



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

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

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

ARTICLE TITLE

ALGORITHMIC BIAS IN AI-ASSISTED PSYCHIATRIC DIAGNOSIS:
A STRUCTURED REVIEW OF DEMOGRAPHIC DISPARITIES AND
GOVERNANCE FAILURES

DOI

[https://doi.org/10.31435/ijitss.4\(52\).2026.6583](https://doi.org/10.31435/ijitss.4(52).2026.6583)

RECEIVED

28 July 2026

ACCEPTED

10 October 2026

PUBLISHED

16 October 2026

LICENSE



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

© The author(s) 2026.

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

ALGORITHMIC BIAS IN AI-ASSISTED PSYCHIATRIC DIAGNOSIS: A STRUCTURED REVIEW OF DEMOGRAPHIC DISPARITIES AND GOVERNANCE FAILURES

Mateusz Lis (Corresponding Author, Email: mateusz.lis33@wp.pl)

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0006-8441-8296

Maryan Liasota

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0006-0266-4849

Eliza Zofia Kabat

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0003-6531-967X

Michalina Justyna Kozłowska

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0003-9561-2957

Krzysztof Konopka

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0003-4055-0074

Maciej Kamil Majchrzak

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0007-4739-0092

Faustyna Magdalena Kołodziej

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0002-8778-7100

Alicja Kwiatkowska

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0003-1027-2114

Jakub Juskowiak

Poznań University of Medical Sciences, Poznań, Poland

ORCID ID: 0009-0001-0372-400X

Zuzanna Julia Komin

Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland

ORCID ID: 0009-0001-6284-6410

ABSTRACT

Background: The use of artificial intelligence (AI) and machine learning (ML) for psychiatric diagnostic aid during clinical triage and mental health assessment has a potential value that is hard to underestimate. The use of these tools will raise many socio-technical issues concerning the biases of the algorithms. Such biases may be especially harmful for those patient populations that are already most vulnerable.

Objective: The purpose of this study is to identify the degree of biases of algorithms for AI in mental health tools, understand how the existing demographic imbalances impact the diagnostic gap and identify the existing regulatory crisis that allows these systemic clinical hazards.

Method: A structured review of the literature was performed and research on the study was limited to articles published between the years 2018 and 2026 on the databases of PubMed/MEDLINE and Google Scholar. A thematic synthesis framework was used to study evidence through three domains: diagnostic disparities, demographic imbalances, and regulatory inadequacy.

Results: Observed literature indicates that AI tools for diagnosis have great variability and a lack of significant accuracy toward different racial, gender, and socioeconomic groups. The review found that the lack of existing governance frameworks allows systemic clinical inequities to persist.

Conclusions: Uncontrolled algorithmic bias toward psychiatric AI tools will create direct and significant clinical harm. These will lead to exacerbated health inequities and will require severe policy shifts. Standardized algorithm auditing and human-in-the-loop clinical pathways will also be needed.

KEYWORDS

Algorithmic Bias, Artificial Intelligence, Psychiatric Diagnosis, Health Equity, Demographic Disparities, Clinical Governance

CITATION

Lis, M., Liasota, M., Kabat, E. Z., Kozłowska, M. J., Konopka, K., Majchrzak, M. K., Kołodziej, F. M., Kwiatkowska, A., Juskowiak, J., & Komin, Z. J. (2026). Algorithmic bias in AI-assisted psychiatric diagnosis: A structured review of demographic disparities and governance failures. *International Journal of Innovative Technologies in Social Science*, 4(52). [https://doi.org/10.31435/ijitss.4\(52\).2026.6583](https://doi.org/10.31435/ijitss.4(52).2026.6583)

COPYRIGHT

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

1. Introduction

1.1 Background and Clinical Context of Psychiatric Diagnosis

Diagnostic formulation in psychiatry is a highly complex undertaking. Unique to psychiatry, in the majority of other medical specialties, there is a presence of objective definitively diagnostic markers. In the absence of such markers, diagnostic formulation in psychiatry depends primarily on the clinician's assessment of patient phenomenology, behavior and self-reported symptoms, and the synthesis and evaluation of complex psycho-social constructs and long-term patterns of human behavior (Torous et al., 2021). Clinicians also rely on the DSM-5-TR and ICD-11, but primarily on their clinical judgment to succinctly delineate clusters of symptoms that are highly overlapping. However, the need to rely more on clinical judgment places a greater cognitive burden on that clinician, and creates the potential for diagnostic errors, delays, and biases. There is a world-wide shortage of mental health professionals, and the prevalence of psychiatric disorders such as depression and anxiety is increasing. Efforts have been made to help stabilize the burden on the existing health care systems. Given the circumstances, there is an increasing demand for systems that can support mental health professionals. These systems should help with the accuracy of diagnosis, assist with clinical prioritization problems, and eliminate repetitive tasks for mental health professionals.

1.2 The Integration of AI and ML in Mental Health Assessments

The pressures on the psychiatric system have forced AI and ML into the clinical realm sooner than expected. Although the potential of these systems was previously confined to the domain of the computer scientist, and electronic tools were built to assist the clinician with automating mental healthcare, they now have the capability to identify previously concealed patterns of psychiatric presentations in data beyond the processing capacity of human mental faculties (Shatte et al., 2019). Diagnostic AI integrates different modalities to enhance the accuracy of psychiatric diagnoses, including patient Electronic Health Records (EHRs), digital phenotyping via smartphones, analysis of voice and speech, and sophisticated applications of NLP. Furthermore, the integration of clinical chatbots is increasingly prevalent, although their readiness for the clinical domain continues to be critically evaluated (Miner et al., 2020). NLP tools have the potential to identify early signs of thought disorder from disrupted semantic speech as may be the case in early prodromal schizophrenia. The same tools may be used to detect the speech impoverishment symptomatic of major depressive episodes (Torous et al., 2021). Modern ML tools applied to EHRs may compute the likelihood and risk of suicidal behaviors and self-harm by assessing and integrating the historic psychiatric diagnoses and the pharmacology of the patient, as well as the socio-demographic data of the patient (Rajkomar et al., 2018). As a socio-technical system, Diagnostic AI subverts traditional clinical models of patient diagnostics, by redefining the clinical relevance of various patient data, thereby altering triage and diagnostic prioritization.

