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	AI IN PSYCHIATRIC DIAGNOSTICS – A REVIEW
DOI	https://doi.org/10.31435/ijitss.3(51).2026.6494
RECEIVED	22 May 2026
ACCEPTED	11 August 2026
PUBLISHED	20 August 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

AI IN PSYCHIATRIC DIAGNOSTICS – A REVIEW

Wiktor Rybicki *(Corresponding Author, Email: wiktorandrzejrybicki@gmail.com)*
Nowodworskie Centrum Medyczne, 2 Miodowa Street, 05-100 Nowy Dwór Mazowiecki, Poland

Radosław Dutczak
Medical University of Silesia in Katowice, ul. Księcia Józefa Poniatowskiego 15, 40-055 Katowice, Poland

Aleksandra Sobieska
Medical University of Warsaw, ul. Żwirki i Wigury 61, 02-091 Warszawa, Poland
ORCID ID: 0009-0009-1538-6237

Patrycja Mularczyk
Nowodworskie Centrum Medyczne, 2 Miodowa Street, 05-100 Nowy Dwór Mazowiecki, Poland
ORCID ID: 0009-0000-1733-7185

Julia Jaguszewska
Independent Researcher, Poland
ORCID ID: 0009-0002-5500-7894

Weronika Martyna Górska
Independent Researcher, Poland
ORCID ID: 0009-0005-3155-9573

Mateusz Okoń
Uniwersytecki Szpital Kliniczny Nr 1 w Lublinie, ul. Stanisława Staszica 16, 20-081 Lublin, Poland
ORCID ID: 0009-0005-1439-9643

Alicja Gawlik
Zespół Opieki Zdrowotnej w Dębicy, Krakowska 91, 39-200 Dębica, Poland
ORCID ID: 0009-0005-1342-435X

Alicja Korpacka
Our Lady of Perpetual Help Hospital, Gdyńska 1/3, 05-200 Wołomin, Poland
ORCID ID: 0009-0002-3638-5312

Iga Maria Owczarek-Danowska
Warszawski Uniwersytet Medyczny, Warsaw, Mazovia, Poland
ORCID ID: 0009-0002-9037-7807

Aleksandra Wojtera
Raszeja City Hospital in Poznań, ul. Mickiewicza 2, 60-834 Poznań, Poland
ORCID ID: 0009-0007-7288-676X

ABSTRACT

Introduction: Artificial intelligence (AI) is increasingly being used in psychiatry, with side effects on solutions stemming from the subjectivity of diagnosis, limited care, and biological complexity, which is subject to threats. Mental disorders affect 293 million people worldwide and pose a burden on human health [9].
Aim: The aim of this review is to summarize the current state of knowledge on AI applications in psychiatric diagnostics, with specific focus on: (1) analysis of communication traffic of AI algorithms, (2) analysis of the results of AI-based primary control, (3) extension of methodological and ethical implications, and (4) extension of research.
Methods: A review of the research literature was conducted in the field of Basic Language Processing (NLP) in digital phenotyping, AI-assisted neuroimaging, and the ethical and legal implications of implementing these technologies. Meta-analyses, specific reviews, and original empirical studies completed between 2015 and 2026 were analyzed.
Results: A meta-analysis reported a cumulative AI diagnostic accuracy of 85% and a therapeutic efficacy of 84% in specific applications [8]. NLP enabled independent assessment, achieving an 86% (AUC 0.93) in studies on psychosis risk states [18]. Chatbots (Woebot, Wysa, Youper) demonstrate the consequences of problem occurrence and anxiety [9]. A review of 555 neuroimaging models revealed that 83.1% of the symptoms appear as a consequence rather than being triggered by a utility [32]. The most important ethical concerns were identified, including algorithm opacity ("black box"), liability, and data privacy [54, 56, 61].
Conclusions: AI in psychiatric diagnostics has demonstrated transformative potential, particularly in the areas of NLP and digital phenotyping, but current neuroimaging models require methodological improvements. The development of comprehensive ethical frameworks and extensions, simple algorithms, and model validation in large, population-based cohorts are essential. The ultimate success of AI in psychiatry will depend on striking a balance between technological innovation and respect for fundamental ethical values, while maintaining a paramount clinical role in diagnostic and therapeutic procedures.

KEYWORDS

Artificial Intelligence, Psychiatry, Diagnostics, Material Language Transmission, Neuroimaging, Ethics

CITATION

Wiktor Rybicki, Radosław Dutczak, Aleksandra Sobieska, Patrycja Mularczyk, Julia Jaguszewska, Weronika Martyna Górska, Mateusz Okoń, Alicja Gawlik, Alicja Korpacka, Iga Maria Owczarek-Danowska, Aleksandra Wojtera. (2026) AI in Psychiatric Diagnostics – a Review. *International Journal of Innovative Technologies in Social Science*. 3(51). doi: 10.31435/ijitss.3(51).2026.6494

Introduction

The development of artificial intelligence (AI) in recent years has opened up new possibilities in the diagnosis and treatment of mental disorders, representing one of the most promising, yet controversial, developments in contemporary psychiatry. Mental disorders are among the most serious public health challenges worldwide – according to the Global Burden of Disease Study, approximately 293 million people aged 5–24 (one in ten) struggle with mental health problems, with the most common being anxiety disorders, depression, post-traumatic stress disorder (PTSD), and obsessive-compulsive disorder [9]. Globally, mental disorders account for over 418 million disability-adjusted life years (DALYs) [9]. This situation is further complicated by persistent barriers to accessing mental health care, such as stigma, high treatment costs, limited number of qualified therapists, and geographic barriers [9]. In response to these challenges, AI-based tools are seen as potentially capable of increasing the accessibility, objectivity, and scalability of psychiatric interventions [2, 5]. Traditional diagnostic methods in psychiatry, although fundamental to clinical practice, face significant limitations. They rely heavily on subjective clinical assessments, interviews, and patient self-assessments, which can lead to diagnostic inconsistencies and the elusiveness of the complex picture of mental disorders [8]. These challenges are particularly evident in disorders with overlapping symptoms, such as schizophrenia, depression, bipolar disorder, and anxiety disorders, where distinguishing between nosological entities using conventional methods can be difficult [8]. Furthermore, psychiatric diagnosis relies heavily on the analysis of spoken and written language – from verbal communication between the patient and clinician, through symptom documentation, to the assessment of treatment outcomes [6]. Language, as a carrier of clinical information, is an area where generative AI models – such as ChatGPT – can find particularly wide application [6]. In the field of research into the biological basis of mental disorders, technological advances enable increasingly advanced analysis of genetic and neurobiological factors. In the case of bipolar disorder, whose heritability is estimated at 60–80%, research to date has identified a relatively small portion of the genetic variants responsible for its etiology [1]. A landmark study by O'Connell and colleagues (2025), using data from participants of diverse ethnicities (European, East Asian, African American, and Hispanic) across 158,036 bipolar disorder cases and 2.8 million controls, identified 298 genome-wide significant loci—a four-fold increase over previous findings [1]. This study also demonstrated differences in the genetic architecture of bipolar disorder depending on patient recruitment source and disorder subtype (type I or II), and identified specific cell types—including GABAergic interneurons and medium spiny neurons—as important in the pathophysiology of this disorder [1]. These findings underscore the biological complexity of mental disorders and justify the need for advanced data analysis methods offered by AI.

In parallel with genetic studies, neuroimaging provides important information on the structural and functional correlates of mental disorders. Studies on posttraumatic stress disorder (PTSD) indicate significant volumetric changes in the cerebellum, a structure traditionally associated with motor functions and now considered crucial for higher cognitive and emotional functions [3]. A mega-analysis conducted by Huggins et al. (2024) on a sample of 4,215 adults (1,642 with PTSD and 2,573 controls) from 40 centers showed that PTSD is associated with significant reductions in the total volume of the cerebellum and selected subregions, primarily in the posterior lobe (lobule VIIB, crus II), vermis (VI, VIII), and medullary nucleus [3]. Importantly, these effects were more pronounced when analyzing symptom severity than diagnosis itself, suggesting a continuous nature of neurobiological changes in PTSD [3]. These findings, along with the observation of reduced cingulate gyrus and prefrontal cortex volume in this population, demonstrate that AI-assisted neuroimaging data analysis may contribute to a better understanding of the pathophysiology of mental disorders and the identification of potential biomarkers.

Systematic literature reviews of AI in psychiatry indicate its promising effectiveness in both diagnosis and treatment. A meta-analysis by Rony et al. (2025), encompassing 14 studies, found a cumulative diagnostic accuracy of 85% (95% CI: 80–87%) for AI and a therapeutic efficacy of 84% (95% CI: 82–86%) for AI [8]. Machine learning (ML) models achieved the highest diagnostic accuracy (85%), followed by hybrid models (84%) and deep learning (DL) models (82%) [8]. In therapy, ML models excelled in personalizing treatment plans and monitoring symptoms, achieving an 85% success rate [8]. A systematic review by Dehbozorgi et al. (2025), covering 15 studies, confirmed that AI tools such as Wysa chatbots significantly improve self-reported mental health symptoms, while highlighting ethical challenges related to data privacy and algorithm transparency [7]. These studies also indicate high user acceptability of AI interventions, which may be crucial for their implementation in clinical practice [7]. AI applications in cognitive behavioral therapy (CBT) via chatbots seem particularly promising. A systematic review by Farzan et al. (2025) found that the three most widely studied chatbots—Woebot, Wysa, and Youper—provide significant reductions in symptoms of depression and anxiety [9]. Woebot demonstrated particularly high effectiveness in reducing symptoms of depression and anxiety with high levels of user engagement, Wysa demonstrated similar benefits in populations with chronic pain and perinatal mental health issues, and Youper led to a 48% reduction in depression symptoms and a 43% reduction in anxiety symptoms [9]. High levels of user satisfaction and the formation of a therapeutic alliance with chatbots comparable to traditional therapy suggest that these tools may be a valuable complement to standard interventions [9]. Similar conclusions are drawn from a systematic review by Sun et al. (2025), which found that AI algorithms achieve high diagnostic accuracy in detecting autism spectrum disorder based on facial and gaze patterns in social interaction assessments (AUC = 0.862), with better results obtained in unstructured play environments compared to structured diagnostic assessments [2]. This study also highlighted the dominant role of facial features as a source of information in ASD diagnosis [2].

