Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ethical Challenges in the Use of Artificial Intelligence for Psychiatric Patient Management: A Systematic Review

Walaa Badawy

wlaabadawy320@yahoo.com

Menoufia University

abeer Saleh Alshehri

King Khalid University

Article

Keywords: artificial intelligence, psychiatry, mental health, ethics, confidentiality, privacy, surveillance, algorithmic bias, therapeutic relationship, accountability, digital mental health, systematic review

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11341749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ethical Challenges in the Use of Artificial Intelligence for Psychiatric Patient Management: A Systematic Review

Abstract

This systematic review examines ethical challenges arising from the use of artificial intelligence (AI) in psychiatric patient management. Six databases (PubMed, Embase, Web of Science, Scopus, PsycINFO and CINAHL) were searched to 1 December 2025. Fourteen empirical studies published between 2020 and 2025 (ten qualitative, two cross-sectional survey and two mixed-methods studies) were appraised with the Mixed Methods Appraisal Tool and synthesized thematically. Across stakeholders—patients, clinicians, young people, and experts—AI was viewed as promising for improving access, efficiency, and personalization of care, yet raised concerns about privacy, surveillance, bias, loss of empathy, and unclear accountability. Participants described the ethical acceptability of AI as conditional on strong data protection, transparency, human oversight, and equity-focused implementation. Because the evidence base is small, largely perception-based, drawn mainly from high-income countries and often concerned with hypothetical tools, these findings should be regarded as exploratory; they nevertheless point to the need for human-centred, ethically governed AI integration in mental healthcare.

Keywords

artificial intelligence, psychiatry, mental health, ethics, confidentiality, privacy, surveillance, algorithmic bias, therapeutic relationship, accountability, digital mental health, systematic review

Introduction:

Mental disorders represent a major and growing global health challenge. Recent estimates from the World Health Organization indicate that more than one billion people worldwide are living with a mental disorder, with anxiety and depressive disorders being among the most prevalent (Limenih et al., 2024). Mental disorders account for a substantial proportion of years lived with disability and are associated with high social and economic costs across the life course (Kieling et al., 2024). Despite the availability of effective pharmacological and psychosocial interventions, a substantial treatment gap persists: in many countries, most individuals who require mental health care receive no or inadequate treatment (Paudel et al., 2025; Wainberg et al., 2017). Shortages in specialist workforce, limited service capacity, and structural barriers to access have prompted interest in technological innovations that could extend the reach and efficiency of psychiatric services (Berardi et al., 2024).

Artificial intelligence (AI) has emerged as a central component of this digital transformation. In mental health care, AI techniques, including machine learning, natural language processing, and deep learning, are being applied to a wide range of tasks: diagnostic support, prognosis and relapse prediction, suicide risk stratification, personalized treatment selection, monitoring of symptoms via smartphones and wearables, and delivery of conversational agents or chatbots for psychoeducation and low-intensity support (Ali et al., 2025). In psychiatry specifically, AI-based tools have been developed to predict suicide attempts from electronic health records and social media signals, to detect early psychosis, and to estimate treatment response trajectories (D'Hotman & Loh, 2020). Parallel advances in large language models and conversational AI have led to the proliferation of mental health–oriented chatbots and general-purpose systems that are increasingly used by patients and clinicians (Bucher et al., 2025; Feng et al., 2025). These innovations raise

hopes for more scalable, data-driven, and personalized psychiatric care, particularly in settings constrained by workforce shortages and service inequities (Ajegbile et al., 2024).

However, the integration of AI into psychiatric patient management also generates complex ethical challenges. Mental health data are among the most sensitive categories of health information, encompassing highly stigmatized diagnoses, histories of self-harm or trauma, and fine-grained behavioural traces from smartphones and online activity. Misuse or inadequate protection of such data may exacerbate stigma, discrimination, and human-rights violations (Conway & O'Connor, 2016; Grando et al., 2020; Parenteau et al., 2023). Moreover, many AI systems operate as opaque "black boxes," making it difficult for clinicians and patients to understand how predictions are generated, to contest decisions, or to identify errors (Xu & Shuttleworth, 2024). Concerns have been raised about algorithmic bias and inequity, particularly where training data under-represent certain diagnostic groups, age categories, or racial and cultural minorities, risking systematic misclassification or unequal access to benefits (Joseph, 2025). In the psychiatric context, where diagnoses are often probabilistic, social determinants are central, and power differentials between patients and services are pronounced, these issues have profound implications for autonomy, justice, trust, and safety (Kirkbride et al., 2024).

Policy bodies and professional organizations have responded by articulating high-level principles for trustworthy and responsible AI in health. The World Health Organization's guidance on the ethics and governance of AI for health emphasizes principles such as autonomy, beneficence, non-maleficence, justice, and explainability, and highlights specific risks when AI is deployed in vulnerable populations (Pham, 2025). The OECD's frameworks on trustworthy AI and responsible AI in health similarly stress human-centred values, transparency, robustness, accountability, and inclusive governance (Dhaigude & Kamath, 2025). National initiatives, including AI ethics

principles issued by health and data authorities in several countries, echo these concerns and call for context-sensitive regulation (Mennella et al., 2024). At the same time, a growing body of reviews and conceptual work has catalogued ethical issues in AI for healthcare generally, such as privacy, consent, bias, opacity, and liability, and proposed frameworks or checklists to support ethical design and deployment (Gorelik et al., 2025). Yet these contributions rarely focus in depth on psychiatric care, where therapeutic relationships, coercion histories, capacity fluctuations, and crisis management introduce distinctive ethical tensions.

Therefore, this systematic review aims to identify, categorize, and critically synthesize the ethical challenges associated with the use of AI in psychiatric patient management, encompassing assessment, diagnosis, risk prediction, treatment planning, intervention delivery, monitoring, and follow-up. By mapping these challenges to existing ethical frameworks and highlighting gaps in current safeguards, the review seeks to inform clinicians, developers, policymakers, and educators working toward the responsible and ethically robust integration of AI into psychiatric care.

Methods

Search Strategy and Selection Criteria

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021) to ensure methodological transparency, replicability, and scientific rigor. The review protocol was not prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO); this is acknowledged as a limitation, and all review materials are provided as supplementary files (Supplementary Files S1–S6) to support transparency and verification. The review aimed to synthesize global evidence on ethical issues, risks, and considerations associated with the use of artificial intelligence (AI) in psychiatric patient management, including assessment, diagnosis, risk stratification, treatment planning, intervention delivery, and clinical monitoring.

A comprehensive and systematic literature search was conducted across six major electronic databases: PubMed, Embase, Web of Science, Scopus, PsycINFO, and CINAHL (via EBSCOhost). The final search was completed on December 1, 2025. The search strategy was developed in consultation with an academic health sciences librarian and informed by a preliminary scoping search to identify key terminology used in the AI ethics and digital psychiatry literature.

Boolean operators (AND, OR) were used to combine search terms within and across conceptual blocks. Searches were limited to peer-reviewed journal articles, published in English, and involving human participants or clinical psychiatric use cases. No date limits were applied in the database searches; the publication-date criterion (January 2010 onwards) was applied during screening. To supplement database searches, the reference lists of all included studies and relevant

systematic or scoping reviews were manually screened to identify additional eligible publications. A detailed overview of search terms used in each database is presented in Table 1, and the complete database-specific search strategies, including platforms, search dates, limits and yields, are provided in Supplementary File S3.

