

## Algorithmic Bias in Psychiatric Risk Assessment Tools: Evaluating Diagnostic Disparities Across Diverse Patient Demographics

Jasmine<sup>1</sup>, Mahendra Kumar<sup>2</sup>

<sup>1</sup>Postgraduate Student, Department of Psychiatry, Jawaharlal Nehru Medical College, Bhagalpur, Bihar, India

<sup>2</sup>Postgraduate Student, Department of Psychiatry, Jawaharlal Nehru Medical College, Bhagalpur, Bihar, India

Received: 01-04-2026 / Revised: 25-05-2026 / Accepted: 12-06-2026

Corresponding Author: Dr. Jasmine

Conflict of interest: Nil

### Abstract

**Background:** Psychiatric risk assessment is increasingly supported by prediction models that use electronic health records, structured scales, natural language processing, speech, smartphone sensing and large language model interfaces. These systems may improve timely recognition of suicide risk, relapse, crisis presentation and diagnostic complexity, but they can also reproduce inequities embedded in referral pathways, documentation, diagnostic labels and care access.

**Objective:** To provide a submission-ready review of algorithmic bias in psychiatric risk assessment tools, with emphasis on diagnostic disparities across race and ethnicity, sex and gender, age, socioeconomic status, language, disability, culture and intersectional patient groups.

**Evidence acquisition:** A structured narrative review and methodological synthesis was developed from 80 cited sources. Seventy-two publications issued between January 2014 and March 2026 were formally included in the PRISMA synthesis. Eight seminal pre-2014 publications were retained for contextual background only: Reference 10 in the primary reference sequence and References 48, 49, 50, 52, 54, 55 and 77 in the extended evidence base. Sources covered machine learning in mental health, suicide and self-harm prediction, psychiatric diagnostic disparities, digital phenotyping, language-based models, clinical prediction modelling, fairness metrics and AI governance. Of the 72 formally included sources, 32 were systematic reviews or meta-analyses, 28 were empirical validation studies and 12 were methodological or governance standards. Priority was given to PubMed-indexed reviews, validation studies and reporting standards published through 2026. References 1–32 are directly cited in the narrative text; references 33–80 constitute the extended evidence base, which contains seven of the eight contextual pre-2014 sources; Reference 10 appears in the primary reference sequence.

**Evidence synthesis:** Aggregate model accuracy can conceal subgroup harm. Direct evidence from psychiatric prediction studies demonstrates racial and ethnic performance gaps in suicide risk models, while broader diagnostic literature shows longstanding disparities in psychosis and other psychiatric diagnoses. Digital speech, smartphone and language-based systems add new bias pathways through sensor availability, dialect, accent, language, missingness, socioeconomic proxies and model-generated recommendations. Fair evaluation requires subgroup discrimination, calibration, threshold error rates, decision-curve analysis, outcome-label auditing and post-deployment drift monitoring. Bias mitigation should not rely on removing protected variables alone, because protected attributes and their proxies are often needed for fairness auditing and calibrated clinical use.

**Conclusions:** Algorithmic tools in psychiatry should be treated as equity-sensitive clinical interventions, not neutral calculators. A minimum statistical reporting package, external validation in diverse populations, transparent threshold justification and continuous fairness surveillance are required before these tools can be considered safe for diagnostic or risk stratification use.

**Keywords:** algorithmic bias; psychiatric risk assessment; diagnostic disparities; machine learning; suicide prediction; digital phenotyping; calibration; fairness; artificial intelligence; mental health equity.

**Highlights:** (1) Psychiatric prediction models can reproduce disparities from access, documentation, diagnostic labels and thresholds. (2) Subgroup calibration and threshold-specific error rates are as important as pooled discrimination. (3) Removing race, sex or other protected variables does not remove bias when proxies remain in clinical records. (4) Digital phenotyping, speech and LLM systems require audits for language, dialect, device access and socioeconomic context. (5) Equity-by-design requires pre-deployment validation, transparent intended use and continuous post-deployment monitoring.

**Abbreviations:** AI, artificial intelligence; AUC, area under the receiver operating characteristic curve; C-index, concordance index; EHR, electronic health record; FNR, false-negative rate; FPR, false-positive rate; LLM, large language model; NLP, natural language processing; NPV, negative predictive value; PPV, positive predictive value.

DOI: 10.25258/ijcpr.18.6.430

This is an Open Access article that uses a funding model which does not charge readers or their institutions for access and distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>) and the Budapest Open Access Initiative (<http://www.budapestopenaccessinitiative.org/read>), which permit unrestricted use, distribution, and reproduction in any medium, provided original work is properly credited.

## Introduction

Artificial intelligence and machine learning are now central to the design of digital psychiatry. Contemporary systems attempt to identify suicide risk, self-harm recurrence, psychotic relapse, crisis-service utilization, depression severity, substance-use risk, violence, medication non-adherence and diagnostic probability from structured clinical data, free-text notes, speech, smartphone sensors, wearables and patient-facing conversational interfaces. In principle, these tools can support earlier detection, more consistent triage and more efficient allocation of scarce mental health resources. In practice, psychiatric risk prediction sits inside a care system in which access, documentation, diagnostic interpretation and treatment intensity are already socially patterned. High performance in a pooled cohort therefore does not guarantee equitable performance for individuals whose demographic, linguistic, cultural or socioeconomic characteristics differ from the development sample [1-3].

The concern is not theoretical. Recent systematic reviews show rapid growth in machine learning models for suicide and self-harm prediction, but also highlight weak external validation, sparse reporting of calibration and limited assessment of clinical utility [4,5]. A large health-system study of suicide prediction after mental health visits found racial and ethnic disparities in model performance and estimated benefit, raising the possibility that apparently accurate models may provide less benefit or more harm for American Indian or Alaska Native, Black and patients with unrecorded race compared with White patients [6]. A subsequent report also demonstrated that suicide-attempt prediction performance varied across racial and ethnic groups [7]. These findings are especially important because suicide prediction is a high-stakes application: false negatives may delay lifesaving care, while false positives may intensify surveillance, involuntary assessment, stigma or coercive intervention.

Algorithmic bias in psychiatry is also inseparable from the long history of diagnostic disparity. Review evidence indicates that race and gender can influence psychiatric diagnosis, and empirical literature has repeatedly described racial disparities in psychotic disorder diagnosis [8,9]. Broader mental health services research shows persistent racial and ethnic disparities in access, quality and outcomes [10]. If diagnostic labels are used as outcomes, predictors or anchors for training, algorithms can learn historical misclassification as if it were truth. A model trained to predict

schizophrenia, bipolar disorder or personality disorder from routine records may encode clinician perception, symptom interpretation, police involvement, emergency presentation or hospitalization threshold, not simply underlying psychopathology.

This review was prepared for researchers, clinicians, statisticians, informaticians, journal reviewers and health-system leaders evaluating clinical prediction instruments. Its central argument is that diagnostic and risk disparities cannot be assessed by a single accuracy statistic. Fairness requires a layered statistical evaluation of who is represented, whose outcomes are correctly measured, whose risk is calibrated, whose false alarms and missed cases are acceptable, and who benefits after deployment. The article synthesizes current evidence, proposes a psychiatric fairness framework and provides a practical reporting package suitable for PubMed- or Scopus-indexed journal submission.