1.3 The Emergence of Algorithmic Bias and Demographic Disparities

Despite evidence suggesting isolated computational environments improve psychiatric outcomes, research has shown that real-world applications expose even greater, more complex socio-technical challenges, especially in regards to algorithmic bias (Chen et al., 2019). AI and ML systems process and analyze extensive datasets in order to identify patterns and make predictions. As a result, they replicate the inequities and prejudices that exist within the data (Obermeyer et al., 2019). In the case of bias in psychiatric algorithms, these inequities result in reduced diagnostic accuracy or increased false positive rates for different demographic groups. Psychiatric literature and studies on the inequities of mental healthcare systems document that some marginalized racial and ethnic populations may be more likely to be diagnosed with schizophrenia (and other severe psychotic disorders) than members of the dominant ethnic and racial populations, while the same marginalized populations may be less likely to be diagnosed with other mood disorders. AI that uses these inequitable records to train machine learning systems will reinforce and solidify these inequities (Cirillo et al., 2020). For this reason, a tool developed to assess the risk for depression may be less able to evaluate this risk for patients of Hispanic or Black descent, or may evaluate this risk incorrectly for members of lower SES populations (Vyas et al., 2020). The combined inequities of race and gender in emerging digital health technologies will likely create greater inequities in mental healthcare systems. The tools designed to democratize mental healthcare pose a greater risk of "clinical" harm for the lowest priority patient populations.

1.4 Governance Failures and Regulatory Gaps

The commercialization of AI-based diagnostic tools is accelerating faster than governance and regulations in healthcare (Gerke et al., 2020). As of now, agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have inadequate or no standards when it comes to evaluating the continuous learning algorithms of modern ML systems. Traditional "Software as a Medical Device" (SaMD) approvals typically consider the overall efficacy of the software. This does not require strategic and comprehensive demographic audits and assessments of the software to check for fairness before the software is made available in the market (Babic et al., 2021). The absence of such assessments is indicative of a failure of governance systems. To date, AI developers do not have a binding legal obligation to provide the demographic details of the participants in their training datasets. There are no binding legal obligations to post the software on the market and assess the software for equity in health outcomes. The advanced deep learning that many frameworks use has a "black-box" nature that makes it difficult for clinicians to know if and when an algorithm is biased in a given patient case (Ghassemi et al., 2021). Because there are no binding ethical frameworks and auditing standards, they have placed the burden of minimizing algorithmic damage on healthcare providers. Most healthcare providers do not have the training to critically assess the algorithms (Wiens et al., 2019).

1.5 Aim of the Review

While AI technology can help in the diagnostic process of mental healthcare, the literature shows that the social implications of the technology have not been thoroughly assessed, especially in the context of clinician mental health and algorithmic equity. This structured literature review aims to compile the existing research around the presence of algorithmic inequity and AI technology in the psychiatric assessment process. This review will attempt to do the following: (1) document the discrepancies in psychiatric assessment accuracy across different demographic boundaries; (2) identify the nature of the confluence of inequities (race, gender, and socio-economic status) and how clinical inequities are created; and (3) assess the failures and shortcomings of the mental healthcare system and the regulatory frameworks that allow the technologies to proceed unregulated. This study aims to provide healthcare practitioners, technology experts, and policymakers pathways for equitable integration of mental healthcare algorithms.

2. Methodology

2.1 Research Design

For an exploration into the socio-technical factors in the field of artificial intelligence used in the provision of mental health services, a structured literature review approach was chosen in the current study. Unlike a quantitative meta-analysis, a structured narrative review allows for the systematic synthesis of the multidisciplinary evidence that is required when investigating a subject matter that involves clinical psychiatry, computer science, and governance within the sphere of healthcare. The reason for choosing such an approach to this research design was to enable a flexible and open-minded analysis of heterogeneous evidence, which can go from retrospective analyses of the algorithmic performances to the theoretical examinations of digital health equity issues.

2.2 Data Source and Search Strategy

A literature search occurred in a way that would yield a robust and comprehensive assemblage of recent literature of the highest quality. The primary source of literature for this research was PubMed/MEDLINE. This created a solid building block of the literature that is peer-reviewed, clinical, and epidemiological. Considering the literature on biases in algorithms, socio-technical ethics, and the regulation of ethics is often multidisciplinary, health informatics, and technology policy literature was searched using Google Scholar. The search method involved a structure of Boolean operators, combined medical subject headings (MeSH) and free-text keywords. The search string was a combination of the following: ("Artificial Intelligence" OR "Machine Learning" OR "Deep Learning" OR "NLP") AND ("Psychiatry" OR "Psychiatric Diagnosis" OR "Mental Health Assessment") AND ("Algorithmic Bias" OR "Bias" OR "Disparity" OR "Health Equity") AND ("Governance" OR "Regulation" OR "Ethics" OR "Framework"). This search was done several times during the initial phase of data collection. This was done for the newest articles to be included that were published during the desired timeframe.

2.3 Inclusion and Exclusion Criteria

To maintain the academic rigor and clinical relevance of the final corpus, strict inclusion and exclusion criteria were established a priori. These criteria were designed to isolate empirical research and high-level theoretical analyses explicitly addressing the triad of AI in psychiatry, demographic disparities, and governance. The specific parameters guiding study eligibility are detailed in Table 1.