Table 1: Search Strategy Overview by Database

Database	**Search Terms**
PubMed	(“artificial intelligence” OR “machine learning” OR “deep learning” OR “algorithmic decision-making” OR “large language model” OR “chatbot”) AND (“psychiatry” OR “mental health” OR “suicide risk prediction” OR “psychiatric diagnosis” OR “psychotherapy support”) AND (“ethics” OR “ethical issues” OR “privacy” OR “bias” OR “accountability” OR “transparency” OR “autonomy”)
Embase	(‘artificial intelligence’/exp OR ‘machine learning’/exp OR “deep learning”) AND (‘mental disease’/exp OR ‘psychiatric assessment’/exp OR “risk prediction”) AND (‘ethics’/exp OR “data privacy” OR “algorithmic bias” OR “informed consent”)
CINAHL	(“artificial intelligence” OR “algorithmic tool” OR “clinical decision support”) AND (“mental health services” OR “psychiatric nursing” OR “behavioral health”) AND (“ethical considerations” OR “privacy” OR “autonomy” OR “safety” OR “fairness”)

PsycINFO	(“machine learning” OR “AI-driven”) AND (“psychiatric diagnosis” OR “mental health management” OR “psychological assessment”) AND (“ethical challenges” OR “bias” OR “justice” OR “transparency” OR “informed consent”)
Scopus	ALL(“artificial intelligence” AND “psychiatric” AND (“ethics” OR “privacy” OR “bias” OR “fairness” OR “autonomy”))
Web of Science	(“AI” OR “machine learning” OR “deep learning” OR “chatbot”) AND (“psychiatric care” OR “mental health management”) AND (“ethical issues” OR “responsibility” OR “accountability” OR “equity”)

Eligibility Criteria for Screening

Following the removal of duplicate records, titles and abstracts were screened for initial relevance. Full-text articles that passed this stage were assessed in detail using pre-established inclusion and exclusion criteria. The aim was to identify studies that explored or evaluated the ethical implications, challenges, or concerns associated with the use of artificial intelligence (AI) tools in psychiatric or mental health patient management.

Included studies met the following criteria:

(1) original empirical research (quantitative, qualitative, or mixed-methods designs) reporting primary data;

(2) focused on the use of AI-driven tools, algorithms, or systems in the context of psychiatric assessment, diagnosis, treatment, decision-making, monitoring, or follow-up;

(3) explicitly addressed ethical issues, such as privacy, consent, transparency, bias, autonomy, accountability, safety, or therapeutic relationships;

(4) involved clinical, community, or digital psychiatric settings or populations (e.g., patients, providers, systems);

(5) published in peer-reviewed journals in English between January 2010 and 1 December 2025 (the date of the final search).

Mixed-methods studies were included if their findings contributed meaningfully to understanding ethical challenges in the implementation or use of AI in psychiatry.

We excluded studies that:

(1) focused on AI applications in general medical or non-psychiatric populations without reference to mental health or psychiatric care;

(2) lacked discussion or analysis of ethical dimensions (e.g., purely technical performance evaluations);

(3) presented broad overviews of AI ethics in healthcare without specific relevance to psychiatric patient management;

(4) were editorials, opinion pieces, letters, protocols, conference abstracts, reviews, or theoretical or conceptual papers without primary empirical data;

(5) were dissertations, theses, or non-peer-reviewed grey literature.

Data Extraction Process

The data extraction process was conducted with methodological rigor to ensure consistency, transparency, and relevance in capturing ethical challenges reported across the included studies. Two independent reviewers extracted data from all eligible studies using a pre-designed, standardized data extraction form that was developed in alignment with the review's objectives.

Any discrepancies in data interpretation or coding were resolved through discussion, and a third reviewer was consulted when consensus could not be achieved. The blank extraction form and the completed extraction and synthesis tables are provided in Supplementary Files S5 and S6, respectively.

The following key domains were extracted from each included study:

- **Study Characteristics**: We recorded bibliographic information (first author, year of publication), study location or geographical scope, study design (e.g., conceptual, empirical, qualitative, quantitative, or mixed-methods), sample size (if applicable), and setting (e.g., inpatient psychiatry, community mental health, digital platforms). This information supported the contextualization of ethical issues across diverse psychiatric and technological environments.

- **AI Application Type**: For each study, we extracted the type of AI system discussed or evaluated, including machine learning models, diagnostic decision-support tools, suicide risk prediction algorithms, conversational agents (e.g., chatbots), digital phenotyping systems, or large language model–based interventions. We also noted whether the AI tool was hypothetical, under development, or in clinical use.

- **Mental Health Focus**: We recorded the specific psychiatric condition(s) or mental health domain addressed (e.g., depression, psychosis, suicidality, generalized anxiety, substance use disorder), as well as the clinical purpose (e.g., assessment, diagnosis, treatment planning, monitoring, triage, crisis management).

- **Ethical Challenges Identified**: A central focus of data extraction was the documentation of ethical issues raised in each study. These were thematically coded and grouped under

higher-order categories, including but not limited to: privacy and confidentiality, informed consent, algorithmic bias and fairness, transparency and explainability, autonomy, accountability, safety, therapeutic relationship, and stigma. We also noted whether these challenges were discussed in relation to specific patient populations (e.g., minors, marginalized groups, involuntarily hospitalized patients).

- **Theoretical or Normative Frameworks**: When present, we extracted any ethical theories, principles, or frameworks used by authors to analyze or frame the ethical issues (e.g., principlism, human rights-based approaches, WHO AI ethics principles, deontological or utilitarian perspectives).

- **Proposed Ethical Safeguards or Mitigation Strategies**: Studies were reviewed for any recommendations, guidelines, or practical safeguards proposed to address the identified ethical concerns. These included strategies such as human-in-the-loop models, explainable AI methods, clinician education, audit mechanisms, transparency protocols, co-design with service users, or regulatory oversight models.

- **Stakeholder Perspectives**: For empirical studies, we extracted information about the populations whose perspectives were captured (e.g., clinicians, developers, patients, ethicists), and any insights into how ethical challenges were perceived or negotiated in practice. Qualitative and mixed-methods findings were analyzed to understand experiential or contextual dimensions of ethical risks and decision-making.

- **Equity and Sociocultural Considerations**: We documented any discussion of disparities in access, risks of exclusion or harm to minority or vulnerable groups, digital divides, or

culturally-specific ethical tensions. These factors were flagged for use in subgroup interpretation and to support context-sensitive recommendations.

- **Regulatory and Policy Relevance**: We noted whether the studies engaged with existing laws, policies, or regulatory frameworks concerning data protection, mental health legislation, or AI governance. Author-stated implications for clinical guidelines, health system governance, or AI ethics regulation were also recorded.
- **Methodological Quality Indicators**: For empirical studies, we extracted data on methodological features such as sample recruitment, ethical approval, researcher reflexivity, data analysis rigor, and transparency of reporting.

In instances where data were incomplete, unclear, or missing, attempts were made to contact the corresponding authors for clarification. All extracted data were reviewed jointly by both reviewers to ensure completeness, coherence, and relevance to the review aims. This systematic and multi-dimensional approach enabled the identification of recurring ethical concerns and supported a comprehensive synthesis of how AI intersects with core ethical principles in psychiatric care.

Quality Assessment

Methodological quality was appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018 (Hong et al., 2018), which was designed for systematic reviews that include qualitative, quantitative and mixed-methods studies. After the two screening questions, each study was appraised against the five criteria for its design category (qualitative; quantitative descriptive; or mixed methods, for which the qualitative and quantitative components were also appraised), and each criterion was rated ‘Yes’, ‘No’ or ‘Can’t tell’. Disagreements were resolved by discussion. In line with MMAT guidance, no overall score was calculated and no study was excluded on the basis

of quality; instead, judgements are reported for each criterion (Figure 2; Supplementary File S6), and the number of criteria met is reported descriptively and was considered when interpreting the synthesis.

Data Analysis

Because all included studies were empirical and most were qualitative, findings were synthesized using thematic synthesis (Thomas & Harden, 2008), supported by a structured narrative summary of study characteristics and survey findings. This approach allowed a comprehensive exploration of the ethical challenges, tensions, and mitigation strategies associated with the use of artificial intelligence (AI) in psychiatric patient management.

1. Narrative Summary

A structured narrative synthesis was undertaken to summarize findings from the included studies, describing ethical dimensions of AI use in psychiatry. These studies varied widely in focus, ranging from diagnostic support systems and suicide risk prediction tools to chatbot-based therapy and digital phenotyping platforms. Key features extracted included the type of AI application, the psychiatric domain addressed, the ethical issues discussed, and any proposed solutions or frameworks.