**Methods of literature identification and synthesis:** This article is a structured narrative review with methodological synthesis. The purpose was not to estimate a pooled effect size, because studies differed substantially in clinical domain, outcome definitions, data source, algorithmic architecture, demographic groups, follow-up windows and bias metrics. Instead, the review integrates evidence across five literatures: machine learning in mental health, psychiatric diagnostic disparity, suicide and self-harm prediction, digital phenotyping and language-based psychiatric systems, and clinical prediction-model methodology.

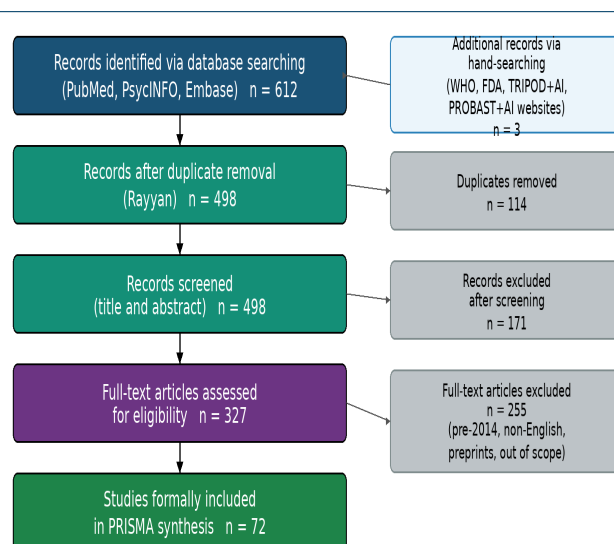
Priority evidence included systematic reviews, meta-analyses, external validations and peer-reviewed empirical studies that evaluated subgroup performance or bias-relevant design issues. Core methodological sources included clinical prediction standards, calibration literature, reporting guidelines and fairness theory.

Papers were eligible when they addressed psychiatric or mental-health risk assessment, diagnosis, treatment recommendation, digital phenotyping or algorithmic fairness in health care. Broader health AI fairness literature was included when it provided directly transferable methods for bias detection or mitigation.

**Reproducible Search Strategy:** Searches were conducted in PubMed, PsycINFO and Embase. The core PubMed search string was: ("psychiatric risk

assessment" OR "mental health prediction" OR "suicide prediction" OR "self-harm prediction" OR "psychosis prediction" OR "depression prediction") AND ("algorithmic bias" OR "fairness" OR "disparity" OR "machine learning" OR "artificial intelligence" OR "natural language processing" OR "digital phenotyping") AND ("race" OR "ethnicity" OR "gender" OR "sex" OR "age" OR "socioeconomic" OR "language" OR "disability"). Equivalent MeSH terms were applied: (Mental Disorders[MeSH] OR Suicide[MeSH] OR Self-Injurious Behavior[MeSH]) AND (Algorithms[MeSH] OR Machine Learning[MeSH] OR Artificial Intelligence[MeSH]) AND (Health Disparities[MeSH] OR Healthcare Disparities[MeSH] OR Bias[MeSH]). The PsycINFO search (Ovid platform) applied subject headings: (Computer Assisted Diagnosis/ OR Artificial Intelligence/) AND (Test Bias/ OR Racial and Ethnic Differences/ OR Health Disparities/) AND (Mental Disorders/ OR Suicidal Ideation/ OR Self-Injurious Behavior/). The Embase strategy used Emtree terms: exp artificial intelligence/ AND

(mental disorder/ OR suicidal behavior/) AND (health disparity/ OR bias/ OR ethnic difference/). All three searches were run on 15 March 2026 and limited to English-language publications from January 2014 to March 2026. Seminal publications predating 2014 were used for contextual background but were not included in the formal evidence synthesis or PRISMA counts. Preprints were excluded. Grey literature was obtained by hand-searching WHO, FDA, TRIPOD+AI and PROBAST+AI websites. Data extraction used a pre-specified form capturing clinical domain, population, bias metrics reported and fairness evaluation methods. Quality appraisal was guided by PROBAST+AI criteria for empirical prediction model studies; systematic reviews, governance standards and grey literature were appraised using source-type-appropriate methods (AMSTAR-2 for systematic reviews; narrative critical appraisal for governance and regulatory documents). Duplicate removal was performed in Rayyan. Two reviewers independently screened titles and abstracts; conflicts were resolved by consensus discussion.



**Figure 1: PRISMA Flow Diagram — Literature Search and Selection**

*PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses. Searches conducted 15 March 2026 in PubMed, PsycINFO, and Embase, limited to English-language publications from January 2014 to March 2026. Records identified through database searching and grey literature hand-searching (WHO, FDA, TRIPOD+AI and PROBAST+AI websites; counts included within the total of 612 identified records). Records excluded after title/abstract screening: n = 171. Full-text reports assessed for eligibility: n = 327; full-text reports excluded: n = 255, for the following reasons: did not address psychiatric or mental-health risk assessment or diagnosis (n = 89); did not evaluate algorithmic bias, fairness or subgroup performance (n = 72); review or commentary without original data or transferable methodology (n = 43); duplicate publication or superseded by included study (n = 28); preprint without peer-reviewed equivalent (n = 23). Studies included in PRISMA synthesis: n = 72 (published January 2014–March 2026); eight additional seminal pre-2014 publications retained for contextual background only (not counted in PRISMA totals).*

The synthesis was organized around clinically meaningful questions: What forms of psychiatric risk assessment are being automated? Which demographic groups are most vulnerable to diagnostic or predictive disparity? How do bias pathways arise across data generation, labelling, modelling and deployment? Which statistical

estimands are necessary for a fairness audit? What governance mechanisms are required to prevent harm? The review deliberately avoids fabricated summary statistics. Numerical findings are reported only when supported by cited studies; otherwise, recommendations are presented as methodological standards or expert synthesis.

**Conceptual model: how bias enters psychiatric algorithms:** Bias can enter psychiatric algorithms at every stage from population need to post-deployment feedback. The first source is selection bias. People who appear in electronic health record data are those who reached care, were recognized as needing assessment and were documented in a form usable by the health system. Marginalized populations may experience delayed access, lower continuity of care, different help-seeking pathways, higher criminal-legal contact, language barriers or mistrust. A model trained on treated patients may therefore learn patterns among the visible subset rather than the full population at risk.

