and minimum reporting checklist. Sustained collaboration among psychologists, clinicians, data scientists, and policymakers will be essential to ensure that AI-assisted assessment strengthens rather than undermines the scientific and ethical foundations of mental health care.

References

1. Oladunmoye EO, Enamudu GP, Sa'ad MT, Oyedele LO, Nakalema F. Role of artificial intelligence in advancing mental health counselling: A critical review. *Int J Acad Appl Res.* 2025;9(2):11-18.
2. Groth-Marnat G, Wright AJ. *Handbook of psychological assessment.* 6th ed. Wiley; 2016.

3. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspect Psychol Sci.* 2011;6(1):21-37.
4. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. *Lancet.* 2018;392(10157):1553-1598.
5. World Health Organization. World mental health today: Latest data. WHO; 2025. Available from: <https://www.who.int/publications/i/item/9789240113817>
6. Insel TR. Digital phenotyping: Technology for a new science of behavior. *JAMA.* 2017;318(13):1215-1216.
7. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: A scoping review of methods and applications. *Psychol Med.* 2019;49(9):1426-1448.
8. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry.* 2021;20(3):318-335.
9. Borsboom D. The attack of the psychometricians. *Psychometrika.* 2006;71(3):425-440.
10. Flake JK, Fried EI. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Adv Methods Pract Psychol Sci.* 2020;3(4):456-465.
11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
12. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People: An ethical framework for a good AI society. *Minds Mach.* 2018;28(4):689-707.
13. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot). *JMIR Ment Health.* 2017;4(2):e19.
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71.
15. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth.* 2020;18(10):2119-2126.
16. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme. ESRC Methods Programme; 2006.
17. Cruz-Gonzalez M, Navarro P, Fernandez-Alvarez J, Molinari G. Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychol Med.* 2025;55. <https://doi.org/10.1017/S0033291725000618>
18. Ali M, Ali S, Abbas Q, Abbas Z, Lee SW. Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *Digit Health.* 2025;11. <https://doi.org/10.1177/20552076251395548>
19. Abd-Alrazaq A, AlSaad R, Shuweihdi F, Ahmed A, Aziz S, Sheikh J. The performance of artificial intelligence-driven technologies in diagnosing mental disorders: An umbrella review. *NPJ Digit Med.* 2022;5(1):1-14.
20. Jackson R, et al. NLP-based assessment of severe mental illness from electronic health records. [As cited in Ali M et al., 2025].
22. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2023 (GBD 2023). IHME; 2025.
23. Oladunmoye EO, Oyedele LO. Artificial intelligence in psychotherapy implementation:

- Bridging technology and practice. ISAR J Med Pharm Sci. 2025;3(5):1-8.
24. Oladunmoye EO, Oboma YI, Enamudu GP. Integrating psychotherapy in medical diagnosis and care: A review of the contemporary landscape. Int J Acad Multidiscip Res. 2024;8(11):283-290.