2. Thematic Synthesis

For qualitative and mixed-methods studies, we applied inductive thematic analysis to identify patterns in the lived experiences, professional reflections, and institutional perspectives on ethical challenges in AI-enabled psychiatric care. Thematic coding was conducted iteratively, beginning with open coding of text segments relevant to ethical concerns, followed by the clustering of related codes into higher-order themes. Findings from the survey and mixed-methods studies were

integrated narratively within the relevant themes. Confidence in the synthesized findings was not formally assessed (for example, with GRADE-CERQual).

Results

Search results

The database search yielded a total of 612 records, with no additional records identified through trial or study registers. Following the removal of 112 duplicates and 10 records excluded for administrative reasons, 490 records proceeded to title and abstract screening. Of these, 430 records were excluded for not meeting the inclusion criteria, leaving 60 full-text reports sought for retrieval. Five reports could not be accessed despite repeated attempts, resulting in 55 full-text articles assessed for eligibility. After detailed evaluation, 41 reports were excluded, including 21 that were non-empirical or conceptual in nature, 14 that examined artificial intelligence applications outside psychiatric or mental health contexts, and 6 that addressed AI technologies but did not include ethical perspectives or primary stakeholder data. Ultimately, 14 studies met all eligibility requirements and were included in the final synthesis (AlMaskari et al., 2025; Blease et al., 2020; Buhr et al., 2025; Cross et al., 2024; Ding & Barbic, 2025; Götzl et al., 2022; Hipgrave et al., 2025; Kim et al., 2023; Lee et al., 2024; Petersson et al., 2025; Sharif et al., 2025; Starke et al., 2023; Stroud et al., 2025; Zhang et al., 2023). Each of the included studies corresponded to a

unique report, yielding a total of 14 full-text reports included in this systematic review (Figure 1). The complete search strategies are provided in Supplementary File S3, the screening and full-text eligibility log (including reasons for exclusion) in Supplementary File S4, and the characteristics of the included studies in Table 2.

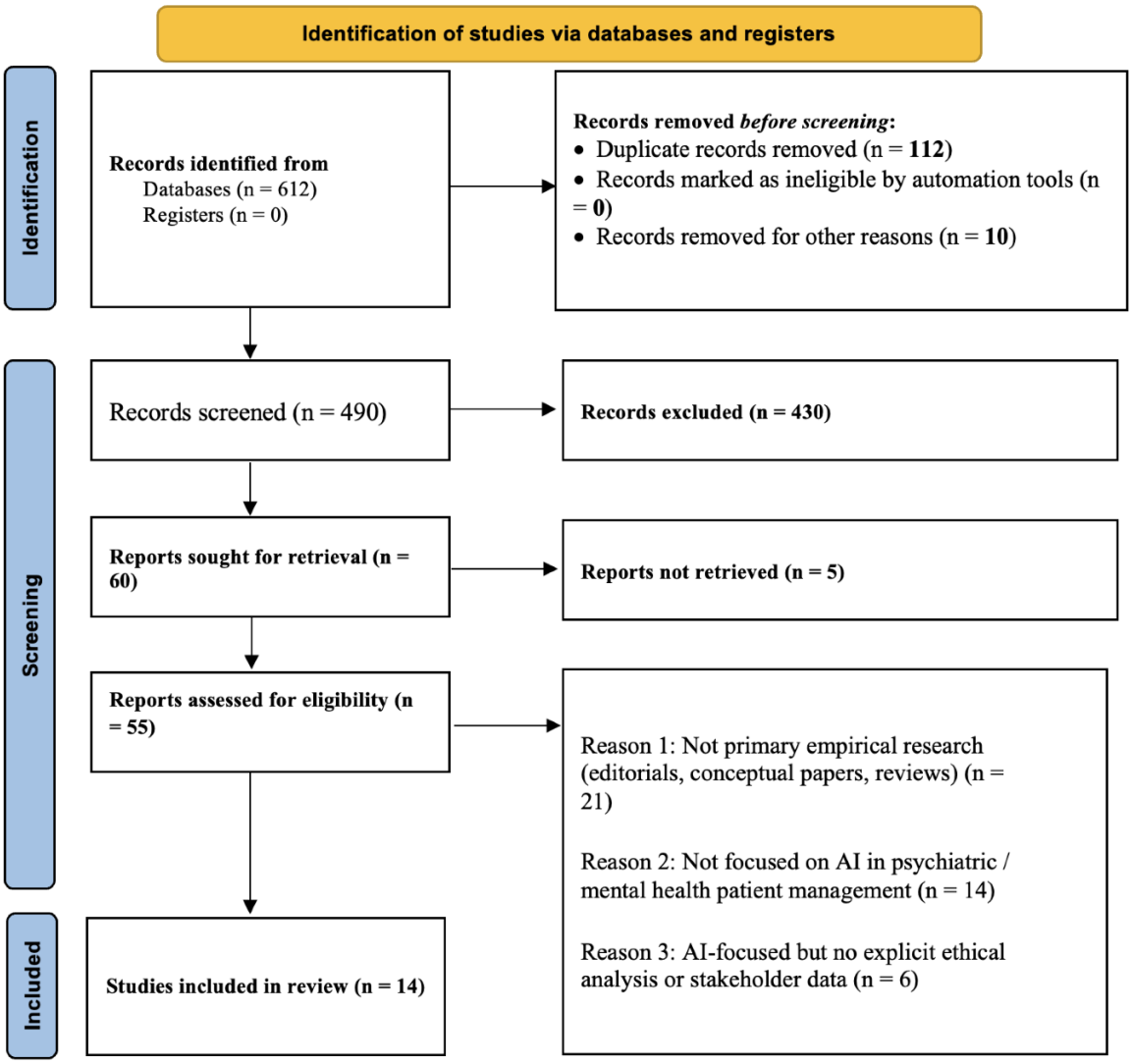


Figure 1: PRISMA Flow Diagram

Methodological Quality

Criterion-level MMAT judgements are summarized in Figure 2, and the supporting justifications are provided in Supplementary File S6. All 14 studies met both MMAT screening questions. The

ten qualitative studies generally performed well: nine met all five qualitative criteria, whereas Blease et al. (2020) did not meet criterion 1.2 because its data consisted of brief written responses to open-ended survey questions that could not be probed. The two quantitative descriptive surveys were weaker: Sharif et al. (2025) met three of five criteria and Cross et al. (2024) one of five, chiefly because of self-selected online samples of uncertain representativeness, unreported response rates and, for Cross et al. (2024), partly ad hoc measures. Both mixed-methods studies met three of five mixed-methods criteria: Götzl et al. (2022) did not explicitly address divergences between its qualitative and quantitative strands, Hipgrave et al. (2025) provided no rationale for combining methods, and both had weaknesses within their individual components (small qualitative samples and largely unvalidated survey items). Across designs, recurrent limitations included convenience or self-selected samples, single-site or single-country recruitment, a focus on hypothetical rather than implemented AI tools, and inconsistent reporting of data saturation. Because the MMAT appraises methodological quality rather than the credibility of each synthesized finding, findings that rested mainly on the lower-rated survey and mixed-methods studies were interpreted with greater caution.

Study (MMAT category)		S1	S2	C1	C2	C3	C4	C5	Criteria met
AlMaskari et al., 2025	QUAL	+	+	+	+	+	+	+	5/5
Blease et al., 2020	QUAL	+	+	+	×	+	+	+	4/5
Buhr et al., 2025	QUAL	+	+	+	+	+	+	+	5/5
Cross et al., 2024	QUAN-D	+	+	×	×	?	×	+	1/5
Ding & Barbic, 2025	QUAL	+	+	+	+	+	+	+	5/5
Götzl et al., 2022	MM	+	+	+	+	+	?	×	3/5
Hipgrave et al., 2025	MM	+	+	×	+	+	+	×	3/5
Kim et al., 2023	QUAL	+	+	+	+	+	+	+	5/5
Lee et al., 2024	QUAL	+	+	+	+	+	+	+	5/5
Petersson et al., 2025	QUAL	+	+	+	+	+	+	+	5/5
Sharif et al., 2025	QUAN-D	+	+	+	×	+	×	+	3/5
Starke et al., 2023	QUAL	+	+	+	+	+	+	+	5/5
Stroud et al., 2025	QUAL	+	+	+	+	+	+	+	5/5
Zhang et al., 2023	QUAL	+	+	+	+	+	+	+	5/5

S1–S2: screening questions (clear research questions; data allow questions to be addressed).