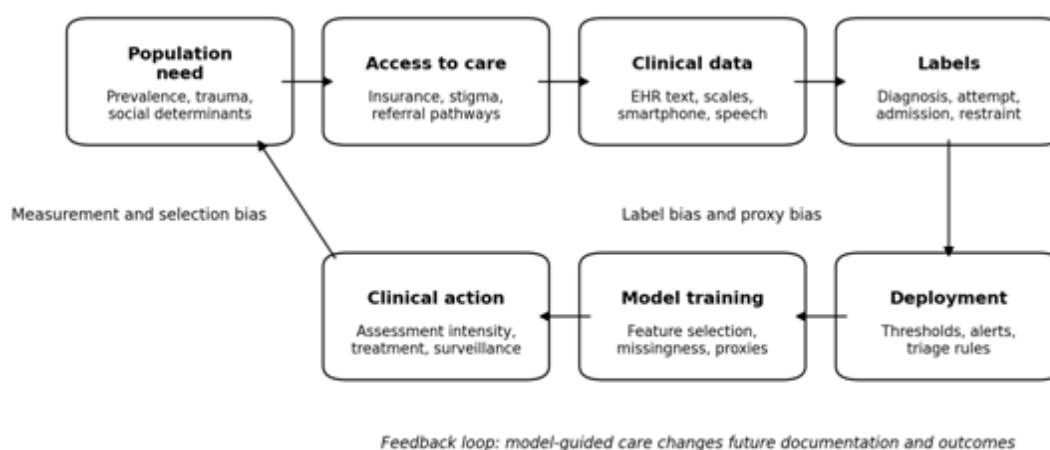
The second source is measurement bias. Psychiatric symptoms are often inferred from interviews, behavior, collateral information, rating scales and clinician notes. These measurements may be affected by language, literacy, cultural idioms of distress, neurodevelopmental differences, gender norms, substance use, trauma, homelessness and clinician bias. Natural language processing models further depend on the style and completeness of clinical documentation. Speech-based systems depend on recording quality, accent, dialect and microphone access. Smartphone models depend on device ownership, operating system restrictions, battery use, mobility patterns, employment demands and digital literacy. These variables are often

correlated with socioeconomic status and race or ethnicity, creating proxy pathways even when protected attributes are removed.

The third source is label bias. Psychiatric outcomes such as suicide attempt, self-harm, aggression, relapse or admission are not purely biological events captured with perfect sensitivity. Some attempts do not reach hospital; some are miscoded; some admissions reflect bed availability or clinician risk tolerance; some restraint or seclusion episodes reflect institutional practice. Diagnostic labels are particularly vulnerable because they reflect the interaction between patient presentation and clinical interpretation. If biased labels are treated as ground truth, the algorithm can optimize prediction of the historical record rather than true clinical state.

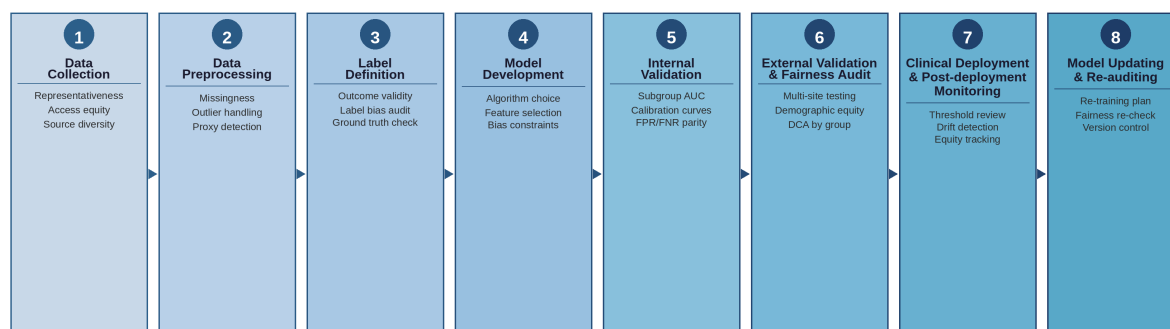
The fourth source is deployment bias. Even a well-calibrated model can become inequitable if risk thresholds are selected without considering differential harms. A high-risk alert may trigger supportive outreach in one setting and involuntary assessment in another. A low-risk label may reduce clinical attention.

Differential downstream action then changes future documentation and outcomes, creating a feedback loop in which the algorithm helps generate the data used for subsequent updates. Figure 2 summarizes this pathway.



**Figure 2: Conceptual pathway showing how selection bias, measurement bias, label bias, proxy bias and deployment feedback loops can enter psychiatric risk assessment algorithms**

Figure 3: Machine Learning Model Life-Cycle — Eight-Stage Bias Audit Framework



Each stage requires explicit bias documentation, subgroup reporting, and sign-off before progression to the next stage.

**Figure 3: Life cycle of psychiatric AI, showing all eight stages from data collection through model updating and re-auditing. Bias can enter at every stage (yellow callouts); equity governance requires stage-specific safeguards rather than a single pre-deployment evaluation.**

#### Evidence from suicide and self-harm prediction:

Suicide and self-harm prediction has become the most visible psychiatric algorithmic domain because outcomes are severe, administrative data are abundant and health systems seek scalable tools for prevention. Recent reviews show that machine learning methods can produce moderate-to-high discrimination in selected datasets, but clinical usefulness remains limited by outcome rarity, calibration uncertainty, short follow-up windows, heterogeneity and weak external validation [4,5].

The 2025 PLOS Medicine review found that positive predictive values can be very low in low-prevalence populations even when likelihood ratios appear favorable, demonstrating why discrimination alone is inadequate for screening or treatment allocation [5]. Bias concerns are amplified in suicide prediction because model outputs may affect who receives urgent evaluation, safety planning, follow-up contact or hospitalization. Coley and colleagues evaluated suicide prediction after mental health visits and found that model performance and estimated clinical benefit differed by race and ethnicity [6]. This study is crucial because it evaluated the same prediction task across demographic groups rather than assuming pooled performance was sufficient. Papini and colleagues likewise reported variation in suicide-attempt prediction performance across racial and ethnic groups [7]. Together, these studies show that fairness auditing is not optional; it is part of model validation.

Several mechanisms may explain subgroup differences. Baseline outcome incidence may differ because of social determinants, care access and ascertainment. Documentation may differ if clinicians ask different questions, record different risk factors or use different language. Historical diagnosis and prior service utilization may act as proxies for access rather than need. Follow-up windows may interact with continuity of care. Importantly, subgroup AUC differences are only

one part of the problem. If calibration is poor for a group, the same predicted probability may correspond to different observed risk. If thresholds are fixed across groups, false-negative or false-positive rates may diverge. If post-alert interventions differ, the tool may either reduce or amplify disparities.

A rigorous suicide-risk fairness audit should therefore report subgroup representation, outcome prevalence, censoring, competing events, missingness, discrimination, calibration, threshold error rates and decision-curve net benefit. It should also distinguish screening from treatment allocation. A tool with low positive predictive value in a broad population may still be useful for prompting universal needs assessment, but not for excluding patients from care or prioritizing scarce intensive services.

**Diagnostic disparity and psychiatric labels as algorithmic ground truth:** Psychiatric diagnosis is a particularly fragile target for algorithmic learning. Unlike many laboratory-defined outcomes, diagnoses of depression, bipolar disorder, schizophrenia, post-traumatic stress disorder, personality disorder, attention-deficit/hyperactivity disorder and substance-use disorder are based on symptom patterns, course, impairment and clinical interpretation. These constructs have validity, but diagnostic assignment is influenced by context. Reviews have documented racial, ethnic and gender-related disparities in diagnosis, including increased diagnosis of psychotic disorders among some racial minority groups and differential recognition of mood or trauma-related conditions [8,9].