In the context of predicting treatment response, AI has particularly significant potential in psychiatry, where response to therapy is highly variable. Saboori Amleshi et al.'s (2025) meta-analysis of schizophrenia showed that AI models achieved a cumulative sensitivity of 70% and specificity of 76% in predicting treatment response [4]. EEG-based models performed the best (sensitivity 89%, specificity 94%), followed by imaging-based models (76% and 80%), suggesting that physiological data may be a particularly valuable source of information for predictive algorithms [4]. This study also revealed significant methodological heterogeneity in defining treatment response and across study populations, indicating the need for standardization in future studies [4]. It is worth emphasizing that the development of AI in psychiatry is not limited to purely technological tools – it also encompasses ethical and social aspects. Generative AI, encompassing models capable of generating text, images, or video, creates new opportunities but also challenges related to security, privacy, and potential misuse [6]. A systematic review by Kolding et al. (2025), covering 40 studies, found that most studies concerned ChatGPT applications in psychiatry, with prompt-based experiments dominating the field, indicating the early stage of development [6]. The authors emphasized the need for increased transparency of methods, the use of experimental research designs, ensuring clinical relevance, and involving users/patients in the design phase of AI tools [6]. Similar postulates regarding the ethical dimension of AI applications in psychiatry appear in the review by Pham et al. (2022), which discusses AI-based interventions in the form of therapeutic chatbots, avatar-based therapies, and intelligent animal-like robots [5]. The authors emphasize the potential of these tools to increase access to psychiatric care, particularly in the context of staff shortages, while also emphasizing the need to develop ethical and regulatory frameworks [5].

Despite promising results, the literature on AI in psychiatry is burdened with significant limitations. Many studies are characterized by small sample sizes, lack of control groups, methodological heterogeneity, and limited transparency in reporting AI models and data processing techniques [7, 8]. Furthermore, most studies focus on Western populations, raising questions about the generalizability of results to other ethnic and cultural groups [2, 8]. Integrating AI tools into existing healthcare systems and ensuring their compliance with legal and ethical requirements also remains a challenge [6, 7]. Given the above considerations, this review aims to synthesize the current state of knowledge on AI applications in psychiatric diagnostics. Specifically, the work focuses on: (1) assessing the diagnostic accuracy of AI algorithms in various mental disorders, (2) analyzing the therapeutic effectiveness of AI-based interventions, (3) identifying key methodological and ethical challenges related to implementing AI in psychiatry, and (4) defining directions for future research in this rapidly evolving field. Through a critical synthesis of available evidence, this work aims to contribute to a better understanding of the potential and limitations of AI in psychiatry, supporting both research development and clinical practice.

Methods

This work is a narrative review, the aim of which was to synthesize and critically analyze the available literature on the application of artificial intelligence (AI) in psychiatric diagnostics. Unlike a systematic review, this work does not follow a pre-registered protocol (e.g., PRISMA-compliant) nor does it aim to exhaustively include all available publications on a given topic. Rather, its goal is to provide a representative and up-to-date overview of the state of knowledge in three key areas: (1) natural language processing (NLP) and digital phenotyping, (2) AI-assisted neuroimaging, and (3) the ethical and legal aspects of implementing these technologies in clinical practice.

Search strategy. A literature search was conducted in the PubMed/MEDLINE, Scopus, and Google Scholar databases, including papers published between January 2015 and July 2026. Keyword combinations in English were used, including: "artificial intelligence AND psychiatry," "machine learning AND psychiatric diagnosis," "natural language processing AND mental health," "deep learning AND neuroimaging AND psychiatric disorders," "digital phenotyping," "large language models AND mental health," as well as terms related to ethical aspects: "AI ethics AND psychiatry," "black box AND clinical decision-making," "algorithmic bias AND healthcare," "informed consent AND artificial intelligence." Additionally, the bibliographic lists of identified reviews and meta-analyses were searched to identify additional relevant sources (snowballing).

Publication selection. Criteria include priority meta-analyses, systematic literature reviews, and original studies with large samples and high methodological quality, published in peer-reviewed journals. Priority was given to recent papers (2023–2026), particularly in rapidly developing areas such as the application of large language models (LLM). Selected classic or groundbreaking papers in a given field (e.g., the 2015 study by Bedi et al.) were also included if they served as a significant reference for subsequent research. Conference proceedings without full text, papers unavailable in English or Polish, and publications not subject to peer review (preprints) were excluded, unless a peer-reviewed version existed and the content was deemed relevant to the topic under discussion.

Data Synthesis. Due to the narrative nature of the review, no formal assessment of the methodological quality of individual source studies was performed using standardized tools (except for reference to risk of bias assessment results reported directly in cited meta-analyses, e.g., using the PROBAST tool [32]), nor was our own quantitative meta-analysis performed. Instead, a qualitative approach was used, consisting of thematic grouping and critical comparison of results from different sources, considering concordances and discrepancies between studies and identifying recurring methodological limitations.

1. NLP and Digital Phenotyping

1.1. Results

This chapter presents the results of a review of the application of natural language processing (NLP) in digital phenotyping. The research synthesis demonstrates the significant potential of NLP techniques in automating, objectifying, and enriching psychiatric assessments, but also reveals significant methodological, technical, and ethical challenges that determine their implementation in clinical practice. The results are organized according to four main analytical axes: (i) methodology for data acquisition and linguistic features; (ii) the effectiveness of NLP models in identifying clinical conditions; (iii) challenges related to implementation in real-world settings; and (iv) directions for further development in the context of personalized medicine.

The first group of results concerns the methodological diversity of data acquisition. The studies utilize data from electronic health records (EHRs), social media, and interactive patient diaries. An analysis of 58 studies in the review by Le Glaz et al. [11] found that the primary data sources for NLP in psychiatry are medical records (EHRs, clinical notes) and social media posts. The latter allows for language analysis in the patient's natural environment but raises questions about sample representativeness and privacy. In schizophrenia research, as indicated by the review by Deneault et al. [19], analyses based on clinical interview transcripts and, increasingly, content from online forums and social media sites also dominate.

A key finding is the demonstration that linguistic features extracted using NLP, such as semantic coherence, part-of-speech usage, syntactic complexity, and speech rate, are significant predictors of clinical outcomes. A study by Argolo et al. [18] on individuals at risk for psychosis (ARMS) showed that advanced measures such as semantic laminarity (measuring the degree of recurrence of topics), semantic recurrence time, and centrality in verbal repetition graphs were most informative about symptoms and allowed for classification of ARMS with a balanced accuracy of 86% (AUC 0.93). In this study, new measures of recurrence quantitative analysis (RQA) proved particularly accurate in reflecting clinical phenomena such as perseveration and thought deviance, and were correlated with negative, positive, and disorganized symptoms.

Results from analyses for depression, including a review by Guo et al. [13] covering 40 studies, show that large-scale language models (LLMs) such as BERT and GPT achieve high performance in detecting suicidal ideation and depressive episodes from text. However, it has been noted that many models focus on binary classification (e.g., depression vs. no depression), which is a limitation given the phenotypic complexity of mental disorders. Importantly, varying text length and using larger semantic windows have also been shown to reduce verbosity artifacts.

An important strand of research is the evaluation of practical NLP implementation in healthcare systems. Based on the experience in the British National Health Service (NHS) described by Au Yeung et al. [10], the implementation of an NLP service based on MedCAT enabled the collection of over 26,086 annotations covering 556 SNOMED CT concepts. This work emphasizes that the key success factor is the involvement of clinical experts in the annotation process, which allows for the incorporation of local context and specialized knowledge unavailable in external crowdsourcing services. An iterative learning and correction process was also identified, which significantly improves the accuracy of Named Entity Recognition (NER) systems. A similar conclusion is drawn from research on the analysis of confidential content in adolescent clinical notes, where Rabbani et al. [15] obtained an AUC of 0.90 for a logistic model detecting confidential content, indicating the potential for supporting processes managing access to sensitive information.

Research on digital phenotyping in eating disorders (Funk et al. [16]) revealed that analyzing textual features from diaries and messages between users and coaches predicts binge eating behavior (AUC 0.72 for known users). Digital Health Interventions Text Analytics (DHITA) methodology has shown that aggregating features such as word usage, embeddings, part-of-speech tagging, and sentiment analysis provides useful indicators for tracking symptom changes in real time. These studies suggest the possibility of automating some aspects of therapeutic interventions, although the authors emphasize limitations related to sample size and the need for experimental validation. An important line of research is provided by analyses related to autism spectrum disorder (ASD). Trayvick et al. [17] concluded in their review that NLP and automated speech analysis can provide objective, quantitative measures of linguistic features that have previously been assessed subjectively. They demonstrated that open-source tools such as Praat and OpenSmile enable quantitative assessment of rate, intonation, and accent, as well as repetitiveness and pedantry, which can support differential diagnosis. These studies, although promising, are plagued by heterogeneous results and small sample sizes, indicating the need for standardization.