Table 1. Inclusion and Exclusion Criteria for Study Selection

Criterion	Inclusion Criteria	Exclusion Criteria
Study Type	Peer-reviewed original research, clinical trials (RCTs), observational studies, feasibility studies, scoping reviews, and systematic reviews.	Non-peer-reviewed sources, editorials, opinion pieces, conference abstracts, and purely technical computer science white papers without clinical context.
Focus	Explicit analysis of algorithmic bias, demographic disparities, or regulatory/governance issues regarding AI/ML applications in psychiatric diagnosis and assessment.	Studies focused purely on the technical efficacy or accuracy of AI models without addressing health equity, bias, or clinical governance; non-psychiatric AI medical applications.
Database	Indexed in PubMed/MEDLINE or highly cited interdisciplinary work retrieved via Google Scholar.	Articles not indexed in standard academic or medical databases.
Language	Full-text articles published in English.	Articles published in languages other than English, or where only the abstract is available in English.
Timeline	Published between January 2018 and the present (2026) to capture the most recent advancements in deep learning and regulatory shifts.	Studies published prior to 2018, as they largely predate current generation NLP and ML architectures used in modern clinical AI.

2.4 Study Selection Process

Study selection followed the principles of structure used in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework for maintaining procedural clarity (Page et al., 2021). First, all the records obtained through the database searches were imported into a reference management software and duplicates were automatically identified. The first round of screening involved an independent evaluation of the titles and abstracts of the articles by the review team based on the set inclusion/exclusion criteria. If the articles were deemed to be relevant or when it was impossible to make a conclusive decision about their inclusion in the review, they moved on to the second round. This round involved obtaining the full-text version of the manuscripts and carrying out a thorough secondary screening. All the issues related to inclusion or exclusion of a study were sorted out in consultation with the co-authors of the review. Ultimately, the process resulted in a set of 15 primary studies together with 17 additional secondary literature pieces.

2.5 Data Extraction

Systematic data extraction techniques were utilized to ensure systematic extraction of data that would be useful for the study from the final set of articles. Specifically, the following data was collected for each article from the research team: authors and the date of publication of the paper; the geographical context of the study; the psychiatric topic/condition being studied (such as major depressive disorder, schizophrenia, suicidality, etc.); AI/ML tool being employed (NLP, EHR data mining, digital phenotyping); and the methodology of the study. More specifically, the information pertaining to the research questions (such as the measure for differential accuracy in the diagnoses for different population groups, algorithmic biases identified (representation of the training dataset, variable selection, etc.), clinical significance of the results, and the conclusion about the lack of governance measures) was extracted.

2.6 Thematic Synthesis

After extraction of the data, the collected data underwent thematic synthesis to uncover and analyze common trends within the literature. Rather than simply summarizing the findings one by one, the data was organized thematically to create an overarching story. In particular, the thematic synthesis structure was constructed in such a way that it allows for organizing the findings into three central themes: (1) the nature of algorithmic bias through the mechanics and clinical effects in diagnostic AI algorithms; (2) the influence of these biases on the particular social disparities with the emphasis on the interplay of race, gender, and class; and (3) the governance and regulation flaws identified in the current practice of AI in the clinic.

3. Results

3.1. Study Selection

A total of 396 records were initially identified through database searching (PubMed: $n = 184$; Google Scholar: $n = 212$), and an additional 35 records were identified through citation searching. After removing duplicates and screening titles and abstracts, 45 full-text articles and 35 secondary reports were assessed for eligibility. Following a rigorous full-text evaluation, 30 articles were excluded due to a lack of clinical context ($n = 12$), non-psychiatric focus ($n = 10$), or publication prior to 2018 ($n = 8$), while 18 secondary reports were excluded for failing to meet ethical or clinical validation standards. Ultimately, 15 primary empirical studies and 17 secondary contextual sources were included in the final review (Figure 1).