C1–C5: the five MMAT criteria for the study's category — QUAL = 1.1–1.5 (qualitative); QUAN-D = 4.1–4.5 (quantitative descriptive); MM = 5.1–5.5 (mixed methods). Criteria met = number of 'Yes' ratings for C1–C5.

+ Yes × No ? Can't tell

Figure 2: Methodological quality of the included studies appraised with the Mixed Methods Appraisal Tool (MMAT, version 2018). S1–S2, screening questions; C1–C5, the five criteria for each study's design category (qualitative 1.1–1.5; quantitative descriptive 4.1–4.5; mixed methods 5.1–5.5).

Main Outcomes

This systematic review synthesized findings from 14 empirical studies (ten qualitative, two cross-sectional survey and two mixed-methods studies) examining how different stakeholders, psychiatrists, other mental health professionals, young people, patients with lived experience, and interdisciplinary experts, perceive the ethical challenges of using artificial intelligence in psychiatric and mental health care (Blease et al., 2020; Stroud et al., 2025; Zhang et al., 2023; Ding & Barbic, 2025; Petersson et al., 2025; AlMaskari et al., 2025; Götzl et al., 2022; Kim et al., 2023; Hipgrave et al., 2025; Cross et al., 2024; Lee et al., 2024; Sharif et al., 2025; Buhr et al., 2025; Starke et al., 2023). The studies covered diverse AI applications, including decision-support systems, predictive models, chatbots, and AI-informed mobile apps, across settings in North America, Europe, the Middle East, and Oceania. A narrative and thematic synthesis of the extracted data revealed five overarching outcome domains:

1. Perceived Benefits and Added Value of AI in Mental Health Care

Across studies, participants consistently identified tangible benefits of AI for psychiatric service delivery. Clinicians and experts emphasized that AI-driven decision-support and predictive tools could enhance diagnostic accuracy, treatment optimization, and early detection of relapse or suicide risk, particularly when integrated into routine workflows (Stroud et al., 2025; Kim et al.,

2023; Starke et al., 2023). Mental health professionals also highlighted the potential of AI to reduce administrative burden and streamline documentation, triage, and information retrieval, thereby freeing time for direct clinical work (Zhang et al., 2023; Hipgrave et al., 2025).

From a service-user perspective, young adults, university students, and community members described AI chatbots and mobile apps as attractive because they offer 24/7 availability, anonymity, and low-barrier access, which may help overcome stigma and long waiting times (Petersson et al., 2025; AlMaskari et al., 2025; Götzl et al., 2022; Lee et al., 2024; Cross et al., 2024). Many perceived AI tools as useful adjuncts for self-monitoring, psychoeducation, and practicing coping skills between therapy sessions (Ding & Barbic, 2025; Götzl et al., 2022; Lee et al., 2024). Overall, there was broad optimism that AI could act as a supportive extension of the mental health workforce and increase the reach and personalization of care when appropriately designed and governed (Blease et al., 2020; Sharif et al., 2025; Buhr et al., 2025).

2. Privacy, Surveillance, and Data Security as Central Ethical Worries

Despite enthusiasm, privacy and data protection were among the most frequently reported concerns, raised in most studies involving patients, young people, and clinicians. Service users worried about who could access their sensitive mental health data, such as universities, employers, insurers, or family members, and whether information shared with chatbots or apps could be used against them (Petersson et al., 2025; AlMaskari et al., 2025; Lee et al., 2024; Cross et al., 2024). Young people and students expressed ambivalence about continuous AI-based monitoring: symptom tracking and digital phenotyping were seen as helpful but also as forms of surveillance that might intensify anxiety or diminish a sense of privacy (Petersson et al., 2025; Götzl et al., 2022).

Clinicians and experts raised similar issues at system level, stressing vulnerabilities of cloud-based LLMs, commercial vendors, and cross-border data flows, and questioning whether current consent processes adequately explain complex data practices (Stroud et al., 2025; Hipgrave et al., 2025; Starke et al., 2023). In Oman, students also linked data protection to concerns about the cultural and religious appropriateness of chatbot advice, underscoring the need for locally appropriate safeguards (AlMaskari et al., 2025). Together, these findings suggest that, for many participants, the acceptability of AI in psychiatry was conditional on stringent privacy, security, and transparent governance mechanisms.

3. Human Relationship, Empathy, and the Limits of Automation

A third major outcome domain concerned the impact of AI on the therapeutic relationship and human empathy. Patients and young adults generally appreciated AI chatbots as accessible sources of information and interim support, but they strongly rejected the idea of chatbots as full replacements for therapists (Petersson et al., 2025; AlMaskari et al., 2025; Götzl et al., 2022; Lee et al., 2024). Many participants described AI as lacking the emotional resonance, nuanced understanding, and moral presence needed in complex or crises (Lee et al., 2024; Ding & Barbic, 2025).

Clinicians echoed these reservations, warning that over-reliance on AI could erode the therapeutic alliance, depersonalize encounters, and encourage transactional forms of care, particularly if organizations use AI to justify reductions in face-to-face services (Stroud et al., 2025; Zhang et al., 2023; Hipgrave et al., 2025). At the same time, both clinicians and patients reported feeling more comfortable when AI tools were explicitly framed as adjunctive supports under human supervision, with clear protocols for escalation and crisis management (Ding & Barbic, 2025;

Cross et al., 2024; Lee et al., 2024). These findings indicate a widely shared expectation among participants that AI should augment rather than replace human relationships in psychiatric care.

4. Trust, Explainability, and Professional Responsibility

The review also identified pervasive uncertainty about how to trust AI systems and who is accountable when things go wrong. Clinicians expressed reluctance to rely on tools whose training data, performance metrics, or decision logic were not transparent, fearing that opaque recommendations could undermine clinical judgment and expose them to legal risk (Stroud et al., 2025; Zhang et al., 2023; Hipgrave et al., 2025). In expert interviews, psychiatrists and neuroscientists debated the ethical value of explainability, with some arguing that "explainable AI" can be a superficial fix if underlying biases and methodological flaws remain unaddressed (Starke et al., 2023).

Both physicians and machine learning researchers highlighted challenges in the early stages of model development, including how research priorities are chosen, how fairness is operationalized, and how interdisciplinary teams align on values and responsibilities (Kim et al., 2023; Buhr et al., 2025). Several studies suggested that AI literacy and prior exposure were associated with greater comfort: clinicians and students who had training or hands-on experience with AI reported higher confidence in critically appraising and safely using these tools (Blease et al., 2020; Zhang et al., 2023; Sharif et al., 2025). Collectively, these accounts suggest that trust in AI may depend not only on technical performance but also on transparent design, independent validation, clear accountability structures, and structured education for mental health professionals.

5. Equity, Bias, and Conditions for Ethical Implementation

Finally, multiple studies underscored concerns that AI could exacerbate existing inequities in mental health care. Participants worried that models trained on narrow datasets might misclassify or underserve certain diagnostic groups, cultural minorities, or socioeconomically disadvantaged populations (Stroud et al., 2025; Cross et al., 2024; Buhr et al., 2025). Youth and experts highlighted the digital divide, noting that not all patients have reliable devices, connectivity, or digital literacy to benefit from AI-enabled services (Götzl et al., 2022; Petersson et al., 2025).

Clinicians and experts argued that ethical implementation requires system-level safeguards, including participatory co-design with service users, robust regulatory oversight, ongoing evaluation of bias and unintended consequences, and integration into existing clinical pathways rather than standalone "black box" tools (Zhang et al., 2023; Ding & Barbic, 2025; Hipgrave et al., 2025; Buhr et al., 2025). Students in Oman further emphasized the importance of cultural and religious congruence, calling for local adaptation and context-sensitive governance (AlMaskari et al., 2025), while mental health professionals in Saudi Arabia highlighted limited AI training and preparedness, uncertainty about responsibility for AI errors, and concerns about professional role displacement (Sharif et al., 2025).