In terms of predicting the development of psychosis, the results of Bedi et al. (2015) [12] cited in Argolo et al. [18] and in the review by Deneault et al. [19] achieved 100% accuracy in predicting psychosis in high-risk young individuals, significantly exceeding clinical assessment. However, the authors emphasize that small sample sizes (34 individuals) and the risk of model overfitting warrant caution and the need for further validation in larger, culturally diverse populations.

The results of studies on large language models (LLMs) reveal ambivalence. On the one hand, in the review by Guo et al. [13], the GPT-4 achieved 85% correct answers in neurological exams, exceeding the human average (73.8%) and demonstrating a high level of empathy in its responses, making it a potentially useful tool in interventions and health education. In the evaluation of depression treatment, the GPT-4, Claude, and Bard models demonstrated high agreement with physicians regarding treatment recommendations. On the other hand, studies have demonstrated the risk of generating inconsistent, hallucinatory, or harmful responses, particularly when prompts are reworded (e.g., recommending risky behaviors in a depressed patient). Additionally, the problem of outdated training data and difficulties in interpreting decisions (black box) has been identified, which constitute a significant barrier in the clinical context.

The analysis of non-English languages, conducted by Névéol et al. [14], reveals a significant resource asymmetry, with English dominating both in terms of the number of models and annotated corpora. In a study of ARMS in Portuguese [18], the use of new techniques such as FastText and RQA yielded accurate results, but the paucity of data in non-English languages was highlighted as a significant challenge. Adapting systems from English, for example, using machine translation, was often ineffective and led to performance degradation.

Regarding cultural and linguistic diversity, research findings indicate that linguistic features can be universal biomarkers for psychosis, but their parameters are language-specific. Deneault et al. [19] emphasize that most studies on NLP in schizophrenia are conducted in English, which limits generalizability. Future work should therefore focus on creating multilingual corpora annotated by experts to increase the reach and comparability of results. Overall, the results indicate that NLP and digital phenotyping represent promising tools for supporting diagnostics and monitoring in psychiatry. NLP models achieve high accuracy in classifying and predicting clinical conditions, often exceeding subjective assessments, but they are still sensitive to data quality, linguistic specificity, and the risk of hallucinations and poor interpretability. Advanced methods such as recurrence analysis and repetition graphs have been observed to provide new value by capturing subtle shifts in thought processes. Finally, the need for integrating models in real-world clinical environments, where clinician involvement in the annotation and verification process is crucial, has been identified, as well as the development of ethical standards and a regulatory framework for the safe use of these technologies.

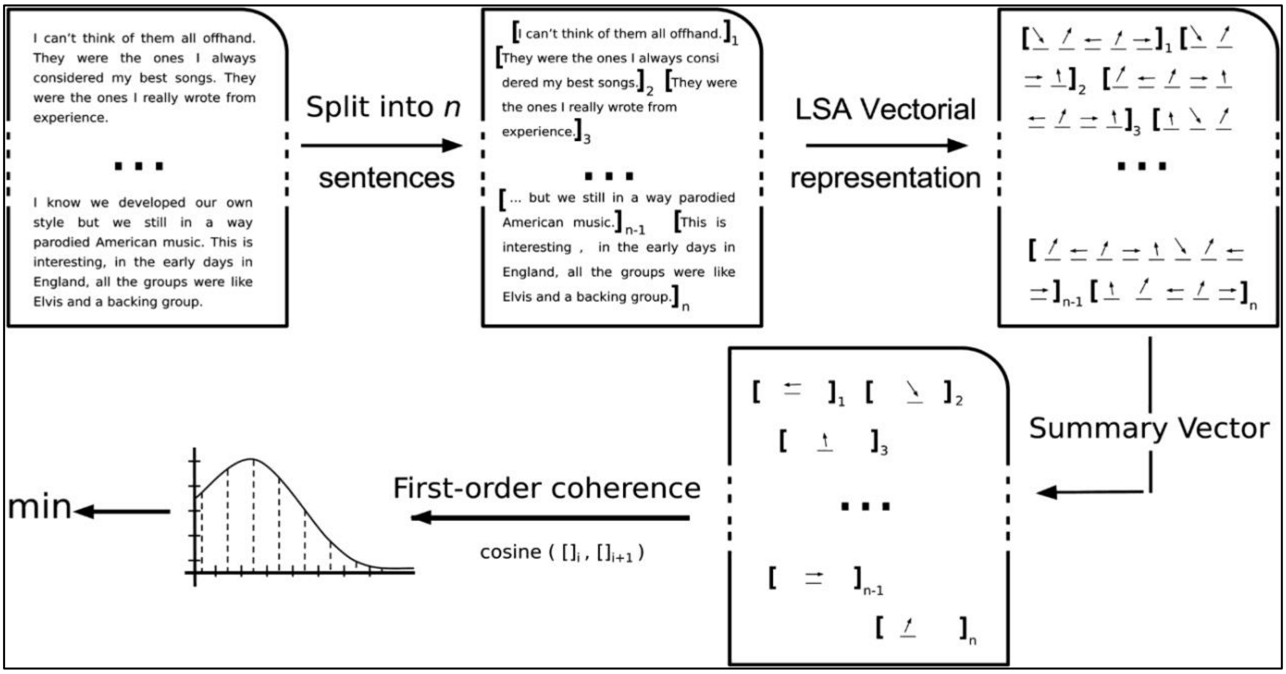

Fig. 1. Pipeline for automated extraction of the semantic coherence features.

Texts were initially split into sentences/phrases. Each word was represented as a vector in high-dimensional semantic space using Latent Semantic Analysis (LSA). Summary vectors were calculated as the mean of each vector in each phrase. Coherence was determined based on the semantic similarity between adjacent phrases, calculated as the cosine of their respective vectors. The semantic coherence feature that best discriminated those who transitioned to psychosis from those who did not was the minimum semantic coherence value (i.e., the coherence at the point of maximal discontinuity) within each transcribed text [12]

1.2. Discussion

Integrating natural language processing (NLP) with digital phenotyping is currently one of the most promising, albeit challenging, avenues for developing psychiatric diagnostics [20]. This discussion addresses key issues related to implementing these technologies in clinical practice, considering both their predictive potential and the limitations resulting from data heterogeneity, language barriers, and ethical issues.

From an implementation perspective, experience from healthcare systems such as the UK National Health Service (NHS) highlights the crucial role of integrating NLP with real-world clinical workflows [10]. As Au Yeung and colleagues note, embedding machine learning workflows in the hospital environment is essential to ensuring model consistency with real-world data and clinical processes [10]. These experiences demonstrate that effective implementation requires not only advanced tools but, above all, the development of a framework for collaboration between clinical teams and data engineers [10].

The availability of appropriate language resources, particularly in the context of languages other than English, remains a significant challenge. Névéol and colleagues point out that English is by far the best equipped language in terms of advanced tools dedicated to the biomedical domain, while data and resources for other languages are often scarce [14]. This asymmetry can lead to the creation of systems with limited utility in non-English-speaking populations, which constitutes a significant limitation for the global application of digital phenotyping tools [14].

In the context of digital phenotyping in psychiatry, research on psychosis spectrum disorders provides particularly interesting observations. Argolo and colleagues demonstrated that subtle speech disorders can be detected using sensitive NLP methods and used in screening for psychosis risk states [18]. Their research highlights the particular usefulness of new indicators, such as centrality in word repetition graphs or recurrence analysis, which reflect clinical phenomena such as perseveration and circularity [18]. Also in the context of schizophrenia, Deneault and colleagues indicate that NLP is primarily used in predictive and diagnostic modeling, as well as in the analysis of specific linguistic phenomena [19]. The most frequently identified application of NLP in this group of disorders is diagnostic and predictive modeling, which reflects the general trend towards using these technologies for early detection and prognosis of the disease [19]. At the same time, the authors emphasize that small sample sizes used to train models constitute a significant limitation that may hinder achieving satisfactory generalization [19, 30].

In autism research, Trayvick and colleagues highlight the potential of NLP in objective, quantitative assessment of speech patterns, which have previously been described primarily subjectively [17, 31]. The use of automated methods allows for the capture of both acoustic features of speech and linguistic patterns, such as echolalia and pragmatic difficulties, more reliably and reproducibly than traditional clinical assessments [17]. This is particularly important in the context of the heterogeneity of the autism phenotype.

From a methodological perspective, Funk and colleagues propose a structured framework for analyzing text data in digital health interventions, which can be adapted to the needs of digital phenotyping in psychiatry [16]. The authors emphasize that textual features can predict changes in symptoms over time, and an integrated approach to text analysis provides valuable information about the mechanisms underlying intervention effects [16]. At the same time, they point out the need for cautious interpretation of predictive models, which do not necessarily reflect causal relationships [16]. However, systematic reviews point to significant limitations in the current state of knowledge [21]. Le Glaz and colleagues note that NLP models in mental health research often confirm clinical hypotheses rather than provide entirely new information, and that one of the main population categories – social media users – constitutes an imprecise cohort [11]. Furthermore, the authors emphasize that language-specific characteristics may improve the effectiveness of NLP methods, which requires further research on extending these methods to other languages [11, 22]. Guo and colleagues, in their systematic analysis of large language models in mental health, indicate that although these models demonstrate good performance in detecting mental health problems, the risks associated with their use in clinical practice may currently outweigh the benefits [13]. The authors cite inconsistencies in generated text, hallucinations, and the lack of a comprehensive ethical framework as major obstacles to clinical implementation [13]. They also point to the "black box" problem, which hinders the interpretability of model decisions [13]. A particularly important area of NLP application in psychiatry is the analysis of confidential content, especially in vulnerable populations [23]. Rabbani and colleagues demonstrated that NLP algorithms can identify confidential content in adolescent clinical notes with high accuracy, which is crucial for maintaining privacy in the context of mandatory disclosure of health information to patients [15]. Their model achieved an area under the ROC curve of 90% in their test cohort, indicating the potential of NLP to support operational efforts to protect adolescent confidentiality [15].