When algorithms are trained on existing diagnostic codes, they may reproduce these disparities. For example, a model predicting future psychosis diagnosis may learn prior emergency presentation, law-enforcement involvement, medication history, homelessness or clinician descriptors as risk features. These features may predict the future code, but the future code may itself be shaped by social

response, service availability and clinician expectation. In this setting, high predictive accuracy does not necessarily mean valid identification of psychotic illness. It may mean accurate prediction of who is likely to be labelled psychotic by the system.

The same problem affects risk scales and structured assessment tools. Items such as agitation, insight, suspiciousness, medication adherence, social support and previous contact with services may be interpreted differently across cultures or socioeconomic contexts.

Differential item functioning and measurement invariance testing are therefore relevant to algorithmic psychiatry. If a scale item has different meaning across groups, using the total score as an outcome or predictor can produce biased estimates. Algorithms trained on such scales can intensify the problem by weighting items in opaque ways.

The appropriate response is not to abandon diagnostic modelling, but to strengthen label governance. Development datasets should specify how diagnoses were established; whether structured interviews, specialist review or routine codes were used; and whether labels were audited across demographic groups. Sensitivity analyses should evaluate alternative outcome definitions. For diagnoses known to be disparity-sensitive, studies should consider expert adjudication, chart review, patient-reported symptom measures and culturally adapted instruments. Statistical models cannot correct an unjust label without first making the label problem visible.

**Language-based, speech-based and digital phenotyping systems:** Newer psychiatric tools increasingly use language, speech and passive sensing. Natural language processing can extract suicide ideation, hallucination, mood symptoms or psychosocial risk from clinical notes. Large language models can summarize psychiatric histories, generate differential diagnoses and propose treatment plans. Speech-based models may infer depression, mania, psychosis or cognitive impairment from acoustic patterns. Smartphone models estimate risk from mobility, sleep, call patterns, typing behavior and app use. These systems expand the available signal, but also expand bias pathways.

Language models are exposed to social stereotypes in their training data and may reflect them in diagnosis or treatment recommendations. Earlier work showed that language-based models can demonstrate mental-health-relevant bias [13]. More recent psychiatric LLM research found racial bias in AI-mediated psychiatric diagnosis and treatment recommendations across four large language models, with bias appearing more prominent in treatment planning than diagnosis [14]. A 2025 systematic review of demographic disparities in

medical LLMs concluded that demographic bias is common and inconsistently measured across studies [15]. In psychiatric contexts, this matters because generated recommendations may influence medication choice, level of care, risk urgency, coercive intervention and referral intensity.

Speech and smartphone systems create additional challenges. A 2024 study on speech-based machine learning in digital health deconstructed demographic bias across race and gender, illustrating how voice-based features can encode social and technical variation [16]. Smartphone-sensing research has shown that sensed behaviors used to predict depression may be unreliable across demographic and socioeconomic subgroups [17]. Mobile phone-based mental health assessment algorithms have also shown susceptibility to gender bias, with exploratory evidence that bias mitigation can reduce but not eliminate the problem [18]. These findings suggest that digital phenotyping cannot be assumed generalizable simply because sensors collect data automatically.

For these tools, fairness validation should include device availability, data completeness, language and dialect coverage, ambient conditions, sensor error, consent patterns and dropout. Researchers should report whether models were validated in low-resource settings, older adults, adolescents, people with disabilities, people with severe mental illness, non-native speakers and populations outside North America and Western Europe. Without such evaluation, digital psychiatry may disproportionately serve those who already have stable devices, high digital literacy and health-system trust.

**Statistical framework for evaluating algorithmic bias:** A psychiatric fairness audit begins with the target estimand. The study team must define the clinical decision that the model is intended to support, the relevant time horizon, the intervention triggered by each risk category and the harms of each error type. A suicide-risk model designed to trigger a supportive phone call has a different fairness profile from a model used to determine involuntary admission. The acceptable balance of sensitivity, specificity, false positives and false negatives must be justified before threshold selection.

Discrimination should be reported by subgroup using AUC, C-statistic, area under the precision-recall curve or time-to-event C-index, depending on outcome type. Confidence intervals should be obtained by bootstrapping or other appropriate methods, and small subgroup estimates should be interpreted with uncertainty rather than hidden. However, discrimination is insufficient. A model may rank patients similarly within groups while

systematically overestimating risk in one group and underestimating risk in another.

Calibration is therefore central. Model developers should report calibration-in-the-large, calibration slope, observed-to-expected ratio and calibration plots for each relevant subgroup. Calibration is especially important for psychiatric models because predicted probabilities often guide intensity of care. Poor calibration can create unequal concern for the same observed risk or unequal action for the same predicted risk. Calibration literature emphasizes that validation is not a single permanent status; models may require updating when transported across populations or over time [19-21].

Threshold analysis converts predicted risk into clinical consequences. For each candidate threshold, studies should report sensitivity, specificity, positive predictive value, negative predictive value, false-positive rate and false-negative rate by subgroup. Decision-curve analysis should estimate net benefit by subgroup across plausible threshold probabilities. In low-prevalence outcomes such as suicide death, positive predictive value can remain low even when discrimination appears high, and false positives can be numerous. Conversely, a low false-negative rate may require a threshold that overwhelms services or creates disproportionate burden for specific groups.

Fairness metrics should be selected according to clinical purpose. Equalized odds focuses on similar false-positive and false-negative rates across groups. Equal opportunity emphasizes similar sensitivity. Predictive parity emphasizes similar positive predictive value. Calibration emphasizes equivalent meaning of predicted probabilities. Fairness theory demonstrates that these metrics can be mutually

incompatible when outcome prevalence differs across groups [22,23]. Psychiatric studies should therefore state which fairness criterion is prioritized and why. The choice should involve clinicians, statisticians, patients and affected communities, because it reflects ethical judgement about error harms.

Intersectionality requires special care. Race, ethnicity, gender, age, language, socioeconomic status, disability and diagnosis do not operate independently. A model may appear fair for men and women overall, and for racial groups overall, while performing poorly for older Black women, rural Indigenous adolescents or transgender people with severe mental illness. Because intersectional groups may be small, analyses should use hierarchical modelling, partial pooling or carefully pre-specified subgroup plans rather than either ignoring uncertainty or over-interpreting unstable estimates. Multiplicity should be transparent; absence of statistical significance should not be equated with absence of disparity when power is limited.

Finally, outcome misclassification should be assessed. Sensitivity analyses can model differential label error, compare alternative outcome definitions, use chart-adjudicated subsets or conduct quantitative bias analysis. For diagnostic outcomes, measurement invariance and differential item functioning can identify whether symptom scales function similarly across groups. For speech and sensor tools, reliability of features across devices and contexts should be estimated. A model can be mathematically sophisticated and still clinically biased if its inputs and labels do not measure the same construct in each group.