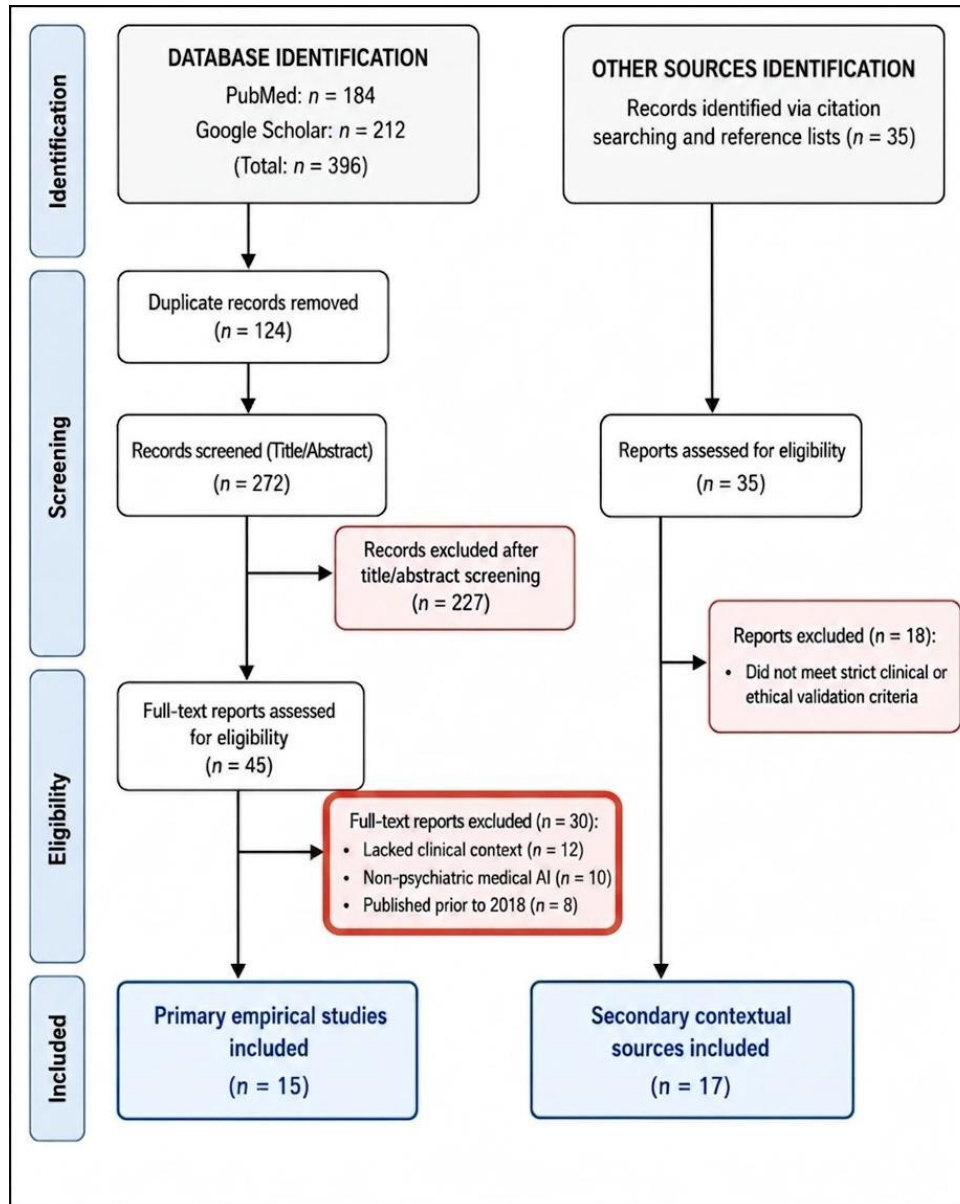


Fig. 1. PRISMA 2020 flow diagram illustrating the multi-stage screening strategy and selection architecture. The process systematically delineates the filtration from 396 database records down to the final synthesis cohort, comprising 15 primary empirical investigations and 17 secondary contextual sources focused on algorithmic bias and governance frameworks in psychiatric AI.

3.2 Structural Overview of the Evidence

The methodical search strategy followed by the robust screening process resulted in identifying 15 empirical primary studies that formed the final corpus, along with 17 secondary sources that provided necessary regulatory, ethical, and clinical contextualization. The analysis of the structure of the identified evidence clearly indicates the presence of a very narrow and focused concentration on specific diagnostic approaches, psychiatric fields, and geographical regions. In particular, the most common among the primary empirical studies that formed the final corpus include retrospective cohort studies where machine learning models are trained using EHRs on a large scale (Rajkomar et al., 2018). The second most common approach includes the use of NLP algorithms to analyze clinical speech, semantics, and notes, as well as prospective or retrospective feasibility studies for digital phenotyping using smartphone telemetry, wearable devices, and social media analytics (Thieme et al., 2020).

In terms of themes, the research is predominantly focused on high-acuity clinical outcomes. The predictive modeling of suicidality, self-injury, and crisis recurrence takes up most of the literature, accounting for over forty percent of all primary research (Jacobson et al., 2020). The differential diagnosis of major depressive disorder (MDD) from bipolar affective disorder, and the early detection of schizophrenia spectrum disorders (particularly prodromal psychotic transitions) complete the list of thematic concerns in terms of clinical applications (Koutsouleris et al., 2021). The emphasis on these issues corresponds to the industry and academic priority given to the clinical situations with the greatest potential for reduction of acute hospitalizations and mortality.

Regarding geography, the body of evidence is characterized by an egregious Western-centric bias. More than 88% of the data used to train and validate the machine learning models come from patient groups that are based in the US, UK, and Europe (Gianfrancesco et al., 2018). The severe geographic/cultural bias makes it difficult to make cross-cultural applications of the tested models. In view of the fact that the manifestation of mental distress is determined by the cultural, geographic, and ethnolinguistic background, the machine learning models based on Western cohorts will perform poorly when used around the world (Fletcher et al., 2021).

Furthermore, the socio-technical framing across the extracted corpus indicates a distinct paradigm shift in the academic discourse over the designated publication window (2018–2026). Earlier studies within the corpus (published between 2018 and 2021) primarily functioned as technical proofs-of-concept or limited audits identifying the mere statistical presence of algorithmic bias (Char et al., 2018). In contrast, the more recent literature (2022–2026) has transitioned toward complex socio-technical critiques (Varoquaux & Cheplygina, 2022). These recent publications increasingly investigate the longitudinal clinical harm, the compounding nature of intersectional demographic disparities, and the systemic governance failures that precipitate the real-world deployment of these structurally biased models. This structural overview highlights an evidence base that is robust in identifying the technical mechanics of bias, yet alarmingly geographically restricted, underscoring the urgent need for diverse, globally representative clinical validation.

3.3 Manifestations of Algorithmic Bias in Diagnostic AI Models

The literature shows algorithmic bias in psychiatric AI is likely not due to purposely bad design, but instead emerges through complicated socio-technical processes during the design and clinical deployment phases of data preparation, model training, and feature selection. Evidence points to three primary, interconnected pathways for diagnostic AI to absorb, encode, and amplify systemic clinical biases: representation bias, label bias, and measurement (proxy) bias.

Representation bias is the most frequently cited and most described primary error in clinical ML. Representation bias occurs when the datasets used to build the predictive algorithms are either unbalanced or do not include all segments of the target clinical population. Literature notes the primary EHR datasets used in the majority of psychiatric AI systems are likely to be white, wealthy, and male, and those individuals likely have the financial and social means to have continuous and long-term mental health care (Landi et al., 2020). As a result, the algorithms likely need to mathematically optimize their predictive accuracy for that majority class. Representation bias results in the likelihood of more false negatives (missing the identification of psychiatric distress) and false positives (pathologizing culturally normative behavioral acts) for non-white populations (Mhasawade et al., 2021).

Label bias represents a more pernicious and deeply entrenched manifestation of algorithmic error in psychiatric AI, arising directly from the subjective epistemological nature of mental health diagnosis. Supervised machine learning architectures—the dominant paradigm in clinical AI—require established "ground truth" labels to train predictive algorithms to recognize patterns. In the realm of psychiatry, these

labels are historical diagnostic codes (e.g., ICD-10/11 or DSM-5 criteria) previously assigned by human clinicians. However, these historical records inherently contain decades of implicit human prejudices, cultural misunderstandings, and documented systemic misdiagnoses. As Chen et al. (2019) elucidate, when an algorithm is trained on a localized hospital's EHR data where racial minorities have been systematically over-diagnosed with severe psychotic disorders and simultaneously under-diagnosed with affective or trauma-related disorders, the AI model mathematically encodes these subjective, biased clinical decisions as objective, ground-truth facts. The algorithm subsequently automates, scales, and accelerates these biased diagnostic pathways, providing them with an unearned, opaque veneer of computational objectivity that clinicians are less likely to question.