Overall, the main outcomes suggest that participants saw promise in AI for psychiatry but described this promise as conditional on strong privacy protections, preservation of human relationships, transparent and accountable systems, and deliberate efforts to promote fairness and inclusion in both design and deployment. Study-level characteristics and extracted findings for all included studies are presented in Table 2.

Discussion :

This systematic review synthesized evidence from fourteen qualitative, survey and mixed-methods studies exploring ethical concerns surrounding the integration of artificial intelligence (AI) into psychiatric patient management. The findings collectively highlight a landscape characterized by both optimism about AI's transformative potential and persistent apprehension about its ethical, clinical, and societal implications. These results align with the broader literature on digital mental health, which underscores that while AI may enhance efficiency, personalization, and access to care, its implementation must be guided by rigorous ethical oversight and context-sensitive governance frameworks (Olawade et al., 2024).

One of the central themes emerging from the review is the tension between AI-enabled efficiency and preservation of the therapeutic relationship. Consistent with previous research, clinicians in the included studies expressed fears that AI systems may unintentionally dehumanize mental healthcare and disrupt the empathic connection fundamental to psychiatric practice (Montemayor et al., 2022). Relational competence and therapeutic rapport are widely regarded as central to mental healthcare (Beyene et al., 2024). Thus, the prospect of AI-mediated interventions—particularly chatbots and automated triage systems—raises concerns that the relational dimension of care could be diminished if AI tools replace rather than complement clinician–patient interactions. This finding echoes calls for "augmented intelligence" models in psychiatry, wherein AI supports but never supplants human expertise (Lee et al., 2024; Stroud et al., 2025).

Another major issue identified was data privacy and digital surveillance, particularly among young adults and patients with heightened sensitivity to stigma. These concerns parallel broader ethical debates on the use of digital phenotyping, social media analytics, and continuous monitoring technologies in mental health research (Martinez-Martin et al., 2021). Psychiatric data are uniquely sensitive due to their link with vulnerability, discrimination, and social identity; thus, breaches or

misuse of such information may have disproportionate consequences compared with general medical data (Martinez-Martin et al., 2021). Recent scholarship emphasizes that informed consent procedures for AI tools must evolve beyond static forms toward dynamic, transparent, and user-centered models that clearly explain data flows, algorithmic purposes, and potential downstream uses (Hao et al., 2025; Ziv & Nakash, 2025).

Moreover, several studies raised issues surrounding algorithmic fairness and structural bias. This is reinforced by mounting evidence that machine-learning models can systematically misclassify or underserve individuals from minority racial, gender, or socioeconomic groups due to imbalanced training datasets or biased feature selection (Wang et al., 2023). In psychiatry, such inequities may exacerbate existing disparities in diagnosis, involuntary treatment, or access to services (Zabelski et al., 2024). Scholars highlight the need for diverse data sets, bias-detection audits, and culturally contextualized model development processes to mitigate inequitable outcomes (Varsha, 2023). Transparent reporting of algorithmic performance across demographic subgroups should also become standard practice.

Clinicians across the included studies expressed uncertainty about accountability, liability, and trust in AI-augmented decision-making. This concern is widely shared in the global literature: without clear regulatory frameworks specifying responsibility when AI contributes to harm, clinicians may become reluctant to adopt these tools, despite their potential benefits (Abdelwanis et al., 2026). Commentators have emphasized that clinicians remain ultimately responsible for decisions informed by AI systems, even when algorithmic recommendations appear authoritative (Zheng & Zhang, 2025). To address this tension, AI tools must be embedded within governance structures that provide clinicians with interpretability, audit trails, and mechanisms for challenging algorithmic outputs (Moreno-Sánchez et al., 2026).

The review also highlighted the critical role of AI literacy and workforce preparedness. Several included studies reported that participants with prior exposure to AI technologies held more positive attitudes, yet many still lacked confidence in evaluating algorithmic validity or incorporating outputs into complex clinical judgments. Literature on digital competency frameworks for healthcare professionals stresses the importance of training programs that encompass not only technical proficiency but also ethical reasoning and critical appraisal of AI systems (Laakkonen et al., 2024). Such competencies are essential for preventing over-reliance on algorithmic outputs and ensuring that AI augments rather than replaces clinical reasoning.

Finally, the findings emphasize that ethical AI implementation in psychiatry requires multilevel governance, including institutional policies, national regulatory standards, and participatory co-design with service users. Patient and community engagement has been shown to improve trust, transparency, and cultural congruence in digital mental health tools (Putica et al., 2025). Ethical frameworks such as the WHO Guidance on AI for Health and the OECD Principles on AI further underscore the need for fairness, explainability, safety, and accountability—elements that align closely with the concerns identified in this review (Mennella et al., 2024).

Implications of the Study

This review has several practical implications for psychiatric and mental health care. Because these implications rest on a small, perception-based evidence base, they should be regarded as provisional. First, the included studies suggest that ethical acceptability cannot be assumed: participants indicated that AI tools are more likely to be trusted when they protect privacy, are transparent about data use, and preserve a meaningful human role in care. Clinicians and service leaders should therefore treat AI systems as adjuncts rather than replacements, embedding them within models of “augmented intelligence” where final responsibility and relational work remain

with human professionals. Second, AI literacy needs to be incorporated into undergraduate, postgraduate, and continuing education for psychiatrists, nurses, psychologists, and allied staff. Training should address not only how tools operate, but also how to critically appraise data quality, detect bias, and communicate AI-generated information to patients. Third, developers and policymakers should use these findings to design governance frameworks and regulatory standards that mandate impact assessments on privacy, equity, and therapeutic relationships before deployment. Finally, the themes identified here provide a foundation for co-design with service users, ensuring that future AI systems respond to the needs and values of patients, families, and communities rather than solely to technical or organizational priorities.

Limitations of the Study

Several limitations should be considered when interpreting this review. First, the protocol was not prospectively registered, which increases the risk of undisclosed post hoc decisions; to mitigate this, the search strategies, eligibility decisions, extraction forms and appraisal judgements are provided in Supplementary Files S1–S6. Second, the search was restricted to English-language, peer-reviewed journal articles; grey literature, dissertations and conference proceedings were excluded, and five potentially eligible full-text reports could not be retrieved. Relevant evidence may therefore have been missed, and research from non-English-speaking and low- and middle-income settings is likely under-represented; all included studies were conducted in high-income countries or, in one case, an international sample of psychiatrists. Third, the search strings were relatively brief and did not exploit the full range of controlled vocabulary and synonyms in every database, so retrieval may have been incomplete. Fourth, the evidence base is small (14 studies) and dominated by qualitative and cross-sectional designs with convenience or self-selected samples; the findings are therefore vulnerable to selection and social desirability bias and cannot

be used to estimate the prevalence of attitudes or experiences in broader populations. Fifth, most studies examined perceptions of hypothetical or prospective AI tools rather than experiences of systems deployed in routine psychiatric care, so the ethical challenges identified largely reflect anticipated rather than observed harms. Sixth, the included studies were heterogeneous in AI application, population and setting, and one study (Kim et al., 2023) examined clinical machine-learning development across medicine rather than in mental health care specifically, although psychiatry was among the participating specialties; findings from that study were used only for cross-cutting issues of development ethics and accountability. Seventh, quality was appraised with the MMAT, but confidence in each synthesized finding was not formally assessed (for example, with GRADE-CERQual), and thematic synthesis involves reviewer interpretation. Finally, AI technology and its regulation are evolving rapidly; some concerns described may already be outdated, while challenges associated with large language models and multimodal systems are only beginning to be studied. The conclusions should therefore be viewed as exploratory and interpretive rather than definitive.

Conclusion

This systematic review indicates that stakeholders across psychiatry—patients, young people, clinicians, and experts—perceive potential for AI to improve access, efficiency, and personalization of mental health care, while describing important ethical conditions for its use. The conditions most frequently raised were robust protection of privacy and data security, preservation of the therapeutic relationship and human empathy, transparency and accountability in decision-making, and attention to fairness and inclusion. Because these findings derive from a small number of predominantly qualitative and cross-sectional studies of stakeholder perceptions, often concerning hypothetical tools, they should be interpreted as exploratory. They suggest that future

efforts should extend beyond technical optimization toward human-centred, ethically informed design and governance developed with service users and front-line clinicians, and that evaluations of AI tools deployed in routine psychiatric care should examine these ethical dimensions directly.