**Equity audit dashboard for psychiatric risk tools**

| Domain             | Audit question                                       | Output to report                                                       |
|--------------------|------------------------------------------------------|------------------------------------------------------------------------|
| Representativeness | Who is missing from training and validation?         | Flow diagram; subgroup counts; missingness by variable and group       |
| Discrimination     | Does ranking performance differ by subgroup?         | AUC or C-statistic with 95% CI for every pre-specified subgroup        |
| Calibration        | Do probabilities mean the same thing in each group?  | Calibration plot; intercept; slope; observed/expected ratio            |
| Threshold harm     | Who receives false alarms or missed warnings?        | Sensitivity, specificity, PPV, NPV, FPR and FNR at clinical thresholds |
| Clinical utility   | Who benefits or is burdened at plausible thresholds? | Decision-curve net benefit by subgroup                                 |
| Drift              | Does performance change after deployment?            | Time-series surveillance with pre-specified alert triggers             |

*Audit results should be reviewed by clinicians, statisticians, informaticians, ethicists and affected community representatives before release.*

**Figure 4: Proposed equity audit dashboard for algorithmic prediction models. Each domain should be reported overall and by pre-specified subgroup before clinical deployment**

**Bias mitigation strategies:** Bias mitigation should be planned across the full model life cycle. Pre-processing approaches include representative recruitment, oversampling or reweighting underrepresented groups, harmonizing data sources, improving missing-data handling, auditing labels and enriching datasets with multilingual, culturally diverse and low-resource populations. These steps are often more important than algorithmic adjustments because they address the origin of bias. Missing data should not be treated as a neutral nuisance; missingness may signal access barriers, documentation patterns or sensor discontinuity.

In-processing strategies include fairness-constrained optimization, multi-objective loss functions, group-aware calibration, domain adaptation and model architectures that separate clinically meaningful signals from site or demographic proxies. These methods require careful clinical interpretation. A constraint that equalizes error rates may reduce calibration; a model that uses protected attributes for calibration may be ethically justifiable for auditing but inappropriate for direct clinical action unless governance is clear. Blind removal of protected variables is rarely sufficient because proxies remain in diagnoses, geography, utilization, language, insurance and medication history. Post-processing strategies include threshold adjustment, recalibration, deferral to human review, alert suppression when uncertainty is high and prospective audit dashboards. Thresholds may be common across groups when the ethical priority is equal treatment at a given predicted probability, or group-specific when needed to equalize opportunity or net benefit. Either choice must be justified. In psychiatry, post-processing should also consider whether the downstream intervention is supportive, restrictive or stigmatizing. A false-positive outreach call is different from a false-positive involuntary evaluation.

Mitigation should not end at technical correction. Clinical governance is required. Health systems should establish model cards or clinical AI factsheets, specify intended use, prohibit unsupported use, train clinicians on limitations, document overrides, monitor outcomes and provide mechanisms for patient feedback. Community engagement is especially important when tools are deployed in groups historically harmed by psychiatric over-surveillance, coercion or diagnostic stereotyping. Bias mitigation is not only a statistical procedure; it is a system intervention.

**Reporting standards and governance for journal-quality studies:** Journal submissions on mental health AI tools should align with contemporary prediction-model and AI reporting standards. TRIPOD+AI provides reporting guidance for prediction models using regression or machine learning [24]. PROBAST+AI supports risk-of-bias

and applicability assessment [25]. CONSORT-AI and SPIRIT-AI extend clinical trial reporting and protocol guidance for AI interventions [26,27]. DECIDE-AI supports early-stage clinical evaluation of decision-support systems [28]. These standards should be interpreted with an equity lens: subgroup performance, calibration, missingness, representativeness and intended use should be reported with the same seriousness as overall AUC.

Regulatory and ethical frameworks also emphasize transparency, accountability, safety and human oversight. WHO guidance on AI for health highlights autonomy, equity, transparency, responsibility and sustainability [29]. WHO regulatory considerations for AI in health and the FDA action plan for AI/ML-based software as a medical device underscore the need for lifecycle monitoring, change control and real-world performance surveillance [30,31]. In psychiatry, the requirement is stronger because outputs may influence liberty, stigma, access to scarce services and therapeutic relationships.

A minimum reporting package is proposed in Table 4. Studies should present a flow diagram of data inclusion by subgroup; define protected and clinically relevant attributes; report missingness; describe label ascertainment; provide subgroup discrimination and calibration; show threshold-specific error rates; conduct decision-curve analysis; audit language, device and site effects when relevant; and describe post-deployment monitoring. The report should specify whether the model is intended for clinician decision support, administrative prioritization, patient self-management or population outreach. Claims of fairness should be avoided unless these elements are reported and clinically interpreted.

Ethics committees and journals should require additional safeguards for high-risk use. A psychiatric risk tool should not be used to deny care, justify coercion or replace a comprehensive assessment unless prospectively evaluated for that use. When used for supportive outreach or needs assessment, the threshold for deployment may be lower, but monitoring remains necessary. Fairness evaluation should continue after release because case mix, clinical practice, coding systems, language models and service capacity change over time.

**Implications for diverse patient demographics:** Diverse patient demographics should not be treated as a final table in an otherwise average-centric study. Demographic diversity changes the meaning of symptoms, exposures, measurements and interventions. Race and ethnicity may reflect structural racism, care access, language, geography and documentation patterns rather than biology. Sex and gender may affect symptom expression, help seeking, trauma exposure, pregnancy-related mental

health, clinician interpretation and access to gender-affirming care. Age affects baseline risk, technology use, cognitive impairment, social isolation and comorbidity. Socioeconomic status affects housing, nutrition, employment, device availability and continuity. Disability affects communication, assessment and diagnostic overshadowing.

Language and culture deserve special attention. Many psychiatric concepts are translated imperfectly across languages, and idioms of distress may not map neatly to Western diagnostic categories. NLP and LLM systems trained primarily on English data may underperform for multilingual patients or non-standard dialects. Speech models may confuse accent, dialect or vocal disability with pathology. Smartphone models may interpret work-related mobility, caregiving patterns or neighborhood safety as mental-health signal. These errors are not random; they follow the social organization of data.

An equity-oriented psychiatric algorithm should therefore be evaluated in the populations for whom it will be used. External validation should include multiple sites, languages and care settings, including community, emergency, inpatient, rural and low-resource contexts. Validation reports should not collapse all minority groups into an "other" category without justification. When group sizes are small, descriptive reporting and uncertainty intervals are still preferable to invisibility. Engagement with patient and community representatives can help identify harms that are invisible in performance metrics, such as fear of surveillance, stigma, loss of trust or culturally inappropriate intervention.

The goal is not to achieve a perfect model, but to prevent average performance from masking unequal clinical consequences. A model may be useful if it improves assessment access for underserved groups, triggers supportive rather than punitive responses and is monitored for drift. Conversely, a technically impressive model may be unacceptable if it increases coercion, misdiagnosis or neglect for specific groups. Equity must be judged by downstream clinical impact, not only by statistical symmetry.

**Clinical implementation framework for equity-by-design:** Equity-by-design requires a deliberate implementation pathway rather than a retrospective fairness paragraph added after model development. The first step is problem selection. Health systems should ask whether the proposed model addresses an inequity, a safety gap or an administrative preference. A tool that identifies patients for proactive outreach may reduce unmet need, whereas a tool that prioritizes scarce appointments without safeguards may simply ration care more efficiently. Problem selection should include people with lived experience, family representatives, clinicians from

community and acute settings, statisticians, privacy experts and leaders responsible for service quality.