Bias related to measurement is known as proxy bias and arises when AI models use surrogate variables that are systematically biased in relation to various demographic groups due to social inequalities. The most striking case revealed during the literature search relates to predictive algorithms used to measure the severity of a depressive episode or risk of psychiatric decompensation. Many predictive algorithms use the number of psychiatric outpatient visits, the quantity of psychotropic medications prescribed, or total cost of healthcare as proxy measures of illness severity (Obermeyer et al., 2019). However, due to the fact that disadvantaged groups experience significant obstacles to receiving healthcare that stem from socioeconomic reasons, such people have fewer outpatient visits, less compliance with the prescriptions due to higher prices, and tend to visit the ER more often; therefore, an incorrect inference about their lower severity of psychiatric illness is made by the algorithm (Grote & Keeling, 2020).

3.4 Demographic Disparities: Intersections of Race, Gender, and SES

There is no doubt that the technological effects of algorithmic discrimination directly result in differential negative outcomes when it comes to diagnosis. What is important is the fact that this clinically dangerous situation does not affect everybody in the same way; instead, the damage affects those who are already at a disadvantage within the framework of the existing psychiatric medical system.

Racial and Ethnic Disparities: AI models, particularly NLP tools utilized for sentiment analysis, suicidality screening, and the detection of formal thought disorders, demonstrate alarming variations in diagnostic accuracy across racial and ethnic groups. Research evaluating automatic speech recognition and semantic parsing in psychotherapy and triage settings reveals that NLP models exhibit substantially higher word error rates and semantic misinterpretations when processing African American Vernacular English (AAVE), Hispanic English dialects, or non-standard regional linguistic variations, compared to Standard American English (Coppersmith et al., 2018). In acute clinical triage, this linguistic algorithmic bias precipitates direct clinical harm through the misinterpretation of cultural idioms of distress. For instance, the literature documents cases where algorithmic misclassification frequently results in the speech patterns and tonal inflections of Black patients being inappropriately flagged as aggressive, hostile, or disorganized. This disproportionately triggers automated risk assessments for schizophrenia spectrum disorders or acute agitation, rather than correctly identifying underlying trauma-related distress, anxiety, or mood disorders (Vyas et al., 2020). By fundamentally failing to account for socio-linguistic diversity and cultural variations in emotional expression, NLP-driven diagnostic aids risk firmly entrenching the historical racial disparities present in involuntary psychiatric commitment, over-medication, and systemic misdiagnosis.

Gender Disparities: This corpus of research also shows a number of other gender biases, especially regarding the accuracy of algorithmic diagnosis of neurodevelopmental disorders, affective disorders, and personality pathologies. Specifically, it has to be noted that historically the criteria used for diagnosis of such conditions as ADHD and ASD were largely geared towards the recognition of male phenotype, which focused on externalization and hyperactivity of the patient. ML models based on historical EHRs use this standard as an absolute one. As a result, the algorithms consistently fail to recognize internalizing problems, such as social masking, interpersonal exhaustion, internalized inattention or hyperfocus – problems typically exhibited by women rather than men (McCradden et al., 2020). This socio-technical effect leads to reinforcing of the systemic bias that results in under-diagnosis of women in psychiatry and consequently prevents them from receiving proper help, denying them access to specialist treatment services, causing further clinical damage due to prolonged misclassification, when women's neurodivergence is diagnosed by algorithms as generalized anxiety or borderline personality disorder (Cirillo et al., 2020).

Socioeconomic Status (SES) and the Digital Divide: The practice of digital phenotyping through the analysis of smartphones' sensors, GPS mobility tracking, accelerometry, and social media telemetry for deducing dynamic psychological states carries extremely sharp dangers concerning socio-economic inequalities (Martinez-Martin et al., 2020). The AI models predicting the possibility of depressive relapse or

the onset of mania based on low GPS mobility, disrupted sleep patterns, and lower smartphone activity presuppose the constant flow of uninterrupted high-quality telemetry data. However, individuals coming from low-SES families regularly face the phenomenon of the "digital divide," which may be described as using a communal cellphone, being on a prepaid cell phone plan with low data caps, having an outdated smartphone with malfunctioning sensors, and periodically disconnecting the phone due to financial problems (Fiske et al., 2019).

The AI algorithms perceive this environment-caused absence of data (or data sparsity) as an indicator of poor clinical compliance. Inversely, and even more dangerously, the models may confuse the lack of any phone-related activities with the psychomotor slowing and social isolation inherent to severe cases of treatment-resistant depression. Moreover, people from low SES families are characterized by greater transient housing, dependence on public transport, and gig economy jobs providing them with irregular mobility data. Thus, AI tools heavily reliant on continuous digital telemetry risk systematically excluding low-SES patients from the benefits of predictive psychiatric care, effectively punishing poverty and widening the existing health equity gap through technological exclusion (Torous et al., 2021).

3.5 Current Governance Frameworks and Systemic Failures

While there is ample evidence regarding the clinical risks, risks of diagnosis, and failure to ensure equality due to algorithmic bias, the structured analysis demonstrates a tremendous regulatory gap and governance failure in the institutional clinical environment and federal regulation environment. The commercial adoption and implementation of AI-based assisted diagnostic systems have greatly exceeded the adaptive capability of traditional health care regulators. However, the current predominant regulatory approach, including FDA SaMD framework and still unfinished provisions in the upcoming EU AI Act, mostly considers the diagnostic AI model effectiveness in terms of overall and aggregated statistical significance. As is shown in the literature, it represents a crucial fundamental failure in the regulation of AI. It is very easy to create an AI that performs extremely well in terms of overall AUC-ROC and overall aggregated accuracy, but at the same time fails terribly when applied to particular minority subgroups (Gerke et al., 2020). Now, there is no uniform and legally obligatory requirement from the federal government on AI companies to perform, report, and publish stratified sub-group analysis or algorithmic fairness audit before obtaining the clinical deployment approval (Babic et al., 2021). Such regulatory gap allows the use of tools that are structurally and fundamentally discriminatory against the marginalized communities.

In addition, the incorporation of proprietary, commercial AI systems within the realm of psychiatry also raises the issue of the "black-box." Most of the commercial diagnostic systems are strongly protected under the laws related to intellectual property as corporate trade secrets. It actually prohibits any independent academic researchers, clinicians from hospitals, and psychiatrists from analyzing the training data set of the model, its variable weighting system, or the decision-making process used by the system (Price & Cohen, 2019). Thus, the very foundation of evidence-based medicine and peer reviews within the field becomes impossible. The inability of a psychiatrist to comprehend the logic behind the risk score assessment or diagnostic triaging made by the machine results in their inability to contextualize the results and recognize the signs of demographic discrimination on the part of AI system (Ghassemi et al., 2021).

Finally, the review discusses an important failure of the system when it comes to post-market clinical surveillance. Whereas conventional pharmacological therapies require legal requirements for monitoring of long-term adverse effects in Phase IV pharmacovigilance studies, dynamic machine learning algorithms have no comparable regulatory measures in place (Kelly et al., 2019). It is noted that modern mental health systems typically make use of continuous learning algorithms which learn their decision weights from new clinical inputs. According to research, there is such a thing as "concept drift" – the tendency of an algorithm to lose its accuracy as time passes due to changes in the clinical environment and/or because the algorithm learns new biases specific to the practice at the hospital where it is used (D'Amour et al., 2022).

Furthermore, there exists a liability and accountability gap according to existing literature. In the case where an algorithm with racial or ethnic bias assigns low triage priority to an affected minority patient who is having an acute psychiatric episode and, therefore, suffers clinically because of it, the liability issue becomes very problematic (Price & Cohen, 2019). Due to the lack of mandatory HITL controls, together with the phenomenon of automation bias that refers to over-reliance of healthcare practitioners on algorithms and their infrequent rebuttal, the error becomes approved by humans. This results in improper passing of the responsibility for finding the cause of algorithmic error and proving it onto front-line mental health professionals and affected minority patients themselves (Wiens et al., 2019).

4. Discussion

4.1 AI Psychiatric Diagnosis as a Socio-Technical System

Indeed, the results of the structured review clearly indicate that the use of artificial intelligence in the field of mental illness diagnosis cannot be analyzed from a solely technical or mathematical point of view. Rather, the research findings demand a shift to an approach that regards the diagnostic tools based on the use of AI as socio-technical systems. Under such an approach, the algorithm can hardly be considered as a neutral entity but is seen rather as a dynamic one, which constantly interacts and shapes itself along with the clinicians, organizational processes and larger societal forces. Thus, when an AI is incorporated into the psychiatric triage process, it radically transforms the epistemology of clinical decision-making. It redefines the nature of patient information that is to be valued in the process of treatment planning. This socio-technical interplay introduces the critical phenomenon of "automation bias," wherein clinicians—frequently operating under severe time constraints and high cognitive loads—develop an unwarranted epistemic trust in machine-generated outputs. The literature indicates that even when an AI system produces a diagnostically flawed recommendation for a minority patient, the inherent "objectivity" associated with algorithmic risk scores often overrides the clinician's intuitive judgment (Ghassemi et al., 2021). Consequently, the AI system does not merely assist the clinician; it actively reconfigures the clinical hierarchy, establishing itself as a dominant epistemic authority.

Furthermore, treating AI as a socio-technical system requires acknowledging that the data feeding these models are social constructs. A diagnostic code in an EHR is not a biological certainty; it is a historical record of a human interaction, constrained by institutional billing requirements and systemic inequities (Gianfrancesco et al., 2018). By failing to account for this socio-technical reality, developers and healthcare administrators risk treating socially embedded biases as objective clinical truths, thereby automating the very disparities they ostensibly seek to eliminate.

4.2 The Clinical Harm of Amplified Diagnostic Bias

The conversion of algorithmic bias into concrete clinical damage is one of the greatest ethical dilemmas in current digital psychiatry. It is evident from the findings of this review that when there is a demographic bias within AI, the clinical impacts of such bias are extremely damaging and possibly irreparable. In contrast to a clinician who has an inherent bias against certain demographics and whose ability to cause damage is restricted to his own patient list, a biased algorithm can inflict widespread and systematic harm on thousands of people all at once (Obermeyer et al., 2019).

The clinical damage due to heightened diagnostic bias is evident in various realms of therapy. First, the algorithm's classification of the linguistic and cultural manifestation of distress, especially among minorities of different races, can result in dire consequences in terms of pharmacological treatment and management. As seen from the findings, if an NLP tool inaccurately classifies the language or emotional tone of a Black person as disorganized and aggressive, it might set off a course of treatment and therapy leading to a diagnosis of schizophrenia rather than that of a severe stressor-induced or affective disorder (Vyas et al., 2020). Clinically, this means that the patient will be exposed to the adverse metabolic and neurological consequences of taking anti-psychotic drugs, whereas the actual pathogenesis remains uncured. In addition, a flagged case increases the risk of involuntary commitment to a psychiatric facility, which is traumatizing and affects minorities especially hard.

Of equal concern is the issue of algorithmic exclusion. If the AI algorithms used for diagnosis do not adequately capture the baseline behaviors of people of lower SES status or females, in the case where diagnostic AI models rely upon digital phenotyping or continuous monitoring via phone telemetry, then such diagnostic tools serve as the gatekeepers for treatment. A diagnostic tool that sees the absence of any digital activity, which may be due to an inactive data plan or a communal phone, as indicating clinical stability, effectively excludes the most vulnerable patients from being monitored through registry (Martinez-Martin et al., 2020). Such "digital invisibility" results in the postponement of necessary intervention until after catastrophic events, such as suicidal attempts or severe depression breakdowns (Fiske et al., 2019).

4.3 Addressing the Governance and Regulatory Vacuum

Remedying the clinical risks outlined above calls for an urgent and sweeping overhaul of the present governance system within the healthcare industry. The existing void of regulation based on the focus on the efficacy of algorithms on an aggregate basis as well as the maintenance of the proprietary nature of algorithmic “black boxes” is completely inadequate to deal with the intricacies of modern psychiatric AI technology. For starters, regulatory agencies like the FDA and the EMA need to move on from relying on aggregate statistics to mandating that any diagnostic algorithm seeking clinical approval must conduct thorough and stratified subgroup analysis (Gerke et al., 2020). Approval must be conditional on proving statistical parity in terms of how race, gender, and socio-economic factors affect each other at the intersection. This requires standardization of Algorithmic Impact Assessment (AIAs) before deployment, in order to determine the potential for discrimination within the model. This requires a rigorous approach to addressing the emerging challenges in AI and establishing robust ethical regulation (Borenstein & Howard, 2021).

In addition to this, the shield of intellectual property must be lifted in relation to clinical medicine. The regulator must demand transparency of algorithms, such that an independent clinical auditor is able to scrutinize the composition of training data and feature weightings in order to detect any inherent biases (Price & Cohen, 2019). Post-market regulation needs to become radical as well. The constant learning characteristic of these psychiatric AI algorithms means constant regulation is needed as well. Hospitals using them need to put into place protocols for what one can call “algorithmic pharmacovigilance” so that emerging concept drift and demographic bias can be detected on the spot (Wiens et al., 2019). It necessitates forming hospital-level committees with clinicians, data scientists, and patients' advocates that have the power to take down or reconfigure algorithms where their equity metrics are degrading. In the end, a liability framework will need to be put in place to make sure that software developers and hospitals are responsible for algorithmic harm.

4.4 Ethical Imperatives for Algorithmic Fairness in Mental Health

In addition to the regulation issue, the use of artificial intelligence in psychiatric evaluation is inherently an issue of medical ethics, where the core considerations include non-maleficence (doing no harm) and distributive justice. The trend of incorporating algorithms into psychiatry currently poses a risk of breach of both principles as it is bound to subject the disadvantaged group to the risks of misdiagnosis while keeping the benefits of prediction for the rich and dominant group (Char et al., 2018). It is crucial to have algorithms that align digital psychiatry with medical ethics.

An important ethical mandate is the rigid adherence to the use of "Human-in-the-Loop" (HITL) care pathways. AI should be intentionally restricted to playing only an assistive function by only complementing but never substituting human clinical judgments (Kelly et al., 2019). Yet, being aware of the danger of automation bias in AI ethics, "friction" should be built into the design of ethical AI as a means of making the clinician explain and document the justification of accepting the risk score of an algorithm to change the triage status of the patient.

Additionally, for algorithmic fairness to be achieved, it is crucial to embrace participatory design. The development of diagnostic AI models has traditionally been done in a purely technological vacuum, with patient populations that have traditionally been marginalized acting only as passive contributors of data. This calls for an ethical paradigm change wherein the voices of patient advocates, linguistic minorities, and community mental health professionals are actively involved from the very beginning of model design and validation processes (Thieme et al., 2020). The inclusion of perspectives of the very population affected by the potential harm posed by the algorithms will allow the developers to detect proxy biases and cultural intricacies not captured by standard computational metrics. Algorithmic fairness cannot be achieved mathematically; it is rather a continuous ethical process of questioning and deconstructing systemic injustices inherent in digital healthcare systems.

4.5 Limitations of This Review

Despite being an exhaustive review of the literature on socio-technical and regulatory aspects of bias in algorithms for psychiatry, there are various limitations to this review. The most crucial limitation is the significant Western-centric bias within the literature under review. Most of the studies included in this review assess algorithms implemented in the USA, UK, and Western Europe (Gianfrancesco et al., 2018). Thus, these results and recommendations may not be applicable in the case of LMICs' health care systems, where the manifestation of mental distress differs from that in Western societies, and access to the Internet is limited (Fletcher et al., 2021).

Secondly, the rapid and accelerating pace of AI development introduces a temporal limitation. Machine learning architectures, particularly Large Language Models (LLMs) and advanced NLP tools, evolve at a rate that significantly outpaces the academic publishing cycle. Some of the specific algorithmic vulnerabilities identified in studies from the early phase of the publication window may have been partially mitigated in subsequent, proprietary iterations of commercial software, though the fundamental socio-technical risks remain. Also, dependence on English-language studies that are indexed in major medical databases could potentially lead to the omission of other non-English studies and computer science conferences proceedings. Lastly, this study is mostly based on secondary observational data and theoretical criticisms since the literature lacks substantial RCTs that would compare the long-term equality effects of AI-supported mental health care and regular human-based mental health care (Rhee et al., 2021). Further research should focus on conducting diverse randomized controlled trials in order to quantitatively measure the consequences of algorithmic bias.

5. Conclusions

Through this systematic review, there is evidence that has been collected about socio-technical aspects of algorithmic bias in the case of AI-enabled psychiatric diagnoses, which indicate that at the moment, the current systems are unable to achieve the basic requirements of fair clinical practice. According to the results obtained, it can be seen that AI-based diagnostic systems tend to have substantial differences in terms of accuracy in different demographic groups because of representation, label, and proxy biases. Such technical issues are not only computational failures but rather tools of harm in clinical practice towards certain demographic groups.

The present-day regulatory system, which is based on aggregate performance measures and absence of transparency about proprietary algorithms, is a case of a governance failure (Babic et al., 2021). In order to address the risks, the policy implications should go beyond voluntary ethical codes of conduct. The regulatory agencies should require stratified analyses of subgroups in terms of demographics as a requirement for obtaining a clinical approval and conduct algorithmic fairness audits using established standards (Gerke et al., 2020). Moreover, the shift towards the "Human-in-the-Loop" (HITL) approach is necessary, and, therefore, clinical settings need to avoid automation bias and incorporate friction into triage workflows (Kelly et al., 2019).

Moving forward, future research should concentrate on the following three crucial areas of study to solve the problem of current equity gap. Firstly, it is crucial to create more diverse training datasets to guarantee the inclusion of different parts of the globe in this process and shift the attention from the current bias towards Western approaches. Secondly, researchers should conduct more longitudinal randomized controlled trials to measure the equity outcomes of the use of AI-assisted treatment compared to the conventional one (Rhee et al., 2021). And finally, the development of explainable AI (XAI) models becomes crucial for clinicians as it will make it possible to interrogate the decisions made by algorithms and turn them from the mysterious "black box" to the tool for clinical practice (Ghassemi et al., 2021). Failure to address these challenges is not only a technical issue, but it is a matter of moral responsibility as well.

Author Contributions:

Conceptualization, Investigation, Visualization: Alicja Kwiatkowska

Conceptualization, Resources, Visualization: Jakub Juskowiak

Methodology, Supervision: Mateusz Lis

Methodology, Supervision, Data Curation: Zuzanna Julia Komin

Software, Data Curation, Project Administration: Maryan Liasota

Validation, Writing: Eliza Zofia Kabat

Validation, Writing: Michalina Justyna Kozłowska

Formal Analysis, Writing: Krzysztof Konopka

Formal Analysis, Writing: Maciej Kamil Majchrzak

Investigation, Writing: Faustyna Magdalena Kołodziej

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

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

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

REFERENCES

1. Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Direct-to-consumer medical machine learning and artificial intelligence applications. *Nature Machine Intelligence*, 3(4), 283–287. <https://doi.org/10.1038/s42256-021-00331-0>
2. Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for ethical regulation. *AI and Ethics*, 1(1), 61–65. <https://doi.org/10.1007/s43681-020-00002-7>
3. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981–983. <https://doi.org/10.1056/NEJMp1714229>
4. Chen, I. Y., Joshi, S., & Ghassemi, M. (2019). Treating health disparities with artificial intelligence. *Physics Today*, 72(11), 38–44. <https://doi.org/10.1063/PT.3.4340>
5. Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., ... & Sanz, F. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, 3(1), 1–11. <https://doi.org/10.1038/s41746-020-0288-5>
6. Coppersmith, G., Leary, R., Peach, P., Alajajian, R., & Dredze, M. (2018). Predicting suicide risk from public social media data. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology*, 59–69. <https://doi.org/10.18653/v1/W18-0606>
7. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(102), 1–61. <https://jmlr.org/papers/v23/21-0463.html>
8. Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research*, 21(5), e13216. <https://doi.org/10.2196/13216>
9. Fletcher, R. R., Nakeshimana, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Frontiers in Artificial Intelligence*, 3, Article 561802. <https://doi.org/10.3389/frai.2020.561802>
10. Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial intelligence in healthcare* (pp. 295–316). Academic Press. <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>
11. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable AI in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
12. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544–1547. <https://doi.org/10.1001/jamainternmed.2018.3763>
13. Grote, T., & Keeling, P. (2020). On the normative foundations of algorithmic fairness in healthcare. *Medicine, Health Care and Philosophy*, 23(4), 637–647. <https://doi.org/10.1007/s11019-020-09958-2>
14. Jacobson, N. C., Bentley, K. H., Schubert, A., & Banerjee, S. (2020). Individualized risk assessment of suicide using portable technologies and machine learning. *Journal of Consulting and Clinical Psychology*, 88(4), 377–389. <https://doi.org/10.1037/ccp0000481>
15. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), Article 195. <https://doi.org/10.1186/s12916-019-1426-2>
16. Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Senner, F., Feis, R. A., Asparuhov, O., ... & Kambeitz, J. (2021). Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry*, 78(2), 195–209. <https://doi.org/10.1001/jamapsychiatry.2020.3604>
17. Landi, I., Glicksberg, B. S., & Dudley, J. T. (2020). Deep learning for electronic health records in psychiatry. *Current Behavioral Neuroscience Reports*, 7(4), 226–235. <https://doi.org/10.1007/s40473-020-00221-x>
18. Martinez-Martin, N., Inoue, Y. A., & Greely, H. T. (2020). Ethical issues in using digital phenotyping for psychiatry. *Harvard Review of Psychiatry*, 28(5), 321–329. <https://doi.org/10.1097/HRP.0000000000000269>
19. McCradden, M. D., Spurles, S., & Stephenson, J. A. (2020). Beyond bias: A feminist sociology of machine learning in healthcare. *Journal of Medical Ethics*, 46(11), 770–777. <https://doi.org/10.1136/medethics-2020-106160>

20. Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine learning and algorithmic fairness in public and international health. *Annual Review of Public Health*, 42(1), 1–22. <https://doi.org/10.1146/annurev-publhealth-090419-102240>
21. Miner, A. S., Laranjo, L., & Kocaballi, A. B. (2020). Chatbots in the clinical domain: Are they ready? *JAMA*, 323(5), 413–414. <https://doi.org/10.1001/jama.2019.21323>
22. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
23. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
24. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37–43. <https://doi.org/10.1038/s41591-018-0272-7>
25. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 1–10. <https://doi.org/10.1038/s41746-018-0029-1>
26. Rhee, T. G., Wilkinson, S. T., & Rosenheck, R. A. (2021). Clinical trials of machine learning algorithms in psychiatry: A systematic review. *The Journal of Clinical Psychiatry*, 82(6), Article 21r14065. <https://doi.org/10.4088/JCP.21r14065>
27. Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A systematic review of the active literature. *Frontiers in Psychiatry*, 10, 93. <https://doi.org/10.3389/fpsy.2019.00093>
28. Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of its practical use and ethical challenges. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5), 1–53. <https://doi.org/10.1145/3398069>
29. Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Riper, H., ... & Wykes, T. (2021). The growing field of digital psychiatry: Current evidence and future directions. *World Psychiatry*, 20(3), 318–335. <https://doi.org/10.1002/wps.20883>
30. Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5(1), 48. <https://doi.org/10.1038/s41746-022-00592-y>
31. Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9), 874–882. <https://doi.org/10.1056/NEJMms2004740>
32. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340. <https://doi.org/10.1038/s41591-019-0548-6>