1.3. Conclusions

A review of the literature on the use of natural language processing (NLP) in digital phenotyping in psychiatry leads to several key conclusions. First, NLP enables the objectification and quantification of previously subjective clinical features [24], such as thought disorders or semantic coherence of utterances, which is supported by research on the psychosis spectrum [18, 19]. These methods, going beyond traditional clinical assessments, allow for the detection of subtle but significant linguistic patterns that can serve as biomarkers of psychosis risk states [18]. Second, access to high-quality, specialized linguistic resources, including expert-annotated corpora, is crucial for the effectiveness of these methods, which is particularly challenging in languages other than English [14, 25]. Adapting tools developed for English requires careful consideration of linguistic and cultural specificities, and simple translation often proves insufficient [14, 26]. Third, the potential of NLP in everyday clinical practice is enormous – from supporting early detection and prognosis, through monitoring treatment response, to supporting therapeutic decisions [11, 13]. However, despite promising research results, widespread implementation of these technologies faces barriers such as a lack of standardization, concerns about data privacy, and the risk associated with the unpredictability of generated content, which is particularly evident in the case of large language models [13, 15]. Ultimately, integrating NLP with digital phenotyping in psychiatry appears to be a promising direction [29], which, however, requires further, carefully designed validation studies and the development of an ethical framework to ensure it can safely support, rather than replace, the diagnostic and therapeutic process [27, 28].

2. Neuroimaging and Biomarkers

2.1. Results

The analysis included 555 AI models from 517 studies published between 1990 and 2022 [32]. Additionally, to deepen the analysis, we considered the results of studies on replicability and generalizability in population-based psychiatric neuroimaging studies [33], the clinical utility of neuroimaging in psychiatric practice [34], advances in molecular and multimodal imaging in neurodegenerative diseases [35], and trends in AI research productivity in clinical neurology and neuroimaging [36].

2.1.1 Risk of Bias Assessment

The majority of the analyzed AI models (83.1%; 95% CI, 80.0%-86.2%) were assessed as having a high overall risk of bias (ROB) according to the PROBAST tool [32]. The highest risk of error was found in the analysis domain, which was primarily due to inadequate sample size (71.7% of models; 95% CI, 68.0%-75.6%), lack of model calibration assessment (100% of models), and failure to take into account data complexity (99.1 % of models; 95% CI, 98.3%-99.9%) [32]. These results are consistent with previous observations indicating that population-based studies in psychiatry are often based on small samples (from several dozen to several hundred participants), which significantly limits the replicability of the obtained results [33]. Marek and colleagues demonstrated that with small effect sizes typical of psychiatric neuroimaging studies ($r \leq 0.10$), samples of hundreds of participants are characterized by high variability of estimation, and only samples of thousands provide adequate precision and statistical power to detect reliable associations [33]. In the participant domain, only 0.4% of models (95% CI, 0.00%-0.01%) were assessed as having high ROB [32]. In the predictor domain, 33.7% of models (95% CI, 29.9%-37.6%) showed high ROB, with the main source being defining predictors with knowledge of the outcome [32]. In the outcome domain, 35.7% of models (95% CI, 31.8%-39.7%) had high ROB due to leakage of outcome knowledge from the testing sets to the predictors in the training sets [32]. This problem is particularly relevant in the context of neuroimaging models, where a large number of features (e.g., voxels, functional connectivity) require dimensionality reduction, which often leads to data leakage and overestimation of model performance [33, 38]. Tejavibulya and colleagues emphasize that bias in machine learning is defined as "outcomes that are systematically biased against an individual or group due to its innate and acquired characteristics" [38].

Risk of error analysis in the analysis domain revealed that 83.1% of the models were rated as high ROB. The main reasons were: inadequate sample size (71.7%), lack of calibration assessment (100%), and failure to consider data complexity (99.1%) [32]. Importantly, 52 models (9.4%) were rated as low ROB in the analysis domain, as calibration and data complexity issues were deemed to have limited relevance for these models [32]. In the context of schizophrenia research, Kotochinsky and colleagues, in a meta-analysis of 16 studies with 3601 participants, showed that deep learning models achieved higher specificity (0.83) and AUC (0.892) compared to support vector machines (SVMs) [39]. Similarly, models using whole-brain input outperformed models based solely on gray matter features (specificity 0.86 vs. 0.80; AUC 0.89 vs. 0.816) [39].

2.1.2. Clinical Utility

None of the analyzed AI models was considered clinically ready [32]. High concerns regarding clinical utility were found in the predictor domain for 35.9% of models (95% CI, 31.9%–39.9%) and in the outcome domain for 39.1% of models (95% CI, 35.0%–43.2%) [32]. These results are supported by a critical assessment of the utility of neuroimaging in routine psychiatric practice. As Vogel and Black point out, for a medical test to be clinically useful, it must pass a hierarchy of requirements – from demonstrating group differences, through reliably distinguishing between individuals with and without the condition, to improving treatment outcomes and patient quality of life at an acceptable cost-to-benefit ratio [34]. Neuroimaging in psychiatry currently typically fails to meet these criteria, primarily due to overlapping distributions of values between diagnostic groups, low reliability of individual measurements, and significant intra-individual variability [34, 49]. Kraus and colleagues argue that "biomarkers exist at the individual level," requiring increased focus on their intra-individual validation [49]. Population-based studies have shown that effect sizes for associations between brain connectivity and mental health symptoms are relatively small ($r \sim 0.10$), meaning that only samples of thousands of participants are adequately powered to provide accurate and precise estimates of these effects [33]. Machine learning models used to predict mental health symptoms from imaging have relatively low predictive accuracy and require sample sizes in the thousands for replicability [33, 38]. Furthermore, these studies are often limited to samples from Western, industrialized, high-income countries, further limiting the generalizability of the results [33, 38]. Grosenick and Liston indicate that multimodal and multivariate (multiview) approaches enable the discovery of complex, latent patterns that are missed in single-modality studies [43]. Zhou and colleagues, using a cross-validated random forest model, achieved an AUC of 0.802 in predicting anxious depression from multimodal neuroimaging data [40].

In the context of autism spectrum disorder research, a systematic review by Parellada and colleagues concluded that no single biomarker has sufficient evidence to inform clinical trials [48]. The most consistent findings were for glutathione, with reduced glutathione levels in autism (median Hedges' $g^* = -1.77$) [48]. Wen and colleagues, using a deep learning-based approach (Surreal-GAN) to identify dimensional neuroimaging endophenotypes, identified 31 genomic loci associated with nine endophenotypes for Alzheimer's disease, autism, late-life depression, and schizophrenia [44].

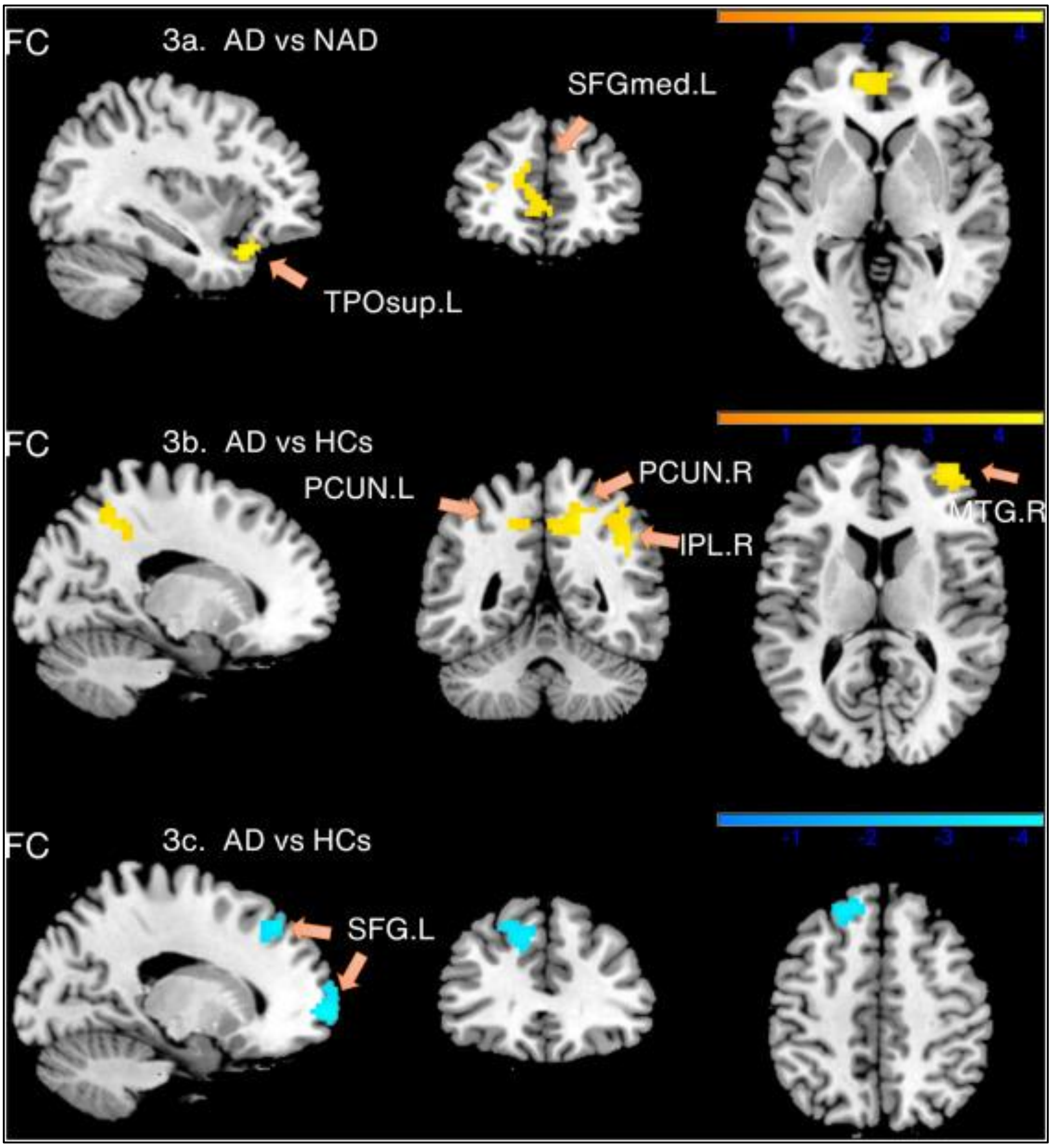

Fig. 2. *Comparison of FC*

Comparison of FC in seeds and other brain regions in anxious MDD patients and HCs and MDD patients with and without anxiety. 3a shows the difference in FC between AD and NAD. 3b and 3c show the positive and negative results of the comparison between AD and HCs, respectively [40].

2.1.3. Reporting Quality

The overall completeness of reporting for the analyzed AI models was 61.2% (95% CI, 60.6%–61.8%), with the lowest completeness found in the technical evaluation domain (39.9%; 95% CI, 38.8%–41.1%) [32]. In the data domain, 68.1% of models did not report quality controls (e.g., how to handle neuroimaging artifacts) [32]. In the technique domain, 46.5% of models had low reporting quality, primarily due to a lack of algorithmic details [32]. All models (100%) were rated as having low reporting quality in the technical evaluation domain, with 98.0% of models not reporting how to publicly evaluate their algorithm, and all models (100%) not including a bias assessment [32]. In the application domain, 27.0% of models (95% CI, 23.3%–30.7%) were characterized by low reporting quality due to a lack of discussion of clinical implications [32]. These reporting deficiencies have significant consequences for model replicability and translation. As Leenings and colleagues emphasize, the lack of standardization in neuroimaging data reporting, inter-scanner variability, and heterogeneity in acquisition protocols significantly hinder comparison of results between centers and study replication [41]. The introduction of standards such as BIDS (Brain Imaging Data Structure) can improve reporting quality and facilitate data exchange between research teams [41]. Baydili et al., in a review of AI applications in psychiatry, indicate that deep learning analyses classify treatment-resistant depression with an accuracy of >90%, and models based on speech analysis using large language models (LLM) achieve an F1-score of 92% in assessing thought disorders in schizophrenia [53].

Suppiah and colleagues, in a review of magnetic resonance spectroscopy (MRS) analysis and AI applications, found that deep learning models such as CNNs achieved 93.3–94% accuracy in classifying Alzheimer's disease from MRS data [37]. In eating disorder research, Monaco and colleagues, in a systematic review of five studies, used SVM to classify anorexia nervosa and bulimia nervosa with >80% accuracy from structural MRI [42].

2.1.4. Advances in Multimodal Imaging

Advanced neuroimaging techniques, including positron emission tomography (PET) with novel ligands and ultra-high-field imaging (7T MRI), now enable the quantitative assessment of molecular pathology, neuroinflammation, metabolism, and brain connectivity in vivo [35]. PET ligands for beta-amyloid (e.g., PiB, florbetapir) and tau (e.g., flortaucipir) allow for topographic quantification of pathology in Alzheimer's disease, while novel ligands for TSPO enable the assessment of neuroinflammation [35]. Multimodal studies combining PET with connectivity imaging (diffusion MRI, fMRI, MEG) provide evidence for the mechanisms of pathology spread along neuronal connections in neurodegenerative diseases [35]. This ability to image multiple aspects of pathophysiology at the molecular, cellular, and network levels may support the development of more sensitive and specific diagnostic biomarkers in the future [35, 43]. In their review of pharmacogenomic biomarkers in psychiatry, Kam and Jeong point out that pharmacogenomic tests for psychiatric medications such as amitriptyline, venlafaxine, and aripiprazole are based on CYP2D6 and CYP2C19 polymorphisms, and the FDA lists 35 psychiatric medications with pharmacogenomic information in their labeling [45]. Kraguljac and colleagues discuss mechanistically justified biomarker targets in schizophrenia, including dopaminergic hyperactivity, NMDA receptor hypofunction, hippocampal hyperactivity, immune dysregulation, and brain network dysconnectivity [46]. Golub and colleagues point to advanced markers of hepatotoxicity, such as osteopontin (OPN), HMGB1 protein, and glutamate dehydrogenase (GDH), which may be important in the context of taking new psychoactive substances [47]. Winchester and colleagues, in a review of AI applications in biomarker discovery in Alzheimer's disease, emphasize that machine learning methods, including random forests and gradient boosting, can effectively combine biomarkers from body fluids (e.g., NfL, Aβ42/40, p-tau181) with imaging and genetic data, achieving competitive predictive accuracy [52]. Lin and colleagues, in a review of precision psychiatry using AI, indicate that deep learning models, such as multilayer feedforward networks (MFNN), achieve an AUC of 0.82 in predicting response to antidepressants [51]. Morrin and Nour, in a commentary, emphasize the importance of interpretable speech biomarkers in psychosis, pointing out that speech features associated with negative symptoms are among the most replicable and clinically translatable [50].

2.1.5. Summary

The analyses conducted indicate that current neuroimaging-based AI models for psychiatric diagnosis are characterized by a high risk of bias, low clinical utility, and insufficient reporting quality. Major limitations relate to the analysis domain, including inadequate sample sizes, lack of calibration assessment, and failure to account for data complexity. These results are consistent with broader observations regarding the low replicability and generalizability of population-based studies in psychiatric neuroimaging, where effect sizes are small and sample sizes of thousands of participants are necessary to obtain reliable estimates. Despite promising advances in multimodal and molecular imaging, which may support the development of more sensitive biomarkers, current AI models are not mature enough for direct clinical application. Efforts are needed to reduce the risk of bias, improve reporting quality, and conduct validation in independent, large, and diverse population cohorts before implementing these models in clinical practice.

2.2. Discussion

Despite the enormous potential of neuroimaging-based artificial intelligence models, their translation into clinical practice faces significant barriers. A key problem is the high risk of bias. In a review of 555 models, as many as 83.1% were assessed as having a high risk of bias, particularly in the domain of data analysis, which undermines the reliability of the results [32]. None of the models were deemed directly applicable to practice, and completeness of reporting was insufficient (61.2%) [32]. This problem is compounded by the fact that correlations between brain connectivity and psychiatric symptoms are small ($r \approx 0.10$), requiring samples of thousands of individuals for reliable estimates [33]. Furthermore, even large samples do not guarantee generalization. Models can learn "shortcuts," associating data with unmeasured variables, such as demographic characteristics, rather than with the target disorder [33]. This translates into clinical reality – despite decades of research, there is still a lack of standardized imaging tests for diagnosing mental disorders [34]. Neuroimaging techniques rely primarily on group analyses, and the distribution of values in patients and healthy individuals often overlaps, preventing reliable individual diagnosis [34].

These challenges are compounded by the "dirty" nature of data – it is incomplete, inconsistent, and subject to measurement bias [38]. At the same time, the complexity of psychiatric data is an inherent feature that should be considered, not eliminated [38]. To address these limitations, a shift in focus to person-centered models is necessary. Biomarkers exist at the individual level, therefore extensive data collection within each individual is necessary, allowing for the capture of reliable individual differences in brain networks and psychopathology, which are imperceptible in group estimates [49].

Current AI models based on neuroimaging are not mature enough for clinical applications. Further progress requires building more diverse cohorts, standardizing protocols, rigorous external validation, and developing explainable models that gain clinicians' confidence.

2.3. Conclusions

This review reveals significant discrepancies between the promise of AI in psychiatry and its actual clinical utility. Most neuroimaging models are characterized by high risk of bias and low utility, resulting from inadequate sample sizes, lack of calibration, and failure to account for data complexity [32]. Similarly, models for schizophrenia demonstrate only preliminary performance, strongly dependent on the algorithm and input features [39]. Replicability remains a key challenge, as small sample sizes lead to overestimation of effects and lack of generalization [33]. Nevertheless, predictive models have potential, especially when applied transdiagnostically and taking into account the complexity of real-world data [38]. It is crucial to strive for model interpretability, which increases clinicians' confidence and enables understanding of the neurobiological basis of disorders [38]. Unified standards for evaluation and publication are necessary to enable comparability of studies [41]. Personalization of models, based on the characteristics of the brain and symptoms of a specific patient, appears to be a necessary direction of development [49]. Current knowledge indicates that AI models, although promising, require methodological refinement and validation before they can be recommended for routine use in psychiatry [32, 34].

3. Ethics, Law, and the "Black Box"

3.1. Results

Analysis of the collected research material reveals a multidimensional picture of the ethical and legal challenges associated with the implementation of artificial intelligence systems in psychiatric diagnostics. This chapter presents a synthesis of the results in key areas: the "black box" problem and algorithm opacity, legal liability issues, data privacy challenges, and aspects of patient consent and the therapeutic relationship.

3.1.1. The Problem of Algorithmic Opacity

A literature review indicates that a key ethical challenge in the use of AI in psychiatry is the lack of understanding of the internal decision-making mechanisms of deep learning systems [54]. The "black box" problem is particularly significant in the psychiatric context, where diagnostic decisions have far-reaching consequences for patients' lives. The opacity of algorithms undermines the possibility of obtaining informed consent, as the patient is unable to understand how the diagnosis was made [58]. This phenomenon is particularly problematic in cases where the algorithm recommends non-standard treatment paths that deviate from established clinical standards.

3.1.2. Legal Liability and Liability Distribution

A literature review reveals a significant lack of legal frameworks enabling precise definition of liability for diagnostic errors committed with the participation of AI systems [56]. Three main models of potential liability distribution have been identified: physician liability (when AI serves as a support), software developer liability (in the case of autonomous systems), and liability of the medical institution implementing the system [59]. Research indicates that a distributed liability model will dominate clinical practice, creating the risk of a "liability gap"—a situation in which no entity feels sufficiently obligated to bear the consequences of algorithmic errors [59]. This problem is particularly acute in psychiatry, where legal decisions are often based on subjective assessments of the patient's condition, which are difficult to algorithmize.

3.1.3. Data Privacy in the Psychiatric Context

The review findings indicate that data privacy and confidentiality are the most frequently discussed ethical issues in the context of conversational AI in psychiatric care [61]. Most of the analyzed articles raise concerns about data security, with the particularly sensitive nature of psychiatric data requiring a higher level of protection than in other areas of medicine [55]. Three categories of risks were identified: inadequate legal regulations for commercial psychiatric chatbots, excessive data collection (including full conversation transcripts), and potential violations leading to patient stigmatization [61]. Additionally, AI systems enable the creation of "social portraits" of patients based on their digital activity, raising questions about the extent to which such inferred data can be considered representative of the patient's actual personality and preferences [55].

3.1.4. Informed Consent and the Therapeutic Relationship

The analysis reveals that the issue of informed consent in the context of psychiatric AI is complex and multifaceted. Patients have the right to know the extent to which decisions regarding their treatment are made by algorithmic systems [58]. Research on family physicians' perspectives on AI indicates that a key challenge is maintaining a balance between increased diagnostic efficiency and maintaining trust in the doctor-patient relationship [57]. Physicians express concerns that excessive reliance on AI recommendations may lead to an erosion of their diagnostic competence and a weakening of communication skills [57]. The phenomenon of "automatic cognitive bias" has also been identified—the tendency to uncritically accept AI recommendations, especially in situations of diagnostic uncertainty [56].

3.1.5. Challenges Related to Algorithmic Bias

The review findings indicate a significant risk of AI systems perpetuating and exacerbating existing inequalities in access to mental health care [59]. Algorithmic bias may result from inadequate representation of minority groups in training data, leading to reduced diagnostic accuracy for these populations. Of particular concern are cases where algorithms recommend more intensive interventions for marginalized groups or, conversely, underestimate the severity of their symptoms [59]. This problem is further complicated by international differences in data protection standards, which hinder the creation of representative, culturally diverse training datasets [54].

3.1.6. Practical Implications

The identified challenges point to the need to develop a comprehensive regulatory framework that takes into account the specific nature of AI applications in psychiatry. Key proposals include: introducing mandatory human oversight of AI diagnostic decisions (human-in-the-loop), developing transparency standards for algorithms tailored to the needs of clinicians and patients, and creating audit mechanisms to monitor and correct algorithmic biases [60]. Developing guidelines regarding the scope of informing patients about AI's involvement in the diagnostic process while maintaining the practical utility of such systems is particularly important [58].

3.2. Discussion

The integration of artificial intelligence (AI) systems in psychiatric diagnostics, while promising, raises profound ethical and legal dilemmas. A central problem remains the "black box" issue, i.e., the opacity of algorithms, which undermines the possibility of obtaining informed consent and complicates attribution of responsibility for diagnostic errors. As Khalique et al. [54] rightly point out, the inability to explain AI's decision-making process contradicts fundamental principles of bioethics. In the context of psychiatric care, where the doctor-patient relationship is based on trust, delegating decisions to an impenetrable system is particularly problematic. Furlan et al. [55] argue that such opacity can lead to the erosion of this relationship, depriving clinicians of the opportunity for authentic dialogue and shared decision-making, which are the foundation of ethical care. From a legal perspective, AI's opacity challenges traditional accountability frameworks. In the case of a diagnostic error, standard models based on physician culpability become inadequate. As Chau et al. [56] point out, the blurring of responsibility between the clinician, developer, and institution, combined with the "black box" nature of the case, significantly complicates the assessment of whether the physician's conduct met the applicable standard of care. From the perspective of family physicians, the prevailing feeling is that the ultimate responsibility for clinical decisions will always rest with them, underscoring the need for human oversight [57]. Furthermore, the issue of informed consent becomes extremely complex. In cases where AI significantly influences a therapeutic decision, its role may be considered a material fact that the patient should be aware of for their consent to be fully informed [58]. This poses a challenge, as attempting to explain the operation of complex models may be incomprehensible to the patient. In this context, education of both physicians and patients is essential to building a shared understanding of AI's capabilities and limitations. Ueda et al. [59] emphasize that counteracting algorithmic errors, which may result from incomplete or biased training data in psychiatric diagnostics, is also crucial to maintaining fairness. This can lead to perpetuation of inequalities in access to care and stigmatization of specific patient groups. The literature clearly indicates that the current ethical and legal frameworks are insufficient to address the challenges posed by AI in psychiatry [60]. The overarching goal must be to ensure that AI serves as a tool augmenting, rather than replacing, human knowledge and empathy, with clearly defined standards of validation, auditing and oversight that guarantee its safety, effectiveness and fairness, while counteracting the danger of "technologization" of the therapeutic relationship [61].

3.3. Conclusions

The literature review reveals a fundamental tension between the transformative potential of AI in psychiatry and the profound ethical and legal challenges that remain unresolved. A central problem remains the "black box"—the opacity of algorithms, which undermines trust and complicates the attribution of responsibility [54]. This problem is particularly acute in the context of mental health, where diagnostic and treatment decisions are exceptionally sensitive [61]. As researchers note, the inability to understand how AI reaches its conclusions places clinicians in a legally and professionally precarious position, as they are ultimately responsible for decisions, including algorithmic errors [55]. From a legal perspective, traditional civil liability frameworks, which assume human reasoning, prove inadequate in the face of AI's opaque decision-making, which limits courts' ability to assess whether AI-assisted diagnoses meet accepted standards [56]. This problem is further complicated by the unclear division of responsibility between clinicians, healthcare institutions, and system developers [56]. Of particular concern is the risk of automating biases – algorithms may not only reflect existing social inequalities but also amplify them, which is unacceptable in the context of psychiatric care [58]. Furthermore, consent to AI's participation in the diagnostic process requires a new approach to informed consent – patients should be informed of AI's involvement when it introduces significant differences in risk or outcomes [56]. Researchers emphasize that the success of AI in healthcare depends on maintaining humans in the decision-making loop, which maintains accountability and prevents automated errors [60].

Summary and Conclusions

The presented literature review on the use of AI in psychiatric diagnostics reveals a complex picture of a dynamically developing field characterized by significant transformative potential, but also burdened with significant methodological, ethical, and practical limitations. A synthesis of the accumulated scientific evidence allows for the formulation of several key conclusions, both cognitive and applied. First, an analysis of the use of natural language processing (NLP) in digital phenotyping indicates the high potential of these technologies in objectifying and quantifying previously subjective clinical features. Research on psychosis spectrum disorders [12, 18, 19] and autism [17] provides compelling evidence for the detection of subtle linguistic patterns that can serve as biomarkers of risk states and support the diagnostic process. Advanced semantic analysis methods, such as recurrence analysis and word repetition graphs, seem particularly promising, as they allow for the capture of the dynamics of thought disorders with accuracy exceeding that of traditional clinical assessments [18]. At the same time, however, systematic reviews reveal significant limitations of this methodology, including the dominance of binary classification studies at the expense of capturing phenotypic complexity [13], the paucity of resources for languages other than English [14], and the risk of generating inconsistent and hallucinatory responses by large language models [13].

Second, in the field of neuroimaging and imaging biomarkers, a significant gap is observed between the promise of AI and its actual clinical utility. A review of 555 AI models revealed that 83.1% of them have a high risk of bias, primarily due to inadequate sample sizes, lack of calibration assessment, and failure to account for data complexity [32]. Furthermore, none of the models analyzed were deemed ready for direct clinical application, and reporting completeness was insufficient (61.2%) [32]. These issues are consistent with broader observations regarding the low replicability of population-based studies in psychiatric neuroimaging, where effect sizes are small ($r \approx 0.10$) and require samples of thousands of participants for reliable estimates [33]. Despite advances in multimodal and molecular imaging [35], current AI models based on neuroimaging remain in the early stages of methodological development and require significant refinement before being implemented in routine clinical practice. Third, integrating AI into psychiatry raises profound ethical and legal dilemmas, the solution of which goes beyond purely technological issues. A central problem remains the "black box"—the opacity of deep learning algorithms, which undermines the ability to obtain informed consent, complicates attribution of responsibility for diagnostic errors, and contradicts the fundamental principles of bioethics [54, 58]. This problem is particularly acute in the context of mental health, where diagnostic and therapeutic decisions are exceptionally sensitive and the doctor-patient relationship is based on trust [61]. Traditional civil liability frameworks, which assume human reasoning, prove inadequate in the face of AI's opaque decision-making [56], creating the risk of a "responsibility gap." Furthermore, algorithms may not only reflect existing social inequalities but also amplify them, which is unacceptable in the context of psychiatric care [59]. Fourth, despite the aforementioned limitations, there are areas where AI is already demonstrating promising effectiveness. Meta-analyses indicate a cumulative diagnostic accuracy of 85% for AI and a therapeutic efficacy of 84% [8], with EEG-based models demonstrating particularly high efficacy in predicting treatment response in schizophrenia (sensitivity 89%, specificity 94%) [4]. Therapeutic chatbots such as Woebot, Wysa, and Youper demonstrate significant reductions in depression and anxiety symptoms, with high levels of user engagement and satisfaction comparable to traditional therapy [9]. These tools can be a valuable complement to standard interventions, particularly in the context of staff shortages and geographical barriers to accessing psychiatric care [5, 9].

Based on the synthesis, several recommendations for future research and clinical practice can be formulated. First, it is necessary to shift the focus from group-based models to person-centered models, with extensive data collection within each individual, which allows for the capture of reliable individual differences in brain networks and psychopathology [49]. Second, it is necessary to build more diverse, representative research cohorts, standardize data acquisition protocols, and introduce reporting standards such as BIDS [41] or the PROBAST guidelines [32] to enable comparability and replication of research. Third, the development of AI in psychiatry requires the development of a comprehensive ethical and regulatory framework that takes into account the specificity of psychiatric data and advocates for maintaining a human-in-the-loop [60]. Fourth, increasing the transparency of methods, using experimental research designs, and involving users and patients in the design phase of AI tools [6] are crucial. In summary, AI in psychiatric diagnostics is currently in a phase of intensive development, characterized by significant potential but also significant methodological and ethical challenges. NLP and digital phenotyping offer promising opportunities for objectifying clinical assessments, while AI-assisted neuroimaging requires further methodological refinements before translation into practice. Ethical challenges, particularly those related to algorithm opacity and legal liability, require the urgent development of appropriate regulations. The ultimate success of AI in psychiatry will depend on striking a balance between technological innovation and respect for fundamental ethical values, while maintaining the paramount role of the clinician in the diagnostic and therapeutic process. Further research should focus on validating models in large, diverse population cohorts, developing interpretable algorithms, and developing ethical standards that ensure the safe and equitable use of AI's potential in psychiatry.

Conflicts of Interest: No conflicts of interest to declare.

REFERENCES

1. O'Connell, K., et al. (2025). Genomics yields biological and phenotypic insights into bipolar disorder. *Nature, 639*, 968–975. https://doi.org/10.1038/s41586-024-08468-9
2. Sun, C., McEwan, A., Boulton, K. A., Demetriou, E. A., et al. (2025). Artificial intelligence for tracking social behaviours and supporting an autism spectrum disorder diagnosis: Systematic review and meta-analysis. *eBioMedicine, 120*, 105931. https://doi.org/10.1016/j.ebiom.2025.105931
3. Huggins, A. A., Baird, C. L., Briggs, M., Laskowitz, S., Hussain, A., Fouda, S., et al. (2024). Smaller total and subregional cerebellar volumes in posttraumatic stress disorder: A mega-analysis by the ENIGMA-PGC PTSD workgroup. *Molecular Psychiatry, 29*, 611–623. https://doi.org/10.1038/s41380-023-02347-9
4. Saboori Amleshi, R., Ilaghi, M., Rezaei, M., et al. (2025). Predictive utility of artificial intelligence on schizophrenia treatment outcomes: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews, 170*, 105968. https://doi.org/10.1016/j.neubiorev.2024.105968
5. Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly, 93*, 249–253. https://doi.org/10.1007/s11126-022-09973-8
6. Kolding, S., Lundin, R. M., Hansen, L., & Østergaard, S. D. (2025). Use of generative artificial intelligence (AI) in psychiatry and mental health care: A systematic review. *Acta Neuropsychiatrica, 37*, e37. https://doi.org/10.1017/neu.2024.50
7. Dehbozorgi, R., Zangeneh, S., Khooshab, E., Hafezi Nia, D., et al. (2025). The application of artificial intelligence in the field of mental health: A systematic review. *BMC Psychiatry, 25*, 132. https://doi.org/10.1186/s12888-025-06842-9
8. Rony, M. K. K., Das, D. C., Khatun, M. T., Ferdousi, S., Akter, M., et al. (2025). Artificial intelligence in psychiatry: A systematic review and meta-analysis of diagnostic and therapeutic efficacy. *Digital Health, 11*, 20552076251330528. https://doi.org/10.1177/20552076251330528
9. Farzan, M., Ebrahimi, H., Pourali, M., & Sabeti, F. (2025). Artificial intelligence-powered cognitive behavioral therapy chatbots: A systematic review. *Iranian Journal of Psychiatry, 20*(1), 102–110. https://doi.org/10.18502/ijps.v20i1.17395
10. Au Yeung, J., Shek, A., Searle, T., et al. (2024). Natural language processing data services for healthcare providers. *BMC Medical Informatics and Decision Making, 24*(1), 356. https://doi.org/10.1186/s12911-024-02784-w
11. Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., et al. (2021). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research, 23*(5), e15708. https://doi.org/10.2196/15708
12. Bedi, G., Carrillo, F., et al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophrenia, 1*(1), 1–7. https://www.nature.com/articles/npjschz201530
13. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health, 11*, e57400. https://doi.org/10.2196/57400
14. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics, 9*(1), 12. https://doi.org/10.1186/s13326-018-0179-8
15. Rabbani, N., Bedgood, M., Brown, C., et al. (2023). A natural language processing model to identify confidential content in adolescent clinical notes. *Applied Clinical Informatics, 14*(3), 400–407. https://doi.org/10.1055/a-2051-9764
16. Funk, B., Sadeh-Sharvit, S., Fitzsimmons-Craft, E. E., et al. (2020). A framework for applying natural language processing in digital health interventions. *Journal of Medical Internet Research, 22*(2), e13855. https://doi.org/10.2196/13855
17. Trayvick, J., Barkley, S. B., McGowan, A., et al. (2024). Speech and language patterns in autism: Towards natural language processing as a research and clinical tool. *Psychiatry Research, 340*, 116109. https://doi.org/10.1016/j.psychres.2024.116109
18. Argolo, F., Ramos, W. H. P., Mota, N. B., et al. (2024). Natural language processing in at-risk mental states: Enhancing the assessment of thought disorders and psychotic traits with semantic dynamics and graph theory. *Brazilian Journal of Psychiatry, 46*, e20233419. https://doi.org/10.47626/1516-4446-2023-3419

19. Deneault, A., Dumais, A., Désilets, M., & Hudon, A. (2024). Natural language processing and schizophrenia: A scoping review of uses and challenges. *Journal of Personalized Medicine, 14*(7), 744. https://doi.org/10.3390/jpm14070744
20. Tonon, A. C., Nexha, A., da Silva, M. M., Gomes, F. A., et al. (2024). Sleep and circadian disruption in bipolar disorders: From psychopathology to digital phenotyping in clinical practice. *Psychiatry and Clinical Neurosciences*. https://doi.org/10.1111/pcn.13729
21. Constant, A., Koksal, C. E., & Palaniyappan, L. (2026). Setting digital psychiatry in motion: Towards dynamic digital markers for digital phenotyping. *NPP—Digital Psychiatry and Neuroscience, 4*(1), 6. https://doi.org/10.1038/s44277-026-00059-y
22. Kamath, J., Leon, B. R., Jain, N., Keisari, E., & Wang, B. (2022). Digital phenotyping in depression diagnostics: Integrating psychiatric and engineering perspectives. *World Journal of Psychiatry, 12*(3), 393–409. https://doi.org/10.5498/wjp.v12.i3.393
23. Slack, S. K., & Barclay, L. (2023). First-person disavowals of digital phenotyping and epistemic injustice in psychiatry. *Medicine, Health Care and Philosophy, 26*, 605–614. https://doi.org/10.1007/s11019-023-10174-8
24. Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S., & Adrien, V. (2023). Digital phenotyping: Data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research, 25*, e44502. https://doi.org/10.2196/44502
25. Shen, F. X., Baum, M. L., Martinez-Martin, N., Miner, A. S., Abraham, M., Brownstein, C. A., Cortez, N., Evans, B. J., Germine, L. T., Glahn, D. C., Grady, C., Holm, I. A., Hurley, E. A., Kimble, S., Lázaro-Muñoz, G., Leary, K., Marks, M., Monette, P. J., Onnela, J. P., . . . Silverman, B. C. (2024). Returning individual research results from digital phenotyping in psychiatry. *The American Journal of Bioethics, 24*(2), 69–90. https://doi.org/10.1080/15265161.2023.2180109
26. Sterling, W. A., Sobolev, M., Van Meter, A., Guinart, D., Birnbaum, M. L., Rubio, J. M., & Kane, J. M. (2022). Digital technology in psychiatry: Survey study of clinicians. *JMIR Formative Research, 6*(11), e33676. https://doi.org/10.2196/33676
27. Orsolini, L., Fiorani, M., & Volpe, U. (2020). Digital phenotyping in bipolar disorder: Which integration with clinical endophenotypes and biomarkers? *International Journal of Molecular Sciences, 21*(20), 7684. https://doi.org/10.3390/ijms21207684
28. Washington, P., Park, N., Srivastava, P., et al. (2020). Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(8), 759–769. https://doi.org/10.1016/j.bpsc.2019.11.015
29. Liu, J. J., Borsari, B., Li, Y., et al. (2025). Digital phenotyping from wearables using AI characterizes psychiatric disorders and identifies genetic associations. *Cell, 188*(2), 515–529.e15. https://doi.org/10.1016/j.cell.2024.11.012
30. Raugh, I. M., James, S. H., Gonzalez, C. M., & Chapman, H. C. (2021). Digital phenotyping adherence, feasibility, and tolerability in outpatients with schizophrenia. *Journal of Psychiatric Research, 138*, 436–443. https://doi.org/10.1016/j.jpsychires.2021.04.022
31. Mouchabac, S., Conejero, I., Lakhlifi, C., Msellek, I., Malandain, L., Adrien, V., . . . Maatoug, R. (2021). Usprawnianie procesu podejmowania decyzji klinicznych w psychiatrii: Wdrożenie fenotypowania cyfrowego może złagodzić wpływ indywidualnych uprzedzeń poznawczych pacjentów i lekarzy. *Dialogues in Clinical Neuroscience, 23*(1), 52–61. https://doi.org/10.1080/19585969.2022.2042165
32. Chen, Z., Liu, X., Yang, Q., Wang, Y., Miao, K., Gong, Z., Yu, Y., Leonov, A., Liu, C., Feng, Z., & Hu, C.-P. (2023). Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: A systematic review. *JAMA Network Open, 6*(3), e231671. https://doi.org/10.1001/jamanetworkopen.2023.1671
33. Marek, S., & Laumann, T. O. (2024). Replicability and generalizability in population psychiatric neuroimaging. *Neuropsychopharmacology, 50*(1), 52–57. https://doi.org/10.1038/s41386-024-01960-w
34. Vogel, A. C., & Black, K. J. (2024). Brain imaging in routine psychiatric practice. *Missouri Medicine, 121*(1), 37–43. https://pmc.ncbi.nlm.nih.gov/articles/PMC10887461/
35. Cope, T. E., Weil, R. S., Düzel, E., Dickerson, B. C., & Rowe, J. B. (2020). Advances in neuroimaging to support translational medicine in dementia. *Journal of Neurology, Neurosurgery & Psychiatry, 91*(8), 882–894. https://doi.org/10.1136/jnnp-2019-322402
36. Ersoy, S., Ersoy, E. H., Danis, A., & Turkoglu, S. A. (2025). Trends and global productivity in artificial intelligence research in clinical neurology and neuroimaging: A bibliometric analysis from 1980 to 2024. *Cerebral Cortex, 35*(6), bhaf148. https://doi.org/10.1093/cercor/bhaf148
37. Seriramulu, V. P., Suppiah, S., & Lee, H. H. (2024). Review of MR spectroscopy analysis and artificial intelligence applications. *Medical Journal of Malaysia, 79*(1), 103–109. https://www.e-mjm.org/2024/v79n1/alzheimers-disease.pdf
38. Tejavibulya, L., Rolison, M., Gao, S., Liang, Q., Peterson, H., Dadashkarimi, J., Farruggia, M. C., Hahn, C. A., Noble, S., Lichenstein, S. D., Pollatou, A., Dufford, A. J., & Scheinost, D. (2022). Predicting the future of neuroimaging predictive models in mental health. *Molecular Psychiatry, 27*(8), 3129–3137. https://doi.org/10.1038/s41380-022-01635-2

39. Kotochinsky, M., Fonseca, P. E. O., Lopera, V. R., Mora, L., Amador, W. F. O., Sirena, E. C. T., Guimarães, F. B. M., Herlyn, D. L., Sherpa, N. N., Lezana, A. G., & Fagundes, T. P. (2025). Comparative diagnostic performance of artificial intelligence models in structural MRI for schizophrenia: A systematic review and meta-analysis. *Asian Journal of Psychiatry, 107*, 104759. https://doi.org/10.1016/j.ajp.2025.104759
40. Zhou, E., Wang, W., Ma, S., Xie, X., Kang, L., Xu, S., Deng, Z., Gong, Q., Nie, Z., Yao, L., Bu, L., Wang, F., & Liu, Z. (2023). Prediction of anxious depression using multimodal neuroimaging and machine learning. *NeuroImage, 285*, 120499. https://doi.org/10.1016/j.neuroimage.2023.120499
41. Leenings, R., Winter, N. R., Dannlowski, U., & Hahn, T. (2022). Recommendations for machine learning benchmarks in neuroimaging. *NeuroImage, 257*, 119298. https://doi.org/10.1016/j.neuroimage.2022.119298
42. Monaco, F., Vignapiano, A., Di Gruttola, B., Landi, S., et al. (2025). Neuroimaging and machine learning in eating disorders: A systematic review. *Eating and Weight Disorders—Studies on Anorexia, Bulimia and Obesity, 30*, 46. https://doi.org/10.1007/s40519-025-01720-1
43. Grosenick, L., & Liston, C. (2026). Multimodal representation learning for parsing biological heterogeneity in psychiatric neuroimaging. *Biological Psychiatry, 99*(10), 862–873. https://www.biologicalpsychiatryjournal.com/article/S0006-3223(26)00086-7/abstract
44. Wen, J., Skampardoni, I., Tian, Y. E., Yang, Z., Cui, Y., Erus, G., Hwang, G., Varol, E., Boquet-Pujadas, A., Chand, G. B., Nasrallah, I. M., Satterthwaite, T. D., Shou, H., Shen, L., Toga, A. W., Zalesky, A., & Davatzikos, C. (2025). Neuroimaging endophenotypes reveal underlying mechanisms and genetic factors contributing to progression and development of four brain disorders. *Nature Biomedical Engineering, 9*, 1920–1937. https://doi.org/10.1038/s41551-025-01310-7
45. Kam, H., & Jeong, H. (2020). Pharmacogenomic biomarkers and their applications in psychiatry. *Genes, 11*(12), 1445. https://doi.org/10.3390/genes11121445
46. Kraguljac, N. V., McDonald, W. M., Widge, A. S., Rodriguez, C. I., Tohen, M., & Nemeroff, C. B. (2021). Neuroimaging biomarkers in schizophrenia. *American Journal of Psychiatry, 178*(6), 509–521. https://doi.org/10.1176/appi.ajp.2020.20030340
47. Golub, A., Ordak, M., Nasierowski, T., & Bujalska-Zadrozny, M. (2023). Advanced biomarkers of hepatotoxicity in psychiatry: A narrative review and recommendations for new psychoactive substances. *International Journal of Molecular Sciences, 24*(11), 9413. https://doi.org/10.3390/ijms24119413
48. Parellada, M., Andreu-Bernabeu, Á., Burdeus, M., San José Cáceres, A., Urbiola, E., Carpenter, L. L., Kraguljac, N. V., McDonald, W. M., Nemeroff, C. B., Rodriguez, C. I., Widge, A. S., & Sanders, S. J. (2023). In search of biomarkers to guide interventions in autism spectrum disorder: A systematic review. *American Journal of Psychiatry, 180*(1), 50–61. https://doi.org/10.1176/appi.ajp.21100992
49. Kraus, B., Zinbarg, R., Braga, R. M., Nusslock, R., Mittal, V. A., & Gratton, C. (2023). Insights from personalized models of brain and behavior for identifying biomarkers in psychiatry. *Neuroscience & Biobehavioral Reviews, 152*, 105259. https://doi.org/10.1016/j.neubiorev.2023.105259
50. Morrin, H., & Nour, M. M. (2026). The promise of artificial intelligence–powered speech biomarkers in psychiatry. *JAMA Network Open, 9*(6), e2620251. https://doi.org/10.1001/jamanetworkopen.2026.20251
51. Lin, E., Lin, C.-H., & Lane, H.-Y. (2020). Precision psychiatry applications with pharmacogenomics: Artificial intelligence and machine learning approaches. *International Journal of Molecular Sciences, 21*(3), 969. https://doi.org/10.3390/ijms21030969
52. Winchester, L. M., Harshfield, E. L., Shi, L., Badhwar, A., Al Khleifat, A., Clarke, N., Dehsarvi, A., Lengyel, I., Lourida, I., Madan, C. R., Marzi, S. J., Proitsi, P., Rajkumar, A. P., Rittman, T., Silajdžić, E., Tamburin, S., Ranson, J. M., & Llewellyn, D. J. (2023). Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia. *Alzheimer's & Dementia, 19*(12), 5860–5871. https://doi.org/10.1002/alz.13390
53. Baydili, İ., Tasci, B., & Tasci, G. (2025). Artificial intelligence in psychiatry: A review of biological and behavioral data analyses. *Diagnostics, 15*(4), 434. https://doi.org/10.3390/diagnostics15040434
54. Khalique, N., Nasir, M., Ahmed, S., & Siddiqui, K. (2026). The role of artificial intelligence in improving the public health care delivery system in India: A legal-ethical audit. *Indian Journal of Community Medicine, 51*(1), 32–41. https://doi.org/10.4103/ijcm.ijcm_7_25
55. Furlan, E., Zamperetti, N., Piccinni, M., Marra, A., Cotogni, P., & Giannini, A. (2026). The Ethics Committee of the Italian Society of Anesthesia, Analgesia, Resuscitation and Intensive Care (SIAARTI)—Artificial intelligence in end-of-life decision-making processes: Ethical reflections. *Journal of Anesthesia, Analgesia and Critical Care, 6*, Article 72. https://doi.org/10.1186/s44158-026-00215-6
56. Chau, M. T., Spuur, K. M., White, S., Pyper, A., & Crossman, M. (2026). Malpractice in the machine age: Legal and ethical responses to machine learning in medical imaging. *Radiography, 32*(3). https://www.radiographyonline.com/article/S1078-8174(26)00015-5/fulltext
57. Ratkevičiūtė, K., & Aliukonis, V. (2025). Exploring opportunities and challenges of AI in primary healthcare: A qualitative study with family doctors in Lithuania. *Healthcare, 13*(12), 1429. https://doi.org/10.3390/healthcare13121429

58. Cohen, I. G., & Slottje, A. (2025). Artificial intelligence and the law of informed consent. In *Research handbook on health, AI and the law*. Edward Elgar Publishing. https://www.ncbi.nlm.nih.gov/books/NBK613199/
59. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology, 42*, 3–15. https://doi.org/10.1007/s11604-023-01547-x
60. Khan, M. M., Shah, N., Shaikh, N., Thabet, A., Alrabayah, T., & Belkhair, S. (2025). Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges. *International Journal of Medical Informatics, 195*, 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780
61. Rahsepar Meadi, M., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the ethical challenges of conversational AI in mental health care: Scoping review. *JMIR Mental Health, 12*, e60432. https://doi.org/10.2196/60432