The second step is data readiness. Development teams should map the full data-generating process before modelling. This map should include who enters the dataset, who is excluded, why records are missing, how diagnoses are assigned, how outcomes are observed, how language is documented and how prior service use may reflect access rather than need. A general lesson from health AI fairness research is that proxy outcomes can create inequitable allocation even when algorithms are technically accurate; a widely cited population-health algorithm underestimated illness burden in Black patients because healthcare cost was used as a proxy for need [11]. Similar proxy problems can occur in psychiatry when service use, medication history, emergency attendance or hospitalization are used as substitutes for clinical severity. Fairness in machine learning therefore requires explicit attention to causal meaning, not only predictive association [12].

The third step is model development with pre-specified equity endpoints. Before training begins, investigators should define protected and clinically relevant attributes, minimum subgroup sizes, intended estimands, fairness metrics, threshold-selection rules and conditions under which a model will be rejected. Development datasets should be partitioned to preserve temporal and site separation where possible. Hyperparameter tuning should not optimize only global AUC. Calibration, net benefit and subgroup error rates should be evaluated as co-primary performance domains. When a subgroup is too small for stable estimation, the limitation should trigger additional data collection, federated validation or cautious restricted use rather than silent exclusion.

The fourth step is clinical workflow integration. Algorithmic outputs should be embedded in a workflow that specifies who receives the alert, what action is expected, what uncertainty means and when human review is mandatory. Clinicians should see clear information about the model's intended use, development population, validation performance and known limitations. Alerts should support comprehensive assessment and care planning, not replace clinical judgement or patient narrative. Override decisions should be documented in a non-punitive manner so that the health system can learn when the model is misleading or when clinicians are ignoring useful information.

The fifth step is surveillance after deployment. Performance should be monitored at regular intervals and after major changes in coding practice, service pathways, population mix, model version, language model provider or clinical guideline. A drift dashboard should include total alert volume, action rates, outcome rates, calibration, false

positives, false negatives, adverse events and patient complaints by subgroup. Model retirement criteria should be specified in advance. If subgroup calibration worsens or alert burden becomes unequal, the model should be recalibrated, restricted or suspended. In this framework, equity is not a single validation result; it is a continuous safety obligation.

### Limitations

Several limitations of this review should be acknowledged. First, the search was restricted to English-language publications, which may have excluded relevant work from non-English-speaking settings where distinct patterns of algorithmic bias could be present. Second, as a structured narrative review rather than a systematic review with meta-analysis, the synthesis reflects expert judgement in selecting and weighting evidence; quantitative pooling was not possible due to heterogeneity in outcome definitions, algorithmic architectures and bias metrics across included studies. Third, the literature base is predominantly from high-income countries, particularly the United States and Western Europe, limiting generalizability to low- and middle-income settings where resource availability, healthcare infrastructure and patient demographics differ substantially. Fourth, published fairness evaluations remain sparse, meaning that absence of reported bias in the reviewed studies should not be interpreted as absence of bias. Fifth, the rapidly evolving nature of psychiatric AI means that some methodological standards and model architectures referenced here may have been superseded by the time of publication, and ongoing monitoring of new literature is recommended. Sixth, intersectional subgroup analyses were rarely available in the primary studies, so conclusions about compounded disparities remain largely inferential. Finally, the review did not include unpublished studies or preprints, which may bias the evidence base toward positive or statistically significant findings.

**Future Directions:** Future studies should move beyond retrospective model development toward prospective, equity-focused evaluation. First, psychiatric AI studies need larger and more diverse external validation cohorts with pre-specified subgroup and intersectional analyses. Second, diagnostic-label audits should become standard when models predict or use psychiatric diagnoses. Third, researchers should evaluate clinical utility by subgroup, not only overall accuracy. Fourth, language, speech and sensor studies should measure device, dialect, multilingual and disability-related performance. Fifth, deployment trials should record how clinicians act on model outputs and whether those actions differ across demographic groups.

There is also a need for methods that link fairness metrics to patient-centered outcomes. Many statistical fairness measures operate on observed labels, but observed labels may be affected by the prediction itself and by interventions triggered after prediction. Counterfactual and causal frameworks are therefore important for psychiatric AI systems. Researchers should ask not only whether an alert predicts hospitalization, but whether issuing the alert improves safety, autonomy and care experience for different groups. Without causal evaluation, a model may appear effective because it predicts the consequences of a biased system. Open science can support fairness by allowing independent audit, reproducibility and broader benchmarking [32]. However, privacy and stigma concerns in psychiatry require careful governance. De-identified data sharing, synthetic data, federated validation, secure enclaves and transparent model documentation may help balance reproducibility with confidentiality. International collaboration is essential because most psychiatric AI evidence remains concentrated in high-income settings, while the global burden of mental illness and suicide prevention need is far broader.

### Conclusion

Algorithmic bias in psychiatric risk assessment is a clinical safety issue, a statistical validity issue and an equity issue. Psychiatric algorithms do not operate on neutral data; they learn from systems shaped by differential access, documentation, diagnosis, surveillance and treatment. Evidence from suicide prediction, diagnostic-disparity research, language models, speech systems and smartphone sensing shows that pooled accuracy can conceal subgroup harm.

For diverse patient demographics, the most important question is not whether an algorithm is accurate on average, but whether its predictions mean the same thing, produce comparable benefits and avoid disproportionate harms across groups.

A journal-quality evaluation of these tools should include subgroup representation, label assessment, discrimination, calibration, threshold errors, clinical utility, intersectional analysis and post-deployment monitoring. Bias mitigation must combine better data, appropriate modelling, transparent thresholds, governance and community accountability. Until these requirements are met, psychiatric prediction models should be used cautiously, with human oversight and as supports for care rather than replacements for comprehensive clinical assessment.

**Table 1: Psychiatric algorithmic use cases, likely bias mechanisms and recommended equity audits**

| Use case                                             | Common data sources                                                      | Main bias mechanisms                                                                     | Recommended equity audit                                                                         |
|------------------------------------------------------|--------------------------------------------------------------------------|------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Suicide and self-harm prediction                     | EHR diagnoses, prior attempts, prescriptions, notes, emergency visits    | Outcome rarity, differential ascertainment, documentation intensity, access to follow-up | Subgroup AUC, calibration, PPV/NPV and decision-curve net benefit; evaluate alert consequences   |
| Psychosis or bipolar diagnostic prediction           | Diagnostic codes, medication history, admission records, clinician notes | Historical diagnostic disparity, coercive-care pathways, social proxies                  | Label audit; alternative outcome definitions; measurement invariance; subgroup calibration       |
| Depression or anxiety monitoring                     | Screening scales, patient portals, smartphone sensors, wearables         | Digital divide, socioeconomic proxies, language, device missingness                      | Missingness by subgroup; device and language performance; subgroup reliability analysis          |
| Speech-based assessment                              | Audio recordings, acoustic and prosodic features                         | Accent, dialect, recording quality, disability, sex and gender voice differences         | Subgroup performance; recording-condition sensitivity; fairness across language and accent       |
| LLM-supported diagnosis or treatment recommendations | Prompts, clinical summaries, generated advice, EHR context               | Training-data stereotypes, prompt sensitivity, demographic framing, hallucination        | Scenario audits; demographic perturbation tests; clinician and patient review of recommendations |

**Table 2: Minimum statistical and fairness estimands for diagnostic prediction systems**

| Estimand or metric      | What it asks                                                        | Why it matters in psychiatry                                           | Suggested reporting                                                                                         |
|-------------------------|---------------------------------------------------------------------|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Subgroup representation | Who was included and excluded?                                      | Under-representation can make pooled estimates misleading.             | Flow diagram and subgroup counts for training, validation and excluded records.                             |
| Discrimination          | Does the model rank higher-risk patients above lower-risk patients? | Useful for triage but insufficient for clinical probability estimates. | AUC or C-statistic with 95% CI by subgroup; precision-recall AUC for rare outcomes.                         |
| Calibration             | Does predicted risk match observed risk?                            | Guides urgency and intensity of clinical response.                     | Calibration plot, intercept, slope and observed/expected ratio by subgroup.                                 |
| Threshold error rates   | Who is missed and who is falsely alerted?                           | False negatives and false positives have different ethical harms.      | Sensitivity, specificity, PPV, NPV, FPR and FNR at every clinical threshold.                                |
| Clinical utility        | Who benefits after using the model?                                 | A model can be statistically accurate but clinically harmful.          | Decision-curve analysis and net benefit by subgroup.                                                        |
| Label validity          | Is the outcome measured equivalently?                               | Psychiatric labels may encode diagnostic disparity.                    | Chart audit, alternative definitions, quantitative bias analysis and measurement invariance where relevant. |

**Table 3: Evidence map of the reviewed literature**

| Evidence domain                       | Representative literature                                       | Main contribution to this review                                                                             |
|---------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Machine learning in mental health     | Shatte et al.; Le Glaz et al.; Topol                            | Rapid growth of ML/NLP applications and need for clinical validation [1-3].                                  |
| Suicide and self-harm prediction      | Seyedsalehi et al.; Spittal et al.; Coley et al.; Papini et al. | External validation gaps, limited utility at low prevalence and racial/ethnic performance disparities [4-7]. |
| Psychiatric diagnostic disparity      | Garb; Schwartz and Blankenship; McGuire and Miranda             | Historical diagnostic and service disparities create label and measurement bias [8-10].                      |
| Language and LLM systems              | Straw and Callison-Burch; Bouguettaya et al.; Omar et al.       | Language models can encode demographic bias in mental-health outputs and medical recommendations [13-15].    |
| Speech and digital phenotyping        | Yang et al.; Adler et al.; Park et al.                          | Sensor, speech, gender, race and socioeconomic subgroup reliability problems [16-18].                        |
| Clinical prediction and AI governance | Van Calster et al.; TRIPOD+AI; PROBAST+AI; WHO; FDA             | Provides calibration, reporting, risk-of-bias and lifecycle monitoring standards [19,24-31].                 |

**Table 4: Reporting checklist for risk stratification tools submitted to peer-reviewed journals**

| Reporting item          | Minimum content for psychiatric fairness evaluation                                                                                    |
|-------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Clinical purpose        | Intended decision, population, time horizon, setting and action triggered by the model.                                                |
| Data representativeness | Inclusion/exclusion flow, subgroup counts, site mix, language, device availability and missingness.                                    |
| Outcome definition      | How diagnosis or risk outcome was measured; whether routine codes, structured interviews, chart review or adjudication were used.      |
| Overall performance     | Discrimination, calibration, threshold metrics and clinical utility in the total validation sample.                                    |
| Subgroup performance    | The same metrics by race/ethnicity, sex/gender, age, socioeconomic status, language, disability and clinically relevant intersections. |
| Fairness criterion      | Which fairness goal was prioritized, why it was chosen and what trade-offs were accepted.                                              |
| Mitigation and updating | Any reweighting, recalibration, threshold changes, human review pathways and plans for drift monitoring.                               |
| Governance              | Consent, clinician training, override documentation, patient feedback, adverse-event monitoring and accountability.                    |

## Declarations

**Ethics approval and consent to participate:** Not applicable. This is a review article and does not involve recruitment of human participants or use of identifiable patient data.

**Consent for publication:** Not applicable.

**Availability of data and materials:** All sources cited in this manuscript are derived from published literature. Of the 80 numbered references, 72 were published between January 2014 and March 2026 and were formally included in the PRISMA synthesis. Eight seminal pre-2014 publications (references 10, 48, 49, 50, 52, 54, 55 and 77) are retained for contextual background only and were not included in the PRISMA counts. References 1–

32 are cited directly in the narrative text. References 33–80 constitute the extended evidence base and are presented at the end of this manuscript. Readers are directed to Table 3 (Evidence map) for the narrative domains they support.

**Competing interests:** The authors declare that they have no competing interests.

**Funding:** No external funding was received for this work.

**Author contributions:** Conceptualization, literature synthesis, drafting and critical revision: Jasmine (Conceptualization, Literature Synthesis, Drafting); Mahendra Kumar (Critical Revision, Methodology Review, Final Approval).

## References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7.
2. 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. doi:10.1017/S0033291719000151.
3. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. *J Med Internet Res*. 2021;23(5):e15708. doi:10.2196/15708.
4. Seyedsalehi A, Arias-Garcia S, MacKinnon D, et al. Prediction models for self-harm and suicide: a systematic review and critical appraisal. *BMC Med*. 2025;23:549. doi:10.1186/s12916-025-04367-6.
5. Spittal MJ, Guo XA, Kang L, et al. Machine learning algorithms and their predictive accuracy for suicide and self-harm: systematic review and meta-analysis. *PLoS Med*. 2025;22(9):e1004581. doi:10.1371/journal.pmed.1004581.
6. Coley RY, Johnson E, Simon GE, et al. Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatry*. 2021;78(7):726-734. doi:10.1001/jamapsychiatry.2021.0493.
7. Papini S, Pisner D, Shumake J, et al. Performance of a prediction model of suicide attempts across race and ethnicity. *JAMA Psychiatry*. 2023;80(4):399-400. doi:10.1001/jamapsychiatry.2022.5063.
8. Garb HN. Race bias and gender bias in the diagnosis of psychological disorders. *Clin Psychol Rev*. 2021;90:102087. doi:10.1016/j.cpr.2021.102087.
9. Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. *World J Psychiatry*. 2014;4(4):133-140. doi:10.5498/wjp.v4.i4.133.
10. McGuire TG, Miranda J. New evidence regarding racial and ethnic disparities in mental health: policy implications. *Health Aff (Millwood)*. 2008;27(2):393-403. doi:10.1377/hlthaff.27.2.393.
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. doi:10.1126/science.aax2342.
12. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872. doi:10.7326/M18-1990.
13. Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. *PLoS One*. 2020;15(12):e0240376. doi:10.1371/journal.pone.0240376.
14. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *NPJ Digit Med*. 2025;8:332. doi:10.1038/s41746-025-01746-4.
15. Omar M, Sorin V, Agbareia R, Apakama DU, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int J Equity Health*. 2025;24(1):57. doi:10.1186/s12939-025-02419-0.
16. Yang YS, Bent B, Hicks JL, Dunn J. Deconstructing demographic bias in speech-based machine learning models for digital health. *Front Digit Health*. 2024;6:1351637. doi:10.3389/fdgth.2024.1351637.
17. Adler DA, Stamatis CA, Meyerhoff J, Mohr DC, Wang F, Aranovich GJ, et al. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *NPJ Ment Health Res*. 2024;3:17. doi:10.1038/s44184-024-00057-y.
18. Park J, Arunachalam R, Silenzio V, Singh VK. Fairness in mobile phone-based mental health assessment algorithms: exploratory study. *JMIR Form Res*. 2022;6(6):e34366. doi:10.2196/34366.
19. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7.
20. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14:40. doi:10.1186/1471-2288-14-40.
21. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Cham: Springer; 2019.
22. Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153-163. doi:10.1089/big.2016.0047.
23. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: *Proceedings of the 8th Innovations in Theoretical Computer Science*

- Conference. 2017;43:1-43:23. doi:10.4230/LIPIcs.ITCS.2017.43.
24. Collins GS, Dhiman P, Navarro CLA, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
  25. Moons KGM, Wolff RF, Riley RD, et al. PROBAST+AI: updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
  26. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. doi:10.1038/s41591-020-1034-x.
  27. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7.
  28. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924-933. doi:10.1038/s41591-022-01772-9.
  29. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: World Health Organization; 2021.
  30. World Health Organization. Regulatory Considerations on Artificial Intelligence for Health. Geneva: World Health Organization; 2023.
  31. US Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan. Silver Spring, MD: FDA; 2021.
  32. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)*. 2021;2(10):100347. doi:10.1016/j.patter.2021.100347.
  33. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243.
  34. Luxton DD. Artificial intelligence in psychological practice: current and future applications and implications. *Prof Psychol Res Pr*. 2014;45(5):332-339.
  35. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(3):223-230.
  36. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol*. 2018;14:91-118.
  37. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243-250.
  38. Janssen RJ, Mourao-Miranda J, Schnack HG. Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2018;3(9):798-808.
  39. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. *Curr Psychiatry Rep*. 2019;21(11):116.
  40. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Sci*. 2017;6:15.
  41. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci*. 2017;5(3):457-469.
  42. Kessler RC, Warner CH, Ivany C, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers. *JAMA Psychiatry*. 2015;72(1):49-57.
  43. Large M, Kaneson M, Myles N, et al. Meta-analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients. *PLoS One*. 2016;11(6):e0156322.
  44. Simon GE, Johnson E, Lawrence JM, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *Am J Psychiatry*. 2018;175(10):951-960.
  45. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry*. 2017;174(2):154-162.
  46. Turecki G, Brent DA, Gunnell D, et al. Suicide and suicide risk. *Nat Rev Dis Primers*. 2019;5(1):74.
  47. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. 2017;143(2):187-232.
  48. Brent DA, Mann JJ. Family genetic studies, suicide, and suicidal behavior. *Am J Med Genet C Semin Med Genet*. 2005;133C(1):13-24.
  49. Cook BL, Barry CL, Busch SH. Racial/ethnic disparity trends in children's mental health care access and expenditures from 2002 to 2007. *Health Serv Res*. 2013;48(1):129-149.
  50. Williams DR, Neighbors HW, Jackson JS. Racial/ethnic discrimination and health:

- findings from community studies. *Am J Public Health*. 2003;93(2):200-208.
51. Paradies Y, Ben J, Denson N, et al. Racism as a determinant of health: a systematic review and meta-analysis. *PLoS One*. 2015;10(9):e0138511.
  52. Anglin DM, Link BG, Phelan JC. Racial differences in stigmatizing attitudes toward people with mental illness. *Psychiatr Serv*. 2006;57(6):857-862.
  53. Alvarez K, Fillbrunn M, Green JG, et al. Race/ethnicity, nativity, and lifetime risk of mental disorders in US adults. *Soc Psychiatry Psychiatr Epidemiol*. 2019;54(5):553-565.
  54. Breslau J, Aguilar-Gaxiola S, Kendler KS, et al. Specifying race-ethnic differences in risk for psychiatric disorder in a USA national sample. *Psychol Med*. 2006;36(1):57-68.
  55. Miranda J, McGuire TG, Williams DR, Wang P. Mental health in the context of health disparities. *Am J Psychiatry*. 2008;165(9):1102-1108.
  56. Substance Abuse and Mental Health Services Administration. Key Substance Use and Mental Health Indicators in the United States: Results from the 2022 National Survey on Drug Use and Health. Rockville, MD: SAMHSA; 2023.
  57. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Ment Health*. 2016;3(2):e16.
  58. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA*. 2017;318(13):1215-1216.
  59. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*. 2016;41(7):1691-1696.
  60. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol*. 2017;13:23-47.
  61. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med*. 2020;102:101753.
  62. Abd-Alrazaq A, Alajlani M, Ali N, et al. Perceptions and opinions of patients about mental health chatbots: scoping review. *J Med Internet Res*. 2021;23(1):e17828.
  63. Firth J, Torous J, Nicholas J, et al. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *J Affect Disord*. 2017;218:15-22.
  64. Seyednasrollah F, Das S, Zanuso S, et al. Using deep learning and wearable device data to identify psychiatric conditions. *NPJ Digit Med*. 2021;4(1):13.
  65. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178(11):1544-1547.
  66. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167-E179.
  67. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-2378.
  68. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337-1340.
  69. Char DS, Shah NH, Magnus D. Implementing machine learning in health care: addressing ethical challenges. *N Engl J Med*. 2018;378(11):981-983.
  70. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med*. 2018;15(11):e1002689.
  71. Topol EJ. Welcoming new medical residents in the era of AI. *Science*. 2019;364(6439):437.
  72. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act). *Off J Eur Union*. 2024;L 2024/1689.
  73. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37-43.
  74. NHS England. Evidence Standards Framework for Digital Health Technologies. London: NHS England; 2023.
  75. Hasin DS, Sarvet AL, Meyers JL, et al. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*. 2018;75(4):336-346.
  76. Olfson M, Cosgrove CM, Altekruze SF, Wall MM, Blanco C. Deaths of despair: adults at high risk for death by suicide, poisoning, or chronic liver disease in the US. *Health Aff (Millwood)*. 2021;40(3):505-515.
  77. Hasin DS, Goodwin RD, Stinson FS, Grant BF. Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. *Arch Gen Psychiatry*. 2005;62(10):1097-1106.
  78. Alegria M, NeMoyer A, Bague IF, Wang Y, Alvarez K. Social determinants of mental health: where we are and where we need to go. *Curr Psychiatry Rep*. 2018;20(11):95.
  79. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. *Public Health*. 2015;129(6):611-620.

80. Walker ER, Druss BG. A public health perspective on mental and medical comorbidity. JAMA. 2016;316(10):1104-1105.