Declarations

Disclosure of Interest Statement

The authors report no conflicts of interest.

Ethical Approval Statement

This study is a systematic review of previously published literature and did not involve the collection of new data from human participants, patients, or identifiable clinical material. Therefore, ethical approval was not required, in accordance with institutional and journal guidelines. All reviewed studies were assumed to have obtained appropriate ethical approval in their original published forms.

Patient Consent Statement

This manuscript does not include any original clinical material, patient data, or case descriptions. As the study synthesizes findings from pre-existing publications and does not involve direct participation of patients or service users, individual patient consent was not required.

Funding Declaration

The authors received no funding for this work.

Author Contributions

Both authors contributed to the conception and design of the review, study selection, data extraction, analysis and interpretation of the findings, and preparation and critical revision of the manuscript. Both authors reviewed and approved the final manuscript.

Data Availability Statement: All materials underpinning this systematic review are provided as Supplementary Information and are available to editors, referees and readers without restriction: the completed PRISMA 2020 checklist (Supplementary File S1); the PRISMA 2020 flow diagram (Supplementary File S2); the full search strategies for all six databases, including platforms, search dates, limits and per-database yields (Supplementary File S3); the screening and eligibility log, including the eligibility criteria applied at each stage, the full-text assessment log with reasons for exclusion and the reports that could not be retrieved (Supplementary File S4); the blank data extraction form (Supplementary File S5); and the completed data extraction and synthesis tables, including study characteristics, the study-by-theme synthesis matrix and criterion-level MMAT appraisals with justifications (Supplementary File S6). All 14 included studies are published and publicly accessible via the DOIs listed in Table 2. No new primary data were generated.

Supplementary Information

Supplementary File S1: Completed PRISMA 2020 checklist.

Supplementary File S2: PRISMA 2020 flow diagram.

Supplementary File S3: Full search strategies for all databases.

Supplementary File S4: Screening and eligibility log.

Supplementary File S5: Data extraction form.

Supplementary File S6: Data extraction and synthesis tables, including MMAT appraisal.

References:

Abdelwanis, M., Simsekler, M. C. E., Gabor, A. F., Sleptchenko, A., & Omar, M. (2026). Artificial intelligence adoption challenges from healthcare providers' perspectives: A comprehensive review and future directions. *Safety Science*, *193*, 107028.

https://doi.org/10.1016/j.ssci.2025.107028

Ajegbile, M., Olaboye, J., Maha, C., Igwama, G., & Abdul, S. (2024). The role of data-driven initiatives in enhancing healthcare delivery and patient retention. *World Journal of Biology Pharmacy and Health Sciences*, *19*, 234–242.

Ali, M., Ali, S., Abbas, Q., Abbas, Z., & Lee, S. W. (2025). Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health. *DIGITAL HEALTH*, *11*. https://doi.org/10.1177/20552076251395548

AlMaskari, A. M., Al-Mahrouqi, T., Al Lawati, A., Al Aufi, H., Al Riyami, Q., & Al-Sinawi, H. (2025). Students' perceptions of AI mental health chatbots: an exploratory qualitative study at Sultan Qaboos University. *BMJ Open*, *15*(10), e103893. https://doi.org/10.1136/bmjopen-2025-103893

Berardi, C., Antonini, M., Jordan, Z., Wechtler, H., Paolucci, F., & Hinwood, M. (2024). Barriers and facilitators to the implementation of digital technologies in mental health systems: a qualitative systematic review to inform a policy framework. *BMC Health Services Research*, *24*(1), 243. https://doi.org/10.1186/s12913-023-10536-1

Beyene, L. S., Strand, E. B., Misund, A. R., Barder, H. E., Brente, T. L., & Størksen, H. T. (2024). Conceptualizing healthcare professionals' relational competence in mental healthcare: An integrative review. *International Journal of Nursing Studies Advances*, *7*, 100266. https://doi.org/10.1016/j.ijnsa.2024.100266

Blease, C., Locher, C., Leon-Carlyle, M., & Doraiswamy, M. (2020). Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digital Health*, *6, 2055207620968355.* https://doi.org/10.1177/2055207620968355

Bucher, A., Egger, S., Vashkite, I., Wu, W., & Schwabe, G. (2025). “It’s Not Only Attention We Need”: Systematic Review of Large Language Models in Mental Health Care. *JMIR Mental Health*, *12*, e78410. https://doi.org/10.2196/78410

Buhr, E., Fischer, M., Biernetzky, O., Teipel, S., Gruber, O., & Schweda, M. (2025). Use of Artificial Intelligence in Psychiatric Research and Practice: A Qualitative Interview Study with Experts from Psychiatry, Computer Science and Philosophy in Germany. *European Psychiatry*, *68(1), e105.* https://doi.org/10.1192/j.eurpsy.2025.10075

Conway, M., & O’Connor, D. (2016). Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology*, *9*, 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004

Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., Titov, N., & Alvarez-Jimenez, M. (2024). Use of AI in Mental Health Care: Community and Mental Health Professionals Survey. *JMIR Mental Health*, *11, e60589.* https://doi.org/10.2196/60589

D’Hotman, D., & Loh, E. (2020). AI enabled suicide prediction tools: a qualitative narrative review. *BMJ Health & Care Informatics*, *27*(3), e100175. https://doi.org/10.1136/bmjhci-2020-100175

Dhaigude, A. S., & Kamath, G. B. (2025). Mapping responsible artificial intelligence in business and management: Trends, influence, and emerging research directions. *Journal of Open Innovation: Technology, Market, and Complexity*, *11*(4), 100640. https://doi.org/10.1016/j.joitmc.2025.100640

Ding, X., & Barbic, S. (2025). Perception of AI Use in Youth Mental Health Services: Qualitative Study. *Journal of Participatory Medicine*, *17, e69449.*

https://doi.org/10.2196/69449

Feng, Y., Hang, Y., Wu, W., Song, X., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-Driven Conversational Agents in Improving Mental Health Among Young People: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *27*, e69639. https://doi.org/10.2196/69639

Gorelik, A. J., Li, M., Hahne, J., Wang, J., Ren, Y., Yang, L., Zhang, X., Liu, X., Wang, X., Bogdan, R., & Carpenter, B. D. (2025). Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1662642

Götzl, C., Hiller, S., Rauschenberg, C., Schick, A., Fechtelpeter, J., Fischer Abaigar, U., Koppe, G., Durstewitz, D., Reininghaus, U., & Krumm, S. (2022). Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. *Child and Adolescent Psychiatry and Mental Health*, *16(1), 86.* https://doi.org/10.1186/s13034-022-00522-6

Grando, A., Ivanova, J., Hiestand, M., Soni, H., Murcko, A., Saks, M., Kaufman, D., Whitfield, M. J., Dye, C., Chern, D., & Maupin, J. (2020). Mental health professional perspectives on health data sharing: Mixed methods study. *Health Informatics Journal*, *26*(3), 2067–2082. https://doi.org/10.1177/1460458219893848

Hao, X., Demir, E., & Eyers, D. (2025). Beyond human-in-the-loop: Sensemaking between artificial intelligence and human intelligence collaboration. *Sustainable Futures*, *10*, 101152. https://doi.org/10.1016/j.sftr.2025.101152

Hipgrave, L., Goldie, J., Dennis, S., & Coleman, A. (2025). Balancing risks and benefits:

clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Frontiers in Digital Health*, *7, 1606291.* https://doi.org/10.3389/fdgth.2025.1606291

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Joseph, J. (2025). Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1643180

Kieling, C., Buchweitz, C., Caye, A., Silvani, J., Ameis, S. H., Brunoni, A. R., Cost, K. T., Courtney, D. B., Georgiades, K., Merikangas, K. R., Henderson, J. L., Polanczyk, G. V., Rohde, L. A., Salum, G. A., & Szatmari, P. (2024). Worldwide Prevalence and Disability From Mental Disorders Across Childhood and Adolescence. *JAMA Psychiatry*, *81*(4), 347. https://doi.org/10.1001/jamapsychiatry.2023.5051

Kim, J. P., Ryan, K., Kasun, M., Hogg, J., Dunn, L. B., & Roberts, L. W. (2023). Physicians' and Machine Learning Researchers' Perspectives on Ethical Issues in the Early Development of Clinical Machine Learning Tools: Qualitative Interview Study. *JMIR AI, 2, e47449.* https://doi.org/10.2196/47449

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & Griffiths, S. L. (2024). The social determinants of mental health and disorder: evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58–90. https://doi.org/10.1002/wps.21160

Laakkonen, N., Jarva, E., Hammarén, M., Kanste, O., Kääriäinen, M., Oikarinen, A., & Mikkonen, K. (2024). Digital Competence among Healthcare Leaders: A Mixed-Methods Systematic Review. *Journal of Nursing Management*, *2024*(1). https://doi.org/10.1155/2024/8435248

Lee, H. S., Wright, C., Ferranto, J., Buttimer, J., Palmer, C. E., Welchman, A., Mazor, K. M., Fisher, K. A., Smelson, D., O'Connor, L., Fahey, N., & Soni, A. (2024). Artificial intelligence conversational agents in mental health: Patients see potential, but prefer humans in the loop. *Frontiers in Psychiatry*, *15, 1505024.* https://doi.org/10.3389/fpsyt.2024.1505024

Limenih, G., MacDougall, A., Wedlake, M., & Nouvet, E. (2024). Depression and Global Mental Health in the Global South: A Critical Analysis of Policy and Discourse. *International Journal of Social Determinants of Health and Health Services*, *54*(2), 95–107. https://doi.org/10.1177/27551938231220230

Martinez-Martin, N., Greely, H. T., & Cho, M. K. (2021). Ethical Development of Digital Phenotyping Tools for Mental Health Applications: Delphi Study. *JMIR MHealth and UHealth*, *9*(7), e27343. https://doi.org/10.2196/27343

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, *10*(4), e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

Montemayor, C., Halpern, J., & Fairweather, A. (2022). In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI & SOCIETY*, *37*(4), 1353–1359. https://doi.org/10.1007/s00146-021-01230-z

Moreno-Sánchez, P. A., Del Ser, J., van Gils, M., & Hernesniemi, J. (2026). A design framework for operationalizing trustworthy artificial intelligence in healthcare: Requirements, tradeoffs and challenges for its clinical adoption. *Information Fusion*, *127*, 103812. https://doi.org/10.1016/j.inffus.2025.103812

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.glmedi.2024.100099

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Parenteau, A. M., Boyer, C. J., Campos, L. J., Carranza, A. F., Deer, L. K., Hartman, D. T., Bidwell, J. T., & Hostinar, C. E. (2023). A review of mental health disparities during COVID-19: Evidence, mechanisms, and policy recommendations for promoting societal resilience. *Development and Psychopathology*, *35*(4), 1821–1842. https://doi.org/10.1017/S0954579422000499

Paudel, S., Chalise, A., Khatri, D., Poudel, S., & Khanal, A. (2025). Nepal's mental health system from public health perspective: a thematic synthesis based on health system building blocks. *The Lancet Regional Health - Southeast Asia*, *36*, 100588. https://doi.org/10.1016/j.lansea.2025.100588

Petersson, L., Ahlborg, M. G., & Häggström Westberg, K. (2025). “I Believe That AI Will Recognize the Problem Before It Happens”: Qualitative Study Exploring Young Adults’ Perceptions of AI in Mental Health Care. *JMIR Mental Health*, *12, e76973.* https://doi.org/10.2196/76973

Pham, T. (2025). Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. *Royal Society Open Science*, *12*(5). https://doi.org/10.1098/rsos.241873

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

Sharif, L., Almabadi, R., Alahmari, A., Alqurashi, F., Alsahafi, F., Qusti, S., Akash, W., Mahsoon, A., Poudel, D. B., Sharif, K., & Wright, R. (2025). Perceptions of mental health professionals towards artificial intelligence in mental healthcare: a cross-sectional study. *Frontiers in Psychiatry*, *16, 1601456.* https://doi.org/10.3389/fpsyt.2025.1601456

Starke, G., Schmidt, B., De Clercq, E., & Elger, B. S. (2023). Explainability as fig leaf? An exploration of experts’ ethical expectations towards machine learning in psychiatry. *AI and Ethics*, *3*(1), 303–314. https://doi.org/10.1007/s43681-022-00177-1

Stroud, A. M., Curtis, S. H., Weir, I. B., Stout, J. J., Barry, B. A., Bobo, W. V., Athreya, A. P., & Sharp, R. R. (2025). Physician Perspectives on the Potential Benefits and Risks of Applying Artificial Intelligence in Psychiatric Medicine: Qualitative Study. *JMIR Mental Health*, *12, e64414. https*://doi.org/10.2196/64414

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45.

https://doi.org/10.1186/1471-2288-8-45

Varsha, P. S. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. *International Journal of Information Management Data Insights*, *3*(1), 100165. https://doi.org/10.1016/j.jjimei.2023.100165

Wainberg, M. L., Scorza, P., Shultz, J. M., Helpman, L., Mootz, J. J., Johnson, K. A., Neria, Y., Bradford, J.-M. E., Oquendo, M. A., & Arbuckle, M. R. (2017). Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. *Current Psychiatry Reports*, *19*(5), 28. https://doi.org/10.1007/s11920-017-0780-z

Wang, R., Chaudhari, P., & Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proceedings of the National Academy of Sciences*, *120*(6). https://doi.org/10.1073/pnas.2211613120

Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm." *Intelligent Medicine*, *4*(1), 52–57. https://doi.org/10.1016/j.imed.2023.08.001

Zabelski, S., Hollander, M., & Alexander, A. (2024). Addressing Inequities in Access to Mental Healthcare: A Policy Analysis of Community Mental Health Systems Serving Minoritized Populations in North Carolina. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*(4), 543–553. https://doi.org/10.1007/s10488-024-01344-8

Zhang, M., Scandiffio, J., Younus, S., Jeyakumar, T., Karsan, I., Charow, R., Salhia, M., & Wiljer, D. (2023). The Adoption of AI in Mental Health Care-Perspectives From Mental Health Professionals: Qualitative Descriptive Study. *JMIR Formative Research*, *7, e47847.*

https://doi.org/10.2196/47847

Zheng, H., & Zhang, X. (2025). Psychiatry in the age of AI: transforming theory, practice, and medical education. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1660448

Ziv, L., & Nakash, M. (2025). Behind the Algorithm: International Insights into Data-Driven AI Model Development. *Machine Learning and Knowledge Extraction*, *7*(4), 122. https://doi.org/10.3390/make7040122

Table 2: *Characteristics and methodological quality of the 14 included studies*

Study	Country	Design (MMAT category)	Participants	AI application	Main ethical issues reported	MMAT criteria met
AlMaskari et al., 2025	Oman	Exploratory qualitative study; 10 online semi-structured focus groups; thematic analysis (Braun & Clarke) (Qualitative)	60 university students aged 18–25 (40 male, 20 female); convenience sample	AI mental health chatbots in general (support, screening, mood tracking); largely hypothetical	Privacy, data security and sharing without consent; cultural/religious insensitivity of advice; lack of empathy and human connection; over-reliance and isolation; inaccurate or hallucinated advice; transparency and accountability; bias from non-representative training data	5/5
Blease et al., 2020	Multinational (22 countries; 35% USA)	Qualitative descriptive content analysis of free-text responses to three open-ended questions in an anonymous web-based survey (Qualitative)	791 psychiatrists (69.5% male) sampled from an online physician network (Sermo)	Hypothetical future AI/ML replacing or augmenting psychiatrists' key tasks	Loss of empathy; harm to the therapeutic relationship; privacy and confidentiality (few comments); medical/diagnostic error; over-reliance and deskilling; threat to the profession; patient non-acceptance	4/5
Buhr et al., 2025	Germany	Qualitative exploratory study; semi-structured expert interviews (video); structuring qualitative content analysis (MAXQDA) (Qualitative)	15 experts: 10 psychiatrists (dementia or depression), 1 computer scientist, 4 philosophers	Prospective AI/ML for psychiatric diagnosis, nosology and care (biomarkers, digital/behavioural markers)	Diagnostic expansion and pathologisation; dismissal of patients' subjective experience; irreplaceable empathy and communication; accountability and obligation to follow AI; data bias and reductionism; dehumanisation of the doctor–patient relationship; opacity and need for AI education	5/5

Study	Country	Design (MMAT category)	Participants	AI application	Main ethical issues reported	MMAT criteria met
Cross et al., 2024	Australia	Two cross-sectional web surveys (community members; professionals) with descriptive statistics and analysis of free-text responses (Quantitative descriptive)	107 community members (70% with lived experience) and 86 mental health professionals; snowball social-media sample	Actual use of commercial generative AI (mostly ChatGPT) plus hypothetical use cases (chatbots, mood tracking, diagnosis, crisis support)	Loss of human connection and empathy; privacy and data security; inaccurate outputs and misdiagnosis; clinician over-reliance; inadequate ethical guidelines and regulation; commercial data sharing and manipulation	1/5
Ding & Barbic, 2025	Canada (British Columbia)	Inductive qualitative study; semi-structured interviews (TAM-informed); thematic analysis; COREQ (Qualitative)	23 stakeholders: 12 youth (18–24), 6 service providers, 5 non-clinical staff	Hypothetical integration of AI for health, mainly generative AI chatbots (education, navigation/triage, administration)	Missing human element and empathy; lack of clinical judgement and personalisation; privacy and confidentiality; inadequate crisis management and harmful content; misinformation and unregulated tools; equitable access; need for regulation and oversight	5/5
Götzl et al., 2022	Germany	Convergent parallel mixed methods: 2 online focus groups and 5 expert interviews (content analysis) + cross-sectional online panel survey (Mixed methods)	8 young people (15–20) and 5 experts (qualitative); 666 young people aged 16–25 (survey)	Planned AI-informed mHealth app personalising ecological momentary interventions (not yet in use)	Acceptability of sharing increasingly intimate data; transparency of data use; user control over personalisation; trust depends on provider; risk of triggering self-harm and need for crisis referral; imprecise suggestions and false expectations; need for regulation	3/5
Hipgrave et al., 2025	Australia	Structured interviews combining open-ended questions (thematic analysis) and Likert ratings before/after a chatbot demonstration (Mixed methods)	23 mental health clinicians (psychologists, counsellors, social workers, support workers, 2 psychiatrists); convenience sample	Generative AI chatbots (GPTs built on ChatGPT-4o) for client-facing and clinician-facing tasks; hypothetical demonstration	Lack of regulation and accountability; data security and privacy; limited cultural understanding; client over-reliance; incorrect treatment recommendations; missed non-verbal cues and crisis risk; sycophancy and cost-cutting replacing oversight	3/5

Study	Country	Design (MMAT category)	Participants	AI application	Main ethical issues reported	MMAT criteria met
Kim et al., 2023	United States	Qualitative descriptive study; semi-structured Zoom interviews; qualitative content analysis (Qualitative)	21 participants: 11 physicians (incl. psychiatry; 1 clinical psychologist) and 10 AI/ML researchers	Clinical AI/ML tools across medicine (development to implementation); not specific to mental health	Misaligned priorities (innovation vs patient care); questionable need for AI; tools lacking evidence and clinical value; academic incentives; lack of diversity and stakeholder involvement; interdisciplinary knowledge gaps; absence of ethics standards and training	5/5
Lee et al., 2024	United States	Qualitative study; semi-structured remote interviews; rapid qualitative analysis (Qualitative)	29 adults (19–66) with self-reported mild–moderate anxiety	AI conversational agents (chatbots) for mental health support; mostly hypothetical/prospective	Lack of empathy; inability to manage crises such as suicidality; data privacy and confidentiality; trust in who holds data; bias and stereotyping; replacement of therapists and preference for human-in-the-loop	5/5
Petersson et al., 2025	Sweden	Qualitative inductive study; individual semi-structured interviews; conventional content analysis; COREQ (Qualitative)	25 young adults (20–30) with experience of help-seeking for common mental health problems	General/hypothetical AI in mental health care (chatbots, triage, screening, diagnosis, monitoring)	Privacy, data security and personal integrity; algorithmic bias; lack of empathy and human relationship; diagnostic error and over-reliance; medicalisation and self-fulfilling predictions; access barriers; responsibility for monitoring data	5/5
Sharif et al., 2025	Saudi Arabia (Jeddah)	Cross-sectional online survey (Shinners AI Perception questionnaire); descriptive and inferential statistics (Quantitative descriptive)	251 mental health professionals (55% nurses; psychiatrists, psychologists, sociologists)	AI in mental healthcare in general (no specific tool); 24% reported current use	Inadequate AI training and preparedness; responsibility for AI errors; presence of workplace ethical frameworks; professional role displacement; changing professional roles; need for safe and ethical implementation	3/5

Study	Country	Design (MMAT category)	Participants	AI application	Main ethical issues reported	MMAT criteria met
Starke et al., 2023	Switzerland and Germany	Qualitative study; semi-structured expert interviews (phone/video); reflexive thematic analysis (Qualitative)	15 experts (10 physicians; 5 psychology/neuroscience) working on ML in psychiatry (15 of 26 invited)	Machine-learning tools for diagnosis, prognosis and treatment-response prediction; mostly under development	Bias and fairness; publication pressure vs methodological rigour and external validation; data sharing vs privacy; explainability as a 'fig leaf'; power shifting from psychiatrist to algorithm; stigmatising or self-fulfilling predictions; need for ML education and regulatory oversight	5/5
Stroud et al., 2025	United States	Qualitative study; case-based (vignette) semi-structured interviews; inductive coding in duplicate (Qualitative)	42 physicians: 21 psychiatrists and 21 family medicine physicians (29% participation)	Hypothetical tools: pharmacogenomic prescribing support, adjunct CBT chatbot, diagnostic risk score, population suicide-risk prediction	Unreliable recommendations from skewed data; over-reliance and deskilling; algorithmic bias and patient harm; privacy and data security; medicolegal liability; chatbots unable to replicate therapeutic trust; added costs and clinician burden	5/5
Zhang et al., 2023	Canada (mainly Ontario)	Qualitative descriptive study; semi-structured interviews; hybrid deductive–inductive thematic analysis (Qualitative)	20 mental health professionals (social workers, nurses, psychiatrist- and nurse-researchers, educators)	AI in mental health care generally (diagnosis/prediction, apps and chatbots, suicide-risk modelling); mostly prospective	Loss of humanistic care and therapeutic relationship; one-size-fits-all tools; dataset bias and representativeness; equity and accessibility; transparency, trust and need for validated, regulated tools; harm to clients; fear of replacement	5/5

Note. MMAT = Mixed Methods Appraisal Tool (version 2018); criteria met = number of the five design-specific criteria rated 'Yes' (see Figure 2 and Supplementary File S6). DOIs: AlMaskari et al., 2025: 10.1136/bmjopen-2025-103893; Blease et al., 2020: 10.1177/2055207620968355; Buhr et al., 2025: 10.1192/j.eurpsy.2025.10075; Cross et al., 2024: 10.2196/60589; Ding & Barbic, 2025: 10.2196/69449; Götzl et al., 2022: 10.1186/s13034-022-00522-6; Hipgrave et al., 2025: 10.3389/fdgth.2025.1606291; Kim et al., 2023: 10.2196/47449; Lee et al., 2024: 10.3389/fpsyt.2024.1505024; Petersson et al., 2025: 10.2196/76973; Sharif et al., 2025: 10.3389/fpsyt.2025.1601456; Starke et al., 2023: 10.1007/s43681-022-00177-1; Stroud et al., 2025: 10.2196/64414; Zhang et al., 2023: 10.2196/47847.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S5Dataextractionform.docx
- S2PRISMA2020flowdiagram.pdf
- S6Dataextractionandsynthesistables.xlsx
- S3Fullsearchstrategies.docx
- S1PRISMA2020checklist.docx