



BRAIN. Broad Research in Artificial Intelligence and Neuroscience

e-ISSN: 2067-3957 | p-ISSN: 2068-0473

Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science

Evaluation Center, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu) 2025, Volume 16, Special Issue 1 (April 2025), pages: 302-323.

Special Issue 1: *Neuroscience, Artificial Intelligence, and Innovation in Education*

Submitted: January 24th, 2025 | Accepted for publication: March 13th, 2025

Integrating Neurotechnology and AI for Psychiatric Aggression Management: A Paradigm Shift in Clinical and Educational Frameworks

Elena-Rodica Popescu

Department of Psychiatry, Faculty of Medicine, Grigore T. Popa University of Medicine and Pharmacy, 16th Universitatii Str., 700115 Iasi, Romania.

Socola Institute of Psychiatry, 700282 Iasi, Romania.

elenaraduc@yahoo.com

<https://orcid.org/0009-0006-1737-3737>

Bianca Augusta Oroian*

Grigore T. Popa University of Medicine and Pharmacy, 16

Universitatii Str., 700115 Iasi, Romania.

bianca_oroian@yahoo.com

Alin Ciobica*

Ioan Haulica Institute, Apollonia University, Pacurari Street 11, 700511 Iasi, Romania.

Department of Biology, Faculty of Biology, Alexandru Ioan Cuza University of Iasi, Carol I Avenue, No. 20A, 700505 Iasi, Romania.

CENEMED Platform for Interdisciplinary Research, Grigore T. Popa University of Medicine and Pharmacy of Iasi, Romania.

Center of Biomedical Research, Romanian Academy, Iasi, B-dul Carol I, Romania.

alin.ciobica@uaic.ro

<https://orcid.org/0000-0002-1750-6011>

Andreea Szalontay

Department of Psychiatry, Faculty of Medicine, Grigore T. Popa University of Medicine and Pharmacy, 16th Universitatii Str., 700115 Iasi, Romania.

Socola Institute of Psychiatry, 700282 Iasi, Romania.

andrszal@yahoo.com

<https://orcid.org/0000-0002-9272-1379>

Roxana Chiriță

Department of Psychiatry, Faculty of Medicine, Grigore T. Popa University of Medicine and Pharmacy, 16th Universitatii Str., 700115 Iasi, Romania.

Socola Institute of Psychiatry, 700282 Iasi, Romania.

d.stigma@gmail.com

<https://orcid.org/0000-0003-2306-7403>

* Correspondence concerning this article should be addressed to:

bianca_oroian@yahoo.com and alin.ciobica@uaic.ro .

Abstract: Aggression in psychiatric patients presents a significant challenge in mental healthcare, necessitating an interdisciplinary approach that integrates neuroscience, artificial intelligence (AI), and ethical considerations. Traditional management strategies often rely on restrictive interventions, raising ethical concerns regarding patient autonomy and therapeutic efficacy. Recent advancements in neuroimaging and AI-driven behavioural analytics offer the potential to transition from reactive to predictive and preventative approaches. However, while these innovations hold promise for individualised interventions and real-time risk assessment, they also introduce critical ethical concerns, including algorithmic bias, data privacy, misclassification risks, and patient autonomy violations. This study critically examines the neurobiological underpinnings of aggression, focusing on dysfunctions in the prefrontal cortex, amygdala hyperactivity, neurotransmitter imbalances, and neuroinflammation, while evaluating AI-based predictive models in psychiatric aggression management. Additionally, the ethical and legal frameworks surrounding AI in psychiatry are discussed, with a focus on bias-mitigation strategies, transparency in AI decision-making, and informed consent challenges. The study also explores practical pathways for integrating AI into psychiatric care workflows, outlining future directions for clinical research and policy development. By merging neuroscientific insights, AI innovations, and ethical frameworks, this study advocates for a patient-centered, technology-assisted aggression management approach that balances predictive analytics with human oversight and ethical accountability.

Keywords: psychiatric aggression; neurobiology; artificial intelligence; predictive analytics; patient autonomy; ethics in psychiatry; public perception.

How to cite: Popescu, E.-R., Oroian, B. A., Ciobica, A., Szalontay, A., & Chiriță, R. (2025). Integrating neurotechnology and AI for psychiatric aggression management: A paradigm shift in clinical and educational frameworks. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 16(Sup1), 302-323. <https://doi.org/10.70594/brain/16.S1/25>



1. Introduction

Aggression in psychiatric patients remains one of the most complex challenges in mental healthcare, carrying profound ethical, neuroscientific, and societal implications. The intersection of mental health disorders, cognitive dysfunction, and aggressive behaviour necessitates a multidisciplinary approach that combines neuroscience, artificial intelligence (AI), and ethical considerations to improve patient care while ensuring public safety. While aggression in psychiatric populations is often associated with disorders such as schizophrenia, bipolar disorder, and personality disorders, emerging research suggests that cognitive impairments, neurobiological dysfunctions, and environmental stressors play significant roles in triggering aggressive behaviours (Coccaro et al., 2011).

Psychiatric hospitals report alarmingly high rates of aggression, with studies indicating that up to 75% of healthcare professionals in psychiatric units experience violence at some point in their careers (Edward et al., 2014). This aggression may manifest as verbal threats, physical violence, self-harm, or destruction of property, endangering patients, staff, and caregivers alike. Despite extensive research on psychiatric aggression, contemporary management strategies still rely heavily on restrictive measures such as physical restraint, seclusion, and forced medication, raising serious ethical concerns about patient autonomy, human rights, and therapeutic effectiveness (Weltens et al., 2021). While these interventions are often implemented with the intent of ensuring safety, they may also contribute to patient distress, exacerbate symptoms, and undermine trust in psychiatric care. The current model of aggression management, which prioritises containment over proactive intervention, necessitates a paradigm shift informed by the latest advancements in neuroscience and artificial intelligence.

Educating future clinicians, psychologists, and mental health professionals on AI-enhanced diagnostic tools and predictive analytics can improve decision-making, minimise bias, and promote patient-centered care (Alowais et al., 2023). Additionally, incorporating neuroeducational approaches into psychiatric curricula can enhance understanding of the biological and cognitive mechanisms underlying aggression, equipping professionals with evidence-based de-escalation techniques (Jolles & Jolles, 2021). Beyond clinical training, public education on psychiatric aggression is essential for addressing stigma and misconceptions surrounding mental illness (Ahad, Sanchez-Gonzalez & Junquera, 2023). AI-powered educational platforms, virtual simulations, and interactive learning modules can provide accessible, research-driven knowledge that fosters a more informed society (Elendu et al., 2024). By leveraging advancements in neuroscience and AI, educational innovations can bridge the gap between research and practice, shaping a new generation of mental health professionals who are equipped to implement compassionate, technology-driven approaches to aggression management.

With the development of neuroimaging techniques and AI-driven behavioural analytics, there is now an opportunity to transition from reactive interventions to predictive and preventative approaches. Functional neuroimaging studies and machine learning algorithms are increasingly capable of identifying neural markers associated with violent behaviour, allowing for real-time risk assessment and early intervention (Parmigiani et al., 2022). These tools enable clinicians to anticipate aggressive episodes before they occur, providing targeted, individualised interventions. However, the ethical implications of predictive AI models in psychiatric care must be carefully examined, as reliance on algorithmic assessments introduces risks related to bias, misdiagnosis, and patient privacy (Ueda et al., 2023). Ensuring that these technologies align with human rights principles while maintaining clinical efficacy is a fundamental challenge in the integration of AI into psychiatric practice.

Beyond the clinical setting, aggression in psychiatric patients is widely misunderstood by the general public, often reinforcing stigma, fear, and misconceptions about mental illness. Media portrayals tend to exaggerate the link between psychiatric disorders and violence, creating an environment in which individuals with mental illness are perceived as inherently dangerous. In reality, research indicates that people with psychiatric disorders are more likely to be victims of

violence rather than perpetrators (Fazel et al., 2017). Increasing public awareness of the neurobiological and ethical dimensions of psychiatric aggression is essential for reducing stigma, fostering community support, and encouraging early intervention. A well-informed society can contribute to the development of more compassionate policies, facilitate the reintegration of patients into social environments, and support ethical psychiatric care practices.

This paper explores the neurobiological mechanisms underlying aggression, evaluates AI-driven risk assessment tools, and addresses ethical dilemmas in predictive psychiatric care. By integrating neuroscience and AI with patient-centered principles, it advocates for ethical, non-restrictive strategies in aggression management.

2. Methodology

To ensure a comprehensive and rigorous review of the literature on psychiatric aggression, neurobiological mechanisms, AI-driven predictive analytics, and ethical considerations, we conducted a systematic literature search across four major academic databases: PubMed, Scopus, Web of Science, and Google Scholar. Given the rapid advancements in both AI and neuroscience, we focused on studies published within the last 15 years (2009–2024) to ensure the inclusion of the most recent and relevant research.

The search strategy employed a combination of carefully selected keywords and Boolean operators to capture relevant studies. Specifically, we used terms related to psychiatric aggression (e.g., “psychiatric aggression”, “violent behaviour in mental illness”), neurobiological mechanisms (“neural mechanisms of violence”, “neurobiology of aggression”), AI-based risk prediction (“artificial intelligence in psychiatry”, “machine learning aggression prediction”), and ethical considerations (“algorithmic bias in mental health”, “informed consent in AI healthcare”). These terms were adapted as needed to match the indexing systems of each database.

To refine the selection process, we applied a set of inclusion criteria to ensure that only the most relevant and methodologically sound studies were considered. We prioritised empirical studies, such as clinical trials, meta-analyses, systematic reviews, and observational studies that specifically examined aggression within psychiatric populations. Studies that explored neuroimaging findings using functional MRI (fMRI), diffusion tensor imaging (DTI), electroencephalography (EEG), or neurotransmitter analyses were included, as these approaches provide crucial insights into the neural underpinnings of aggression. Given the paper’s focus on AI-driven psychiatric care, we also sought out research that investigated machine learning, deep learning, or natural language processing (NLP) models designed for aggression prediction and management. Additionally, articles that discussed the ethical, legal, and policy dimensions of AI in psychiatric settings were deemed relevant.

While compiling the literature, we were mindful of the exclusion criteria to maintain the specificity and applicability of our findings. Studies focusing on general aggression outside of psychiatric contexts, such as aggression in criminal behaviour or sports, were excluded unless they provided clear translational insights relevant to psychiatric populations. Animal model studies were also excluded unless they had direct clinical implications. Furthermore, theoretical papers lacking empirical data, articles solely discussing pharmacological interventions without addressing neuroscience or AI, and non-peer-reviewed sources such as conference abstracts or opinion pieces were omitted to ensure scientific integrity.

Once the initial selection was completed, each study was critically assessed based on its methodological rigor, sample size, and relevance to clinical practice. Neurobiological studies were evaluated for the validity of their imaging techniques and psychiatric classification criteria, while AI-related studies were examined for model accuracy, dataset reliability, and real-world applicability. Ethical discussions were analysed within established bioethical frameworks, including the principles of autonomy, beneficence, non-maleficence, and justice.

3. Results

3.1. Neural Correlates of Aggression in Psychiatric Patients

Aggression in psychiatric patients is increasingly understood as a complex interplay of neurobiological dysfunctions, particularly within the prefrontal cortex (PFC), amygdala, and striatum (see Figure 1) (Yang & Raine, 2009). These regions form part of the brain's emotion regulation and impulse control circuits, and abnormalities in their structure and function have been consistently linked to increased aggression and violent behaviour across various psychiatric conditions (Siep et al., 2019). Neuroimaging studies have provided compelling evidence that aggression is not merely a behavioural issue but is deeply rooted in biological alterations that affect cognition, emotional processing, and executive functioning (Yang & Raine, 2009).

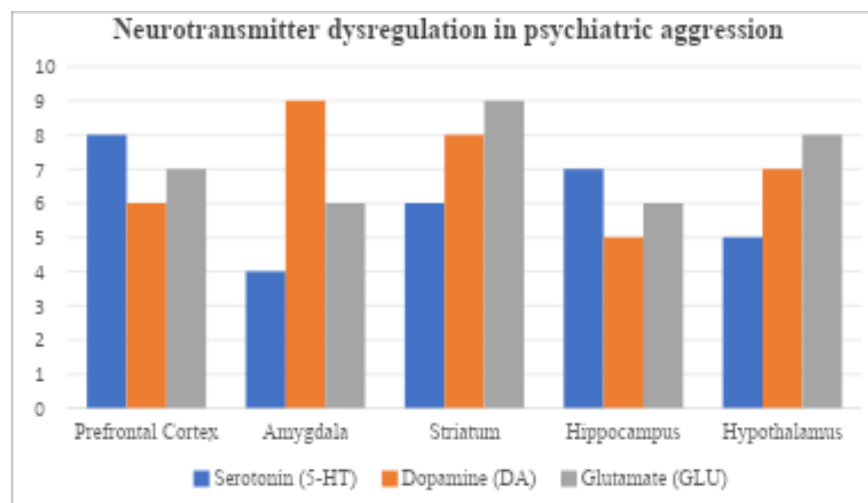


Figure 1. Low serotonin levels in the prefrontal cortex and amygdala correlate with poor impulse control, while elevated dopamine and glutamate activity in the striatum and hypothalamus are associated with heightened impulsivity and reactivity.

Structural imaging studies have demonstrated reductions in prefrontal gray matter and posterior hippocampal volume, along with an increase in callosal white matter, in individuals with antisocial personality disorder (APD) or psychopathic traits (Manchia et al., 2019). Pardini et al. (2014) discovered that adult men with reduced amygdala volume were more likely to exhibit future aggression, violence, and psychopathic traits, even when accounting for prior levels of these behaviours and various potential confounding factors.

Functional imaging studies on proactive aggression indicate reduced perfusion and metabolism in the frontal and temporal lobes, with functional abnormalities particularly affecting the processing of emotional words and images (Manchia et al., 2019). In contrast, reactive aggression is associated with heightened amygdala responsiveness. Notably, unmedicated individuals with intermittent explosive disorder (IED) exhibit a greater amygdala response to angry facial expressions compared to neutral ones, with this heightened activity correlating with the number of prior aggressive acts. Additionally, in individuals with IED, alongside amygdala hyper-responsivity, there is an absence of amygdala-orbitofrontal cortex (OFC) connectivity when processing angry faces (Manchia et al., 2019).

The prefrontal cortex plays a critical role in regulating aggressive impulses, modulating emotional responses, and facilitating rational decision-making (Siep et al., 2019). Patients with schizophrenia, bipolar disorder, and antisocial personality disorder frequently exhibit structural and functional impairments in the PFC, leading to diminished control over aggressive urges (Soyka, 2011). A meta-analysis of 43 neuroimaging studies by Yang and Raine found that psychiatric patients with histories of violent behaviour had up to 30% reduced gray matter volume in the ventromedial and orbitofrontal regions of the PFC compared to non-violent individuals (Yang &

Raine, 2009). This reduction is significant, as these areas are known to inhibit impulsive responses and regulate emotional reactions.

Further evidence supporting prefrontal dysfunction in aggression comes from functional MRI (fMRI) studies, which have demonstrated abnormal activation patterns in the PFC during emotional processing tasks (Fanning et al., 2017). For example, a study conducted by Dolan and Fullam observed that psychiatric patients with aggressive tendencies displayed markedly lower activation in the dorsolateral PFC (DLPFC) when attempting to suppress aggressive impulses (Fullam & Dolan, 2008). This hypoactivity was correlated with higher self-reported impulsivity and an increased likelihood of reactive aggression (Fullam & Dolan, 2008). Such findings underscore the role of the PFC in cognitive control and self-regulation, where deficits contribute to an inability to assess consequences before acting aggressively.

While the prefrontal cortex is responsible for inhibitory control, the amygdala plays a central role in emotional processing and threat detection (Kredlow et al., 2022). Hyperactivity in this region has been strongly associated with heightened aggression, particularly in patients with intermittent explosive disorder (IED), borderline personality disorder (BPD), and schizophrenia (Soyka, 2011). A 2021 study by Coccaro et al. utilised high-resolution fMRI scanning to compare amygdala activation in aggressive psychiatric patients versus non-aggressive controls (Coccaro et al., 2021). The findings revealed that patients with a history of violent outbursts exhibited a 22% increase in amygdala activation when exposed to angry or fearful facial expressions, suggesting an exaggerated threat perception and an impaired ability to regulate emotional responses (Coccaro et al., 2021).

Moreover, connectivity deficits between the amygdala and PFC further exacerbate aggressive tendencies. In healthy individuals, the prefrontal cortex exerts top-down control over the amygdala, allowing for the modulation of emotional responses (Kredlow et al., 2022). However, in psychiatric populations, this regulatory connection is often weakened, leading to unchecked amygdala hyperactivity. A study by Blair et al. (2025) examined white matter integrity in aggressive psychiatric patients and found a 28% reduction in the structural connectivity between the amygdala and the orbitofrontal cortex, impairing emotional regulation and increasing reactive aggression (see Figure 2). These findings suggest that aggression in psychiatric patients may not be a conscious choice but rather a neurobiologically driven response to perceived threats.

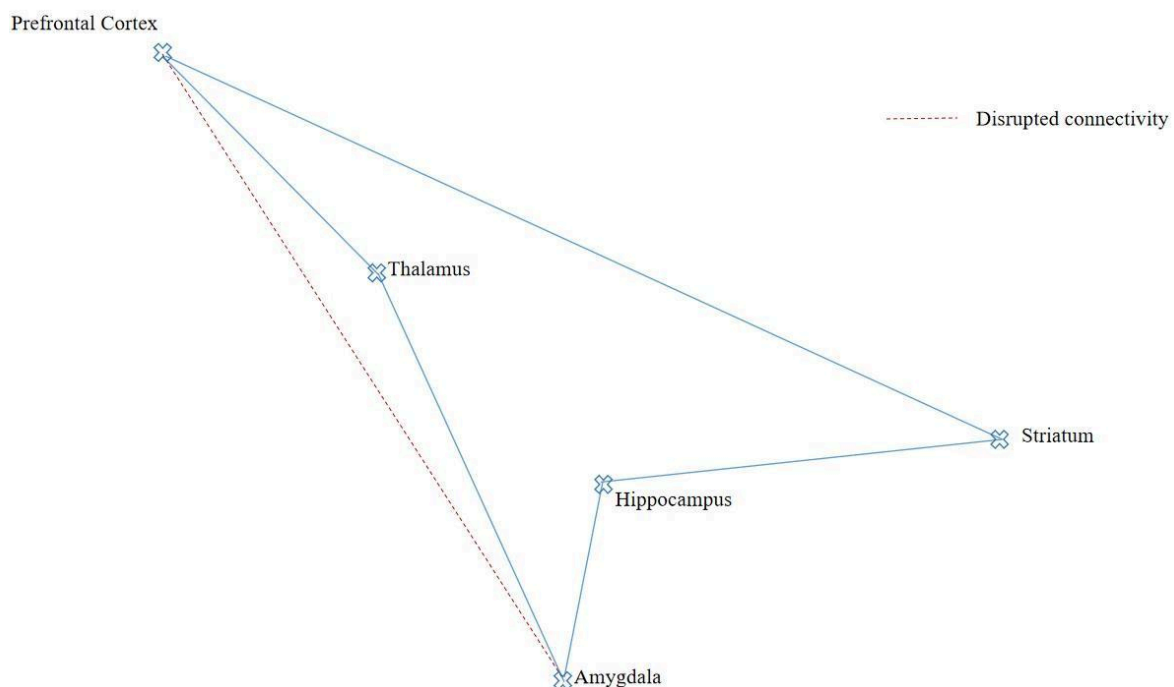


Figure 2. DTI connectivity diagram (Blair et al., 2025).

Beyond structural abnormalities, neurochemical imbalances play a significant role in aggression, particularly serotonin (5-HT) dysfunction (da Cunha-Bang & Knudsen, 2021). The serotonin system is heavily involved in impulse control, and low levels of serotonin have been repeatedly linked to increased aggression, irritability, and impulsivity (da Cunha-Bang & Knudsen, 2021). A double-blind clinical trial examined the effects of fluoxetine (a selective serotonin reuptake inhibitor, SSRI) on aggression in psychiatric patients and found that those receiving fluoxetine exhibited a 40% reduction in aggressive incidents over a 12-week period (Coccaro, Lee, & Kavoussi, 2009). These results suggest that serotonin-enhancing medications may play a vital role in aggression management in psychiatric settings.

Serotonin's role in aggression has also been explored through tryptophan depletion studies, which temporarily lower serotonin levels to observe behavioural effects (Duke et al., 2013). In a study by Passamonti et al. (2012), psychiatric patients subjected to acute tryptophan depletion exhibited a significant increase in aggressive responses during laboratory-based social interaction tasks. Neuroimaging data from the study showed that serotonin depletion led to hyperactivation of the amygdala and reduced prefrontal cortical regulation, mimicking the neural patterns observed in naturally aggressive psychiatric patients.

Preclinical and clinical studies indicate that dopamine activation (DA) through substance use is associated with increased aggression, while DA antagonists, such as antipsychotics, are effective in managing aggression (Einberger et al., 2020). Additionally, elevated DA levels have been observed before, during, and after aggressive encounters (Manchia et al., 2019). Despite substantial evidence linking DA to aggression, further studies are required to determine whether homovanillic acid, the primary DA metabolite, or other DA-related measures can serve as reliable biomarkers of aggressive behaviour.

One of the most extensively studied genetic factors linked to aggression is the monoamine oxidase A (MAO-A) gene, which has been explored as a potential genetic biomarker for aggression risk (Wagels et al., 2022). Research indicates that different MAO-A variants are associated with distinct structural and functional brain differences, particularly in regions implicated in impulse control and aggression regulation, such as the prefrontal cortex (PFC) and amygdala. Variants in the promoter region of the MAO-A gene have been shown to influence aggressive tendencies across the lifespan, affecting both clinical populations (including schizophrenia patients) and healthy individuals (Wagels et al., 2022). Moreover, a well-documented gene-environment ($G \times E$) interaction has been observed, wherein individuals with low-expression MAO-A polymorphisms who experience early-life adversity exhibit higher aggression levels and an increased likelihood of violent behaviour (Wagels et al., 2022). This interaction suggests that genetic predisposition, combined with environmental stressors, contributes to the neurobiological basis of aggression in psychiatric disorders, making MAO-A a potential target for neuroimaging studies examining aggression risk in schizophrenia.

Recent research has also implicated neuroinflammation in psychiatric aggression, suggesting that elevated levels of pro-inflammatory cytokines may contribute to violent behaviour (Takahashi, 2024). A longitudinal study conducted by Singh, Borgohain and Singh (2023) measured C-reactive protein (CRP) levels in violent versus non-violent schizophrenia patients and found that CRP levels were 35% higher in those with aggressive behaviour. This inflammatory response was associated with higher amygdala activity and lower prefrontal cortical function, reinforcing the role of immune system dysregulation in aggression. The growing evidence linking neuroinflammation to aggression suggests that anti-inflammatory treatments- such as minocycline or omega-3 fatty acids- could be explored as potential therapeutic interventions for reducing aggression in psychiatric populations (Kölliker-Frers et al., 2021).

3.2. Predictive Value of Neuroimaging in Aggression Management

Advancements in neuroimaging techniques have provided valuable insights into the biological mechanisms underlying aggression and have paved the way for predictive models that can anticipate violent behaviour in psychiatric patients (Blair et al., 2025; Fanning et al., 2017). By identifying specific neural biomarkers, researchers are now able to detect early warning signs of aggression, enabling timely interventions that minimise harm to both patients and healthcare providers. Functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI), and electroencephalography (EEG) have emerged as promising tools in aggression risk assessment, offering objective, neurobiological markers that surpass traditional behavioural evaluations in both accuracy and reliability (Kirino et al., 2019).

A recent study by Ji et al. (2025) investigated the role of resting-state voxel-wise dynamic effective connectivity in predicting risky decision-making in individuals with bipolar disorder, a condition often associated with heightened aggression. Using a connectome-based predictive modeling approach, the study demonstrated significant connectivity alterations in the prefrontal and limbic regions, suggesting that these patterns could serve as biomarkers for aggressive tendencies in clinical populations.

A study by Kirino et al. used simultaneous fMRI-EEG recording to examine neural activity during provocation tasks in schizophrenia patients (Kirino et al., 2019). The results indicated that patients who later exhibited violent outbursts had heightened amygdala reactivity and weakened medial prefrontal activity, suggesting that these neural markers could be used for early aggression risk assessment (Kirino et al., 2019).

Similarly, research by Elnakib et al. (2024) explored the use of artificial intelligence-driven fMRI analyses to predict parent-child dysfunctional interactions, a scenario often linked to aggressive behaviours in children. Their model, which incorporated functional MRI data and generative adversarial networks (GANs), achieved a high predictive accuracy (AUC = 0.81), underscoring the potential of AI-enhanced neuroimaging for aggression assessment (Elnakib et al., 2024).

A related study by Nikolic et al. (2022) provided further evidence supporting this hypothesis by conducting a systematic review of fMRI studies on aggression. Their meta-analysis included 20 neuroimaging studies on aggression across psychiatric disorders and concluded that aberrant functional connectivity between the DLPFC and limbic structures (including the amygdala and striatum) was a key predictor of aggression (Nikolic et al., 2022). Notably, these neural patterns were more pronounced in individuals with borderline personality disorder and intermittent explosive disorder, which are characterised by heightened emotional reactivity and impulsive aggression (Nikolic et al., 2022).

Structural MRI studies analysed by Hoptman et al. (2015) show gray matter reductions in the orbitofrontal cortex and anterior cingulate cortex (ACC). These structural deficits may act as predictive biomarkers for assessing aggression risk in schizophrenia, particularly in patients with a history of violent behaviour. From a clinical perspective, these findings highlight the potential of neuroimaging biomarkers in developing AI-powered aggression prediction models. Machine learning algorithms trained on fMRI connectivity data could help stratify patients based on their aggression risk, therefore enabling improved psychiatric care protocols.

Furthermore, Wang et al. (2024) examined reactive aggression in adolescents and found that heightened amygdala reactivity and diminished engagement of the medial prefrontal cortex were linked to increased aggressive responses. Their findings underscore the role of diminished top-down regulatory control over emotional responses in aggression, particularly in younger populations.

A study by Kolla, Tully, & Bertsch (2023) reviewed Diffusion Tensor Imaging (DTI) findings in individuals with impulsive and reactive aggression and found that reduced fractional anisotropy (FA) in the uncinate fasciculus, a white matter tract connecting the amygdala and prefrontal cortex, was a significant predictor of aggressive outbursts. This supports the hypothesis

that impaired connectivity between emotion-regulating and executive-functioning brain regions contributes to aggression.

White matter disruptions in key neural circuits involved in aggression have been consistently observed in individuals with conduct disorder (CD), intermittent explosive disorder (IED), and antisocial personality disorder (ASPD). Research has shown that reduced FA in the corpus callosum, particularly in the anterior and midbody regions, is associated with higher levels of impulsive aggression. The corpus callosum plays a vital role in inter-hemispheric communication, and deficits in this structure may impair emotional regulation, contributing to uncontrolled aggressive behaviours (Lee et al. 2016).

Additionally, abnormalities in the cingulum bundle, which connects the prefrontal cortex with the limbic system, have been implicated in aggression (Bubb, Metzler-Baddeley & Aggleton, 2018). Lower FA values in this region correlate with higher aggression scores and reduced cognitive control over emotional impulses (Waller et al., 2017). These findings align with broader research suggesting that aggression is not merely a function of increased emotional reactivity (amygdala hyperactivity) but also of diminished top-down inhibitory control exerted by the prefrontal cortex (Waller et al., 2017; Bubb, Metzler-Baddeley & Aggleton, 2018).

Abnormalities in frontal and temporal EEG activity have been linked to heightened aggression, particularly in individuals with traumatic brain injury (TBI), schizophrenia, and mood disorders (Murphy et al., 2024). These irregularities suggest disruptions in the neural circuits responsible for impulse control and emotional regulation, which may contribute to an increased risk of violent behaviour.

Research has demonstrated that increased beta wave activity in the orbitofrontal cortex is associated with a higher likelihood of aggression. Beta waves are typically involved in cognitive control and decision-making, but excessive activity in aggressive individuals may indicate an overactive but inefficient inhibitory system, leading to failed impulse regulation. Additionally, studies have shown that individuals who later engage in violent behaviour often exhibit reduced theta power in the anterior cingulate cortex (ACC), a region crucial for conflict monitoring and emotional self-regulation (Reyes & Amador, 2009).

In addition to static EEG biomarkers, recent advancements have integrated machine learning algorithms with EEG data to develop real-time aggression prediction models. These models analyse shifts in brainwave activity, detecting increases in high-frequency oscillations (beta and gamma waves) alongside reductions in low-frequency oscillations (theta and delta waves), which often signal impending aggression (Parsaei et al., 2024). With high predictive accuracy, these systems offer the potential for early intervention strategies, allowing clinicians to identify and address aggression risk before escalation.

3.3. AI Use in Aggression Prediction and Management

A key application of AI in aggression management is the real-time assessment of violence risk in psychiatric inpatients (Tay, Li & Sim, 2022). AI-driven models, such as machine learning-based risk prediction algorithms, have been implemented in psychiatric hospitals to flag patients with an increased likelihood of violent behaviour based on historical data, behavioural cues, and neuroimaging patterns. For instance, a study conducted by Parmigiani et al. (2022) demonstrated that deep learning models trained on functional MRI (fMRI) and speech patterns were able to predict aggressive incidents in psychiatric patients with 87% accuracy. These models were subsequently integrated into hospital EHRs, where alerts were generated for clinicians, allowing for early intervention and de-escalation strategies.

In clinical settings, speech and movement analysis using AI has proven effective for early aggression detection. For instance, a study by Badawi et al. (2024) introduced a multimodal system integrating wearable sensors and video analysis to predict agitation in patients with dementia. The system detected pre-agitation patterns at least six minutes before an aggressive event, allowing healthcare professionals to intervene promptly.

A systematic review by Tay, Li, & Sim (2022) evaluated the effectiveness of AI methods in predicting aggression risk among psychiatric inpatients. The review analysed eight studies that employed machine learning algorithms to assess factors associated with aggression. The findings indicated that these AI models achieved acceptable to excellent predictive accuracy, with area under the curve (AUC) values ranging from 0.75 to 0.87. Notably, no single machine learning algorithm consistently outperformed others across the studies. The review also identified relevant factors associated with aggression, including demographic and social profiles, past aggression, forensic history, psychiatric history, psychopathology, challenging behaviours, and management domains (Tay, Li, & Sim, 2022).

The study conducted by Lamsma et al. (2022) assessed the predictive validity and calibration of the Oxford Mental Illness and Violence (OxMIV) tool for identifying the risk of interpersonal violence in patients with schizophrenia spectrum disorders. Conducted in a Dutch psychiatric cohort of 637 individuals, the study sought to externally validate and recalibrate OxMIV for a three-year violence prediction window. Initial analysis demonstrated moderate discrimination, with an area under the receiver operating characteristic curve (AUC) of 0.67 (95% CI: 0.61–0.73). However, the original model systematically underestimated individual risk probabilities, necessitating recalibration. Adjusting the intercept and slope significantly improved the model's calibration while maintaining an optimal balance between sensitivity (47%) and specificity (73%) at a 10% probability threshold (Lamsma et al., 2022). These findings suggest that a recalibrated OxMIV model enhances predictive accuracy and can serve as a valuable adjunct in clinical risk stratification for psychiatric patients. The results support the implementation of AI-driven risk stratification tools in psychiatric hospitals to enhance aggression prediction and facilitate early intervention strategies. Unlike conventional structured risk assessment tools, which are resource-intensive and require specialised training, OxMIV is a scalable, clinically accessible tool relying on routinely collected patient data.

One study by Botelle et al. (2022) investigated the application of transformer-based natural language processing (NLP) models for the extraction and classification of interpersonal violence-related information from electronic health records (EHRs) in a psychiatric healthcare setting. Using de-identified EHRs from the South London and Maudsley NHS Foundation Trust, the researchers developed an annotated dataset to categorise textual references to violence based on patient status (perpetrator, victim, or witness) and type of violence (domestic, physical, or sexual). The study employed the BioBERT model, fine-tuned on these annotated datasets, and evaluated its predictive performance using a 10-fold cross-validation approach. The model demonstrated high classification accuracy, with precision scores ranging from 89% (perpetrator classification) to 98% (sexual violence identification), and recall rates between 89% and 97% (Botelle et al. 2022). These findings confirm that state-of-the-art NLP methodologies can efficiently process large volumes of unstructured clinical text with high reliability, facilitating the automatic extraction of critical data on violence exposure in psychiatric populations. By automating the detection of violence-related information from clinical notes, these NLP approaches reduce the workload on healthcare professionals while improving the precision and efficiency of risk assessments. Integrating such AI-based methodologies into routine psychiatric care can support proactive intervention strategies, inform clinical decision-making, and improve patient safety.

A recent study by Cheng et al. (2023) explored the efficacy of multiple machine learning algorithms in predicting aggressive behaviours among hospitalised patients diagnosed with schizophrenia. Utilising data from 2,037 patients, the researchers compared four different algorithms: Multi-layer Perceptron, Lasso regression, Support Vector Machine, and Random Forest. The Random Forest algorithm notably demonstrated superior predictive performance, achieving an impressive area under the receiver operating characteristic curve (AUC) of 0.955 (95% CI: 0.935–0.970), significantly outperforming the other algorithms ($p < 0.0001$). The analysis identified several critical predictors associated with aggressive incidents, including reduced family support, diminished patient insight into their illness and treatment, longer duration of illness, previous

history of aggressive behaviour, lower social support, poor medication adherence, younger patient age, and higher perceived family burden (Cheng et al. 2023). The findings underscore the clinical utility of machine learning, particularly Random Forest algorithms, as powerful tools in enhancing risk assessment accuracy.

A recent study by Watts et al. (2023) employed multiple machine learning algorithms to longitudinally predict incidents of physical aggression among forensic psychiatric patients diagnosed with schizophrenia. Using evidence-based risk factors, protective variables, and clinical treatment indicators collected at baseline, the researchers developed predictive models evaluated at 4-, 12-, and 18-month follow-ups. Among several machine learning methods tested—including Logistic Regression, Elastic Net, Extreme Gradient Boosting (XGBoost), and Random Forest—the Random Forest algorithm consistently showed the highest predictive accuracy, achieving a balanced accuracy of 86.60% and an area under the receiver operating characteristic curve (AUC) of 0.914 at the 4-month follow-up schizophrenia. Key predictors identified were impulse control, treatment adherence, worsening psychotic and mood symptoms, substance abuse, poor family and social support, negative peer influence, and stress management issues. The models demonstrated significant improvement over existing traditional risk assessment tools (e.g., VRAG and HCR-20) and exhibited strong negative predictive values, effectively identifying patients unlikely to exhibit future aggression schizophrenia (Watts et al, 2023).

The advancement of bioinformatics and machine learning (ML) techniques has enabled the processing of large datasets from genetic and omics analyses, facilitating the identification of complex patterns with clinical relevance. Recent studies have demonstrated the utility of ML in psychiatry, such as Cao et al. (2020) who used ML with fMRI data to distinguish first-episode drug-naïve schizophrenia patients and predict their response to risperidone. Similarly, other ML models integrating imaging and clinical data have been able to predict social and functional impairments in individuals at high risk for psychosis or with recent-onset depression. Additionally, deep learning techniques incorporating genetic and clinical variables have been used to predict antidepressant treatment response in major depressive disorder. A recent study by Suchting et al. (2018) used ML models incorporating demographic, psychometric, and genetic data, including polymorphisms in the FKBP5 gene, to identify predictors of aggression. Their findings suggest that combining clinical and biological markers can effectively distinguish aggressive individuals and assess their likelihood of developing aggressive behaviour. These innovations in ML not only hold promise for enhancing clinical practices but also extend to other sectors, including public health education, where similar approaches are being leveraged to address mental health challenges.

AI-powered educational platforms have emerged as innovative tools to increase public awareness and effectively address stigma surrounding psychiatric disorders (Thakkar, Gupta, & De Sousa, 2024). Historically, educational campaigns on psychiatric conditions have employed mass media, informational pamphlets, and public seminars with varying success. Campaigns combining factual education with personal narratives or lived-experience stories have proven particularly effective in reducing stigma by fostering empathy, improving mental health literacy, and increasing willingness to seek help (Semrau et al., 2023). Nonetheless, traditional methods have limitations, such as limited personalisation, scalability, and difficulties measuring long-term effectiveness.

Integrating AI-driven storytelling into educational platforms can significantly enhance effectiveness by providing personalised content tailored to diverse audiences, thus increasing relatability and engagement. By leveraging AI to create immersive, interactive experiences—such as virtual reality simulations or personalised narrative journeys—users can gain deeper emotional insights into the lived experiences of individuals with psychiatric disorders, thereby reducing misconceptions and stigma. Additionally, AI-powered media representation correction systems can analyse portrayals of psychiatric disorders across media channels, identify inaccuracies or harmful stereotypes, and generate immediate evidence-based corrections (Thakkar, Gupta & De Sousa, 2024). Similarly, social media monitoring powered by AI can proactively detect and respond to misinformation or stigmatising content, enabling targeted educational interventions in real-time.

Adopting a proactive stance in combating stigma involves not only correcting misconceptions but actively reshaping public perception through continuous, adaptive education and community engagement. AI platforms can dynamically adapt educational materials based on user feedback, ensuring sustained effectiveness. Moreover, by utilising AI-driven analytics to identify regions or communities most affected by stigma, targeted interventions can be deployed, maximising resources and impact. These comprehensive approaches, leveraging AI's capabilities, can contribute substantially to creating a more informed, empathetic, and stigma-free society.

3.4. Ethical Considerations in Psychiatric Aggression Management

The management of aggression in psychiatric settings presents a unique ethical dilemma, particularly with the increasing integration of artificial intelligence (AI) and predictive analytics in clinical decision-making (see Figure 3). While AI-driven models offer improved accuracy in predicting violent behaviour, their use raises significant concerns regarding autonomy, bias, consent, and the potential for unjustified restrictions on patient rights (Parmigiani et al., 2022).

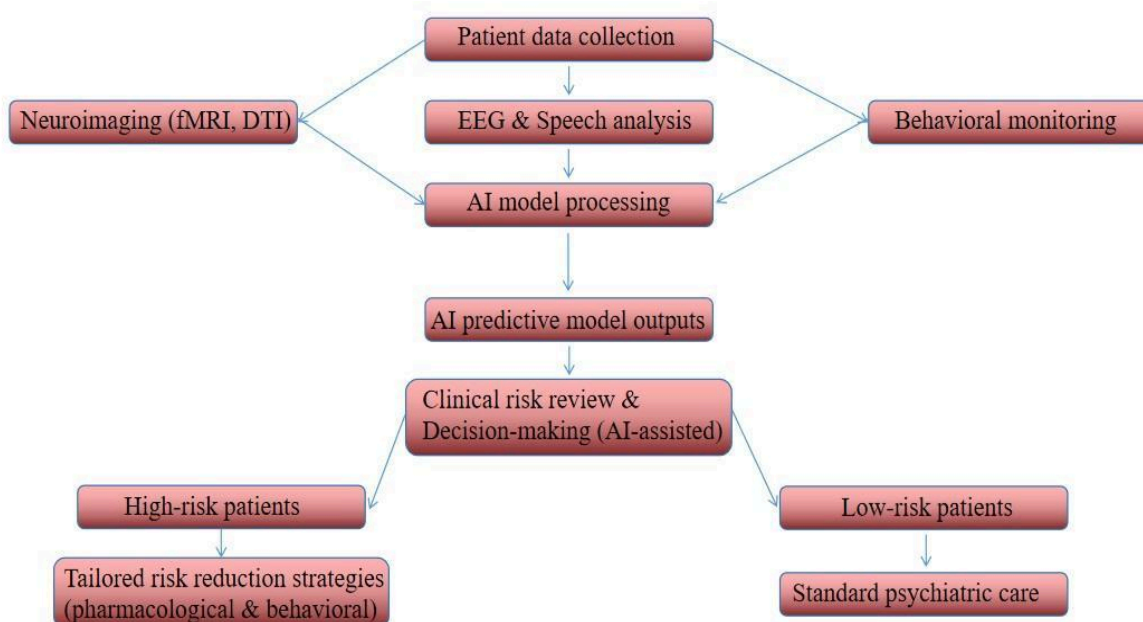


Figure 3. Flowchart illustrating the AI-driven aggression prediction process.

Patient data is collected from neuroimaging (fMRI, DTI), EEG & speech analysis, and behavioural monitoring. The AI model processes these inputs to generate predictive risk outputs, which are reviewed by clinicians. Based on AI-assisted risk assessment, patients are classified as high-risk or low-risk. High-risk patients undergo tailored risk reduction strategies, while low-risk patients receive standard psychiatric care, ensuring a balanced, AI-supported clinical decision-making framework.

3.4.1. The Ethical Dilemma of Predictive AI in Psychiatry

The ability of AI-driven models to predict aggression before it occurs represents both an opportunity and an ethical challenge. In forensic psychiatry, risk assessment tools have traditionally relied on clinical judgement, structured assessment scales, and patient history. The introduction of machine learning algorithms trained on fMRI, speech patterns, and behavioural markers has significantly improved predictive accuracy, with some models achieving up to 87% accuracy in identifying aggression risk (Parsaei et al., 2024; Yousefi et al., 2024). However, preemptive interventions based on AI-driven forecasts raise critical ethical concerns, including the potential for false positives, unnecessary interventions, and violations of patient rights.

Recent efforts to regulate AI in psychiatric settings have resulted in the development of legal frameworks aimed at ensuring that AI-driven psychiatric assessments remain ethical and compliant

with human rights laws (Tavory, 2024). The European Union AI Act from 2024 has categorised AI applications in healthcare as "high-risk," requiring full transparency in decision-making, mandatory bias audits before deployment, and human oversight in clinical decisions (Goktas, & Grzybowski, 2025). Similarly, the World Health Organisation's AI Ethics Guidelines for Mental Health emphasise that AI systems should enhance rather than replace human clinical judgement, and that risk assessment models must prioritise patient rights over predictive accuracy (Mennella et al., 2024). The United States has also introduced regulatory measures, recommending that psychiatric hospitals using AI-based aggression prediction implement explainable AI models to improve clinician trust and interpretability, while requiring explicit patient-informed consent before integrating real-time biometric monitoring (Abgrall et al., 2024).

One of the most pressing concerns is the risk of misclassification. AI models, despite their statistical power, are not infallible; errors in classification could result in patients being unjustly labelled as violent based on predictive probabilities rather than actual behaviour (Cheng et al., 2023; Yousefi et al., 2024). A false positive could lead to unwarranted coercive measures, such as forced medication, isolation, or even involuntary hospitalisation, violating the principle of non-maleficence (do no harm). On the other hand, false negatives- where an AI model fails to detect an imminent act of aggression- could expose clinicians, caregivers, and other patients to harm, raising issues related to beneficence (the duty to prevent harm).

In forensic psychiatry, AI-based risk prediction tools have been shown to overestimate aggression risk in patients from lower socioeconomic backgrounds, suggesting that environmental stressors were misclassified as intrinsic aggression risk factors (Ogonah et al., 2023).

Ethicists argue that risk prediction should not replace human judgement but rather serve as an augmentation tool for clinicians (Maleki Varnosfaderani & Forouzanfar, 2024). AI should act as an assistive rather than authoritative system, ensuring that predictions are used within the context of multidisciplinary decision-making rather than as sole determinants of patient management strategies.

3.4.2. Patient Autonomy vs. Public Safety

A fundamental ethical challenge in psychiatric aggression management is balancing patient autonomy with the need for public and institutional safety. Patients experiencing acute psychosis, mania, or severe personality disorders may exhibit impaired judgement and decision-making capacity, raising ethical dilemmas regarding involuntary treatment (Chen et al., 2023). AI-driven models exacerbate this issue by introducing risk-based restrictions that may limit patient autonomy before aggression even manifests.

The principle of autonomy dictates that psychiatric patients should have the right to make decisions about their own treatment, including refusing medication or hospitalisation (Neilson & Chaimowitz, 2015). However, when a patient is classified as high-risk for aggression, clinicians may feel compelled to intervene preventively, potentially overriding patient autonomy. The ethical justification for such interventions typically rests on the principle of proportionality, which holds that any limitation of autonomy should be the least restrictive necessary to ensure safety (Byskov, 2019).

A review of psychiatric aggression policies in the EU and the US found that preemptive psychiatric interventions based on AI risk assessments increased the use of coercive measures by 14% in hospitals that implemented algorithmic decision-support systems (Chieze et al., 2021). This raises concerns that predictive AI may lead to an over-reliance on restrictive measures, reinforcing rather than mitigating coercive psychiatric practices.

To address this ethical challenge, some institutions advocate for shared decision-making models, in which patients are informed about AI-based risk predictions and included in discussions about their treatment (Mennella et al., 2024). This participatory approach respects patient autonomy while ensuring that psychiatric interventions remain ethical and proportionate.

3.4.3. The Ethical Risks of Algorithmic Bias in Psychiatric Aggression Prediction

AI systems are only as unbiased as the datasets on which they are trained. Numerous studies have highlighted racial, socioeconomic, and gender biases in psychiatric risk assessments, which are exacerbated when AI models inherit these biases from training data (Cross, Choma & Onofrey, 2024). For example, an analysis of machine learning-based psychiatric aggression prediction models found that Black and Hispanic patients were disproportionately flagged as “high-risk” for aggression, despite no significant differences in actual violent incidents (Kuzmanov, 2024; Barak-Corren et al, 2016).

This raises serious ethical concerns regarding fairness and justice, as biased AI models may lead to unwarranted restrictive interventions for marginalised populations (Cross, Choma & Onofrey, 2024; Kuzmanov, 2024). If an AI system disproportionately classifies certain groups as aggressive, it reinforces discriminatory practices in psychiatric care, violating the ethical principle of justice (the fair distribution of healthcare resources and treatment).

To mitigate algorithmic bias in psychiatric aggression management, experts emphasise the necessity of developing AI models that are trained on diverse and representative datasets, ensuring that predictive systems account for multi-ethnic, socioeconomically diverse patient populations rather than reinforcing racial or class-based disparities (Nazer et al., 2023). The underrepresentation of certain demographic groups in training data has been identified as a major factor contributing to bias, as AI models tend to reflect the statistical patterns of their datasets rather than genuine behavioural tendencies (Nazer et al., 2023; Cross, Choma & Onofrey, 2024; Mennella et al., 2024). Addressing this issue requires systematic efforts to include data from a wide range of psychiatric patients, ensuring that risk assessments are equitable and clinically relevant across different populations.

One of the primary challenges in AI-driven risk assessment is the interpretability of predictive models, which requires mental health professionals to be proficient in both algorithmic reasoning and clinical judgement (Lee et al., 2021). A lack of AI literacy among clinicians has been identified as a significant barrier to the responsible use of predictive analytics, with studies indicating that many practitioners lack the necessary training to critically evaluate algorithmic outputs and differentiate between clinically meaningful predictions and statistical artifacts (Khan et al., 2023). To address this, specialised AI-in-psychiatry curricula should incorporate hands-on training in machine learning interpretability, real-world case studies on algorithmic misclassification, and interactive modules on risk stratification to ensure that clinicians understand the probabilistic nature of AI predictions and their ethical implications (Lee et al., 2021).

Additionally, decision-support training programs can help mental health professionals integrate AI-generated insights into their clinical workflow without over-relying on automated risk assessments (Elhaddad & Hamam, 2024). Research has shown that multimodal educational interventions, such as virtual patient simulations, AI-assisted case-based learning, and neuroimaging interpretation workshops, significantly enhance clinicians' ability to apply AI tools responsibly (Maleki Varnosfaderani & Forouzanfar, 2024). Expanding education beyond clinical applications, neuroethics and AI governance courses should be incorporated into medical and psychology curricula to equip future practitioners with a deep understanding of algorithmic bias, patient autonomy, and the impact of AI-driven classifications on psychiatric treatment pathways (Ueda et al., 2023). Such educational initiatives are critical in bridging the gap between AI innovation and ethical, patient-centered psychiatric care, ensuring that emerging technologies enhance clinical decision-making without reinforcing coercive or discriminatory practices.

Furthermore, the implementation of transparent AI decision-making processes is crucial to ensuring that psychiatric risk assessments remain interpretable and ethically justifiable. The use of explainable AI (XAI) models allows clinicians to understand, interpret, and critically evaluate AI-generated predictions, preventing cases where restrictive psychiatric interventions are carried out based solely on opaque, black-box algorithmic outputs (Abgrall et al., 2024). A comprehensive

survey by Markus, Kors & Rijnbeek, (2021) highlights that the opacity of complex AI models can hinder their adoption in healthcare, as clinicians may be reluctant to rely on systems they do not understand. The study emphasises that explainable AI methods can bridge this gap by making AI decision-making processes more interpretable, thereby enhancing trust and facilitating integration into clinical workflows.

Transparency in AI-generated risk assessments fosters clinical accountability and ensures that predictive models can be audited for bias, adjusted when inaccuracies arise, and tailored to individual patient contexts rather than applied as rigid, one-size-fits-all solutions. Additionally, it is essential that risk assessment remains a collaborative effort between human clinicians and AI systems, with psychiatric professionals maintaining decision-making authority over patient treatment. AI should serve as an assistive tool rather than a replacement for clinical expertise, meaning that no patient should be classified solely based on an algorithmic determination. By integrating AI-driven insights with human judgement, clinicians can ensure that ethical considerations, contextual factors, and patient-specific nuances are fully accounted for in aggression risk assessments, ultimately fostering a more fair, reliable, and patient-centered approach to AI-enhanced psychiatric care.

To address these ethical challenges, it is essential to ensure fairness in algorithm design, enhance transparency in AI decision-making processes, and adopt patient-centered approaches to consent and data privacy. Improving transparency in clinical AI applications requires concrete methodological steps, including the adoption of explainable AI techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) (Ladbury et al., 2022). These techniques allow clinicians to interpret individual AI-based predictions by highlighting features or data points that contributed significantly to specific decisions. Moreover, developing user-friendly interfaces that visually present model reasoning can further facilitate clinician understanding and enhance trust in AI systems.

3.4.4. Informed Consent and AI Transparency in Psychiatry

Another major ethical concern in AI-based aggression prediction is the issue of informed consent (Farhud & Zokaei, 2021). Patients in psychiatric facilities may lack the capacity to fully understand AI-based risk assessments and how such predictions impact their treatment. The use of real-time fMRI surveillance, biometric monitoring, and AI-driven behavioural tracking further complicates the issue, as patients may unknowingly be subjected to continuous risk profiling (Farhud & Zokaei, 2021; Nazer et al., 2023).

It is mentioned in the literature that only 32% of psychiatric inpatients were aware that AI models were being used to assess their risk of violence, suggesting a lack of transparency in AI deployment (Darby et al., 2023). In some psychiatric hospitals, real-time biometric monitoring, including continuous EEG, speech analysis, and facial expression tracking, is implemented without explicit patient consent, raising further ethical concerns regarding surveillance and autonomy (Appenzeller, Appelbaum & Trachsel, 2020).

Ethicists argue that psychiatric patients should be explicitly informed when AI models are used in their care, and their consent should be obtained whenever possible (Tavory, 2024). Institutions must implement clear, patient-friendly communication strategies that explain how AI models are used in aggression risk assessment, what data is collected and analysed, and how AI predictions influence clinical decisions.

3.4.5. Ethical Frameworks for AI in Psychiatric Aggression Management

Given the complex ethical challenges associated with AI-driven psychiatric aggression prediction, several ethical frameworks have been proposed to guide the responsible and fair implementation of AI in mental health care. One of the most foundational frameworks is The Four Principles of Biomedical Ethics, established by Beauchamp and Childress, which asserts that AI applications must adhere to the core principles of autonomy, beneficence, non-maleficence, and

justice. This means that AI-driven psychiatric interventions should respect patient self-determination, actively promote well-being, avoid causing harm, and ensure fair and equal treatment for all patients, regardless of demographic background or psychiatric history. Additionally, as AI becomes increasingly integrated into medical decision-making, regulatory bodies have recognised the need for legal oversight to prevent unethical practices.

The EU AI Act approved in 2024 outlines specific requirements for AI models used in healthcare, mandating that they meet transparency, accountability, and human oversight standards. This legislation ensures that AI-based risk assessments in psychiatry remain interpretable, free from systemic bias, and subject to clinician verification, preventing the misuse of opaque, unchallengeable AI-generated classifications in aggression management. At a global level, the World Health Organisation (WHO) has introduced the Ethics and Governance Framework for AI in Mental Health (2024), which emphasises the need for multidisciplinary oversight in AI deployment (Mennella et al., 2024).

This framework advocates for collaboration between ethicists, AI developers, clinicians, and patient representatives, ensuring that psychiatric AI models align with ethical, clinical, and human rights considerations before being widely implemented. Collectively, these ethical guidelines seek to balance technological advancements with moral responsibility, ensuring that AI in psychiatric aggression management remains a tool for support rather than a mechanism for control, coercion, or discrimination.

To address all challenges, researchers and policymakers have proposed evidence-based techniques to mitigate bias in aggression prediction models. One approach is algorithmic fairness testing, which ensures that AI models undergo bias audits using diverse multi-ethnic and multi-socioeconomic datasets before deployment. Reweighting training data to ensure balanced demographic representation has also been identified as a strategy for reducing bias (Chen et al., 2023; Ueda et al., 2023). Explainable AI has emerged as a key tool for improving transparency, allowing psychiatrists to interpret AI-based risk assessments rather than blindly trusting opaque black-box systems (Muhammad & Bendechea, 2024). Hospitals implementing explainable AI in forensic psychiatry have reported increased clinician trust in AI-generated aggression risk scores (Rosenbacke et al., 2024). Human-in-the-loop AI decision-making, in which psychiatric professionals maintain final decision-making authority over patient treatment, has been shown to significantly reduce aggression-related misclassifications (Tay, Li, & Sim, 2022). Several studies demonstrated that a hybrid decision-making approach combining AI risk predictions with clinician oversight resulted in a significant reduction in misclassifications (Maleki Varnosfaderani & Forouzanfar, 2024).

Another practical method involves implementing standardised reporting frameworks, such as the Minimum Information for Medical AI Reporting (MINIMAR), which mandates disclosure of essential details on model development, validation, data provenance, and performance across diverse populations (Subías-Beltrán et al, 2024). Routine auditing of AI algorithms by independent oversight committees, incorporating fairness and bias detection protocols, can also systematically enhance transparency (Ueda et al., 2023). Continuous ethical scrutiny and collaboration among AI developers, clinicians, and ethicists are crucial to achieving responsible and inclusive AI implementation in healthcare, thereby enhancing both the clinical value of AI and the trustworthiness of AI systems among patients and healthcare professionals.

Additionally, performing systematic bias audits before deployment is crucial. Techniques like fairness metrics evaluation (e.g., demographic parity, equalised odds, and predictive equality) should be utilised to quantitatively assess model performance across diverse patient subgroups. Conducting sensitivity analyses or subgroup validations on stratified datasets can further identify and mitigate disparities (Chen et al., 2023). Regular re-auditing after deployment ensures continued monitoring and maintenance of fairness standards in AI systems.

4. Conclusion

Managing aggression in psychiatric patients is a complex challenge that demands a balance between patient rights, therapeutic efficacy, and public safety. This paper has highlighted the neurobiological underpinnings of aggression, emphasising dysfunctions in the prefrontal cortex, amygdala hyperactivity, neurotransmitter imbalances, and neuroinflammation. These findings affirm that aggression is not merely a behavioural issue but a symptom of underlying neurobiological disruptions requiring evidence-based, ethical interventions.

Current management strategies often rely on restrictive measures like seclusion and forced medication, which, while ensuring short-term safety, risk exacerbating symptoms and eroding trust in psychiatric care. The integration of predictive AI models and neuroimaging offers a promising shift toward early detection and individualised interventions, reducing reliance on coercive measures. However, these advancements introduce ethical concerns, including algorithmic bias, misclassification, and threats to patient autonomy. AI should enhance, not replace, clinical judgement, requiring transparent, explainable, and ethically regulated implementation.

Beyond clinical settings, public perception of psychiatric aggression remains shaped by stigma and misinformation. Contrary to media portrayals, research shows that individuals with psychiatric disorders are more often victims than perpetrators of violence. Addressing these misconceptions through public education and policy reform is crucial in fostering a more compassionate and supportive environment.

Future research on clinical AI integration should prioritise randomised controlled trials comparing AI-supported diagnostic tools with conventional methods, rigorously evaluating key performance metrics such as sensitivity, specificity, predictive values, and clinical utility. Special emphasis should be placed on assessing model fairness across diverse demographic groups to mitigate algorithmic biases and promote equitable healthcare outcomes. Additionally, pilot studies should focus on testing interpretability methods, examining clinician decision-making accuracy, and workflow integration when using explainable versus opaque AI systems. Complementary qualitative research involving clinicians and patients can provide valuable insights into practical, ethical, and trust-related factors influencing AI adoption.

Ultimately, a shift toward proactive, patient-centered approaches- combining neuroscience, AI, and ethical considerations- can redefine psychiatric aggression management. By prioritising early intervention over containment, fostering public awareness, and maintaining a strong commitment to patient rights, psychiatric care can evolve into a more humane, effective, and ethically sound practice.

References

- Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K., & Monnet, X. (2024). Should AI models be explainable to clinicians? *Critical Care*, 28(1), Article 301. <https://doi.org/10.1186/s13054-024-05005-y>
- Ahad, A. A., Sanchez-Gonzalez, M., & Junquera, P. (2023). Understanding and addressing mental health stigma across cultures for improving psychiatric care: A narrative review. *Cureus*, 15(5), e39549. <https://doi.org/10.7759/cureus.39549>
- Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1), 689. <https://doi.org/10.1186/s12909-023-04698-z>
- Appenzeller, Y. E., Appelbaum, P. S., & Trachsel, M. (2020). Ethical and practical issues in video surveillance of psychiatric units. *Psychiatric Services*, 71(5), 480–486. <https://doi.org/10.1176/appi.ps.201900397>

- Badawi, A., Elmoghazy, S., Choudhury, S., Elgazzar, S., Elgazzar, K., & Burhan, A. (2024). A novel multimodal system to predict agitation in people with dementia within clinical settings: A proof of concept. *arXiv preprint*, arXiv:2411.08882. <https://doi.org/10.48550/arXiv.2411.08882>
- Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2016). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2). <https://doi.org/10.1176/appi.ajp.2016.16010077>
- Blair, R. J. R., Bashford-Largo, J., Dominguez, A. J., Hatch, M., Dobbartin, M., Blair, K. S., & Bajaj, S. (2025). Using machine learning to determine a functional classifier of retaliation and its association with aggression. *JAACAP Open*, 3(1), 137–146. <https://doi.org/10.1016/j.jaacop.2024.04.007>
- Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., Velupillai, S., & Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: An applied evaluative study. *BMJ Open*, 12(2), Article e052911. <https://doi.org/10.1136/bmjopen-2021-052911>
- Bubb, E. J., Metzler-Baddeley, C., & Aggleton, J. P. (2018). The cingulum bundle: Anatomy, function, and dysfunction. *Neuroscience & Biobehavioral Reviews*, 92, 104–127. <https://doi.org/10.1016/j.neubiorev.2018.05.008>
- Byskov, M. F. (2019). Qualitative and quantitative interpretations of the least restrictive means. *Bioethics*, 33(4), 511–521. <https://doi.org/10.1111/bioe.12548>
- Cao, B., Cho, R. Y., Chen, D., Xiu, M., Wang, L., Soares, J. C., & Zhang, X. Y. (2020). Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity. *Molecular Psychiatry*, 25(4), 906–913. <https://doi.org/10.1038/s41380-018-0106-5>
- Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6), 719–742. <https://doi.org/10.1038/s41551-023-01056-8>
- Cheng, N., Guo, M., Yan, F., Guo, Z., Meng, J., Ning, K., Zhang, Y., Duan, Z., Han, Y., & Wang, C. (2023). Application of machine learning in predicting aggressive behaviors from hospitalized patients with schizophrenia. *Frontiers in Psychiatry*, 14, Article 1016586. <https://doi.org/10.3389/fpsyt.2023.1016586>
- Chieze, M., Clavien, C., Kaiser, S., & Hurst, S. (2021). Coercive measures in psychiatry: A review of ethical arguments. *Frontiers in Psychiatry*, 12, Article 790886. <https://doi.org/10.3389/fpsyt.2021.790886>
- Coccaro, E. F., Keedy, S., Lee, R., & Phan, K. L. (2021). Neuronal responses to adverse social threat in healthy human subjects. *Journal of Psychiatric Research*, 136, 47–53. <https://doi.org/10.1016/j.jpsychires.2021.01.015>
- Coccaro, E. F., Lee, R. J., & Kavoussi, R. J. (2009). A double-blind, randomized, placebo-controlled trial of fluoxetine in patients with intermittent explosive disorder. *Journal of Clinical Psychiatry*, 70(5), 653–662. <https://doi.org/10.4088/JCP.08m04150>
- Coccaro, E. F., Sripada, C. S., Yanowitch, R. N., & Phan, K. L. (2011). Corticolimbic function in impulsive aggressive behavior. *Biological Psychiatry*, 69(12), 1153–1159. <https://doi.org/10.1016/j.biopsych.2011.02.032>
- Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, 3(11), Article e0000651. <https://doi.org/10.1371/journal.pdig.0000651>
- da Cunha-Bang, S., & Knudsen, G. M. (2021). The modulatory role of serotonin on human impulsive aggression. *Biological Psychiatry*, 90(7), 447–457. <https://doi.org/10.1016/j.biopsych.2021.05.016>

- Darby, W. C., MacIntyre, M., Cockerill, R. G., Stephens, D. B., Weinstock, R., & Darby, R. R. (2023). In the courts: Ethical and legal implications of emerging neuroscience technologies used for forensic purposes. In L. W. Roberts (Ed.), *Ethics and clinical neuroinnovation* (Chapter 10). Springer, Cham. https://doi.org/10.1007/978-3-031-14339-7_10
- Duke, A. A., Bègue, L., Bell, R., & Eisenlohr-Moul, T. (2013). Revisiting the serotonin-aggression relation in humans: A meta-analysis. *Psychological Bulletin*, 139(5), 1148–1172. <https://doi.org/10.1037/a0031544>
- Edward, K. L., Ousey, K., Warelou, P., & Lui, S. (2014). Nursing and aggression in the workplace: A systematic review. *British Journal of Nursing*, 23(12), 653–654, 656–659. <https://doi.org/10.12968/bjon.2014.23.12.653>
- Einberger, C., Puckett, A., Ricci, L., & Melloni, R. Jr. (2020). Contemporary pharmacotherapeutics and the management of aggressive behavior in an adolescent animal model of maladaptive aggression. *Clinical Psychopharmacology and Neuroscience*, 18(2), 188–202. <https://doi.org/10.9758/cpn.2020.18.2.188>
- Elendu, C., Amaechi, D. C., Okatta, A. U., Amaechi, E. C., Elendu, T. C., Ezech, C. P., & Elendu, I. D. (2024). The impact of simulation-based training in medical education: A review. *Medicine (Baltimore)*, 103(27), e38813. <https://doi.org/10.1097/MD.00000000000038813>
- Elhaddad, M., & Hamam, S. (2024). AI-driven clinical decision support systems: An ongoing pursuit of potential. *Cureus*, 16(4), e57728. <https://doi.org/10.7759/cureus.57728>
- Elnakib, A., Khalifa, F., Soliman, A., Shalaby, A., & Elhosseini, M. (2024). Editorial: Emerging artificial intelligence technologies for neurological and neuropsychiatric research. *Frontiers in Neuroscience*, 18, 1518442. <https://doi.org/10.3389/fnins.2024.1518442>
- Fanning, J. R., Keedy, S., Berman, M. E., Lee, R., & Coccaro, E. F. (2017). Neural correlates of aggressive behavior in real-time: A review of fMRI studies of laboratory reactive aggression. *Current Behavioral Neuroscience Reports*, 4(2), 138–150. <https://doi.org/10.1007/s40473-017-0115-8>
- Farhud, D. D., & Zokaei, S. (2021). Ethical issues of artificial intelligence in medicine and healthcare. *Iranian Journal of Public Health*, 50(11), i–v. <https://doi.org/10.18502/ijph.v50i11.7600>
- Fazel, S., Wolf, A., Larsson, H., Lichtenstein, P., Mallett, S., & Fanshawe, T. R. (2017). Identification of low risk of violent crime in severe mental illness with a clinical prediction tool (Oxford Mental Illness and Violence tool [OxMIV]): A derivation and validation study. *The Lancet Psychiatry*, 4(6), 461–468. [https://doi.org/10.1016/S2215-0366\(17\)30109-8](https://doi.org/10.1016/S2215-0366(17)30109-8)
- Fullam, R. S., & Dolan, M. C. (2008). Executive function and in-patient violence in forensic patients with schizophrenia. *The British Journal of Psychiatry*, 193(3), 247–253. <https://doi.org/10.1192/bjp.bp.107.040345>
- Goktas, P., & Grzybowski, A. (2025). Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI. *Journal of Clinical Medicine*, 14(5), 1605. <https://doi.org/10.3390/jcm14051605>
- Hoptman, M. J. (2015). Impulsivity and aggression in schizophrenia: A neural circuitry perspective with implications for treatment. *CNS Spectrums*, 20(3), 280–286. <https://doi.org/10.1017/S1092852915000206>
- Ji, S., Zhang, H., Zhou, C., Liu, X., Liu, C., & Yu, H. (2025). Resting-state voxel-wise dynamic effective connectivity predicts risky decision-making in patients with bipolar disorder type I. *Neuroscience*, 564, 135–143. <https://doi.org/10.1016/j.neuroscience.2024.11.024>
- Jolles, J., & Jolles, D. D. (2021). On neuroeducation: Why and how to improve neuroscientific literacy in educational professionals. *Frontiers in Psychology*, 12, 752151. <https://doi.org/10.3389/fpsyg.2021.752151>
- Khan, B., Fatima, H., Qureshi, A., Kumar, S., Hanan, A., Hussain, J., & Abdullah, S. (2023). Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials and Devices*, 1–8. <https://doi.org/10.1007/s44174-023-00063-2>

- Kirino, E., Hayakawa, Y., Inami, R., Inoue, R., & Aoki, S. (2019). Simultaneous fMRI-EEG-DTI recording of MMN in patients with schizophrenia. *PLoS ONE*, 14(5), e0215023. <https://doi.org/10.1371/journal.pone.0215023>
- Kolla, N. J., Tully, J., & Bertsch, K. (2023). Neural correlates of aggression in personality disorders from the perspective of DSM-5 maladaptive traits: A systematic review. *Translational Psychiatry*, 13(1), 330. <https://doi.org/10.1038/s41398-023-02612-1>
- Kölliker-Frers, R., Udovin, L., Otero-Losada, M., Kobiec, T., Herrera, M. I., Palacios, J., Razzitte, G., & Capani, F. (2021). Neuroinflammation: An integrating overview of reactive-neuroimmune cell interactions in health and disease. *Mediators of Inflammation*, 2021, 9999146. <https://doi.org/10.1155/2021/9999146>
- Kredlow, A. M., Fenster, R. J., Laurent, E. S., Ressler, K. J., & Phelps, E. A. (2022). Prefrontal cortex, amygdala, and threat processing: Implications for PTSD. *Neuropsychopharmacology*, 47(1), 247–259. <https://doi.org/10.1038/s41386-021-01155-7>
- Kuzmanov, I. (2024). AI and psychological profiling for violence risk assessment: Enhancing accuracy and addressing ethical challenges. *Journal of Novel Research and Innovative Development (JNRID)*, 2(12)). https://www.researchgate.net/publication/388155585_AI_and_Psychological_Profiling_for_Violence_Risk_Assessment_Enhancing_Accuracy_and_Addressing_Ethical_Challenges
- Ladbury, C., Zarinshenas, R., Semwal, H., Tam, A., Vaidehi, N., Rodin, A. S., Liu, A., Glaser, S., Salgia, R., & Amini, A. (2022). Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: A narrative review. *Translational Cancer Research*, 11(10), 3853–3868. <https://doi.org/10.21037/tcr-22-1626>
- Lamsma, J., Yu, R., Fazel, S., & Genetic Risk and Outcome of Psychosis (GROUP) investigators. (2022). Validation and recalibration of OxMIV in predicting violent behaviour in patients with schizophrenia spectrum disorders. *Scientific Reports*, 12(1), Article 461. <https://doi.org/10.1038/s41598-021-04266-9>
- Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H. C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6(9), 856–864. <https://doi.org/10.1016/j.bpsc.2021.02.001>
- Lee, R., Arfanakis, K., Evia, A., Evia, J. M., Fanning, J., Keedy, S., & Coccaro, E. F. (2016). White matter integrity reductions in intermittent explosive disorder. *Neuropsychopharmacology*, 41, 2697–2703. <https://doi.org/10.1038/npp.2016.74>
- Maleki Varnosfaderani, S., & Forouzanfar, M. (2024). The role of AI in hospitals and clinics: Transforming healthcare in the 21st century. *Bioengineering*, 11(4), Article 337. <https://doi.org/10.3390/bioengineering11040337>
- Manchia, M., Comai, S., Pinna, M., Pinna, F., Fanos, V., Denovan-Wright, E., & Carpiniello, B. (2019). Biomarkers in aggression. *Advances in Clinical Chemistry*, 93, 169–237. <https://doi.org/10.1016/bs.acc.2019.07.004>
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, 10(4), e26297. <https://doi.org/10.1016/j.heliyon.2024.e26297>
- Muhammad, D., & Bendecheche, M. (2024). Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*, 24, 542–560. <https://doi.org/10.1016/j.csbj.2024.08.005>
- Murphy, M., Carrión, R. E., Rubio, J., & Malhotra, A. K. (2024). Peak alpha frequency and electroencephalographic microstates are correlated with aggression in schizophrenia.

- Journal of Psychiatric Research*, 175, 60–67.
<https://doi.org/10.1016/j.jpsychires.2024.04.051>
- Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6), Article e0000278. <https://doi.org/10.1371/journal.pdig.0000278>
- Neilson, G., & Chaimowitz, G. (2015). Informed consent to treatment in psychiatry. *Canadian Journal of Psychiatry*, 60(4), 1–11. Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4459249>
- Nikolic, M., Pezzoli, P., Jaworska, N., & Seto, M. C. (2022). Brain responses in aggression-prone individuals: A systematic review and meta-analysis of functional magnetic resonance imaging (fMRI) studies of anger- and aggression-eliciting tasks. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 119, 110596. <https://doi.org/10.1016/j.pnpbp.2022.110596>
- Ogonah, M. G. T., Seyedsalehi, A., Whiting, D., & Fazel, S. (2023). Violence risk assessment instruments in forensic psychiatric populations: A systematic review and meta-analysis. *Lancet Psychiatry*, 10(10), 780–789. [https://doi.org/10.1016/S2215-0366\(23\)00256-0](https://doi.org/10.1016/S2215-0366(23)00256-0)
- Pardini, D. A., Raine, A., Erickson, K., & Loeber, R. (2014). Lower amygdala volume in men is associated with childhood aggression, early psychopathic traits, and future violence. *Biological Psychiatry*, 75(1), 73–80. <https://doi.org/10.1016/j.biopsych.2013.04.003>
- Parmigiani, G., Barchielli, B., Casale, S., Mancini, T., & Ferracuti, S. (2022). The impact of machine learning in predicting risk of violence: A systematic review. *Frontiers in Psychiatry*, 13, Article 1015914. <https://doi.org/10.3389/fpsyt.2022.1015914>
- Parsaei, M., Arvin, A., Taebi, M., Seyedmirzaei, H., Cattarinussi, G., Sambataro, F., Pigoni, A., Brambilla, P., & Delvecchio, G. (2024). Machine learning for prediction of violent behaviors in schizophrenia spectrum disorders: A systematic review. *Frontiers in Psychiatry*, 15, Article 1384828. <https://doi.org/10.3389/fpsyt.2024.1384828>
- Passamonti, L., Crockett, M. J., Apergis-Schoute, A. M., Clark, L., Rowe, J. B., Calder, A. J., & Robbins, T. W. (2012). Effects of acute tryptophan depletion on prefrontal-amygdala connectivity while viewing facial signals of aggression. *Biological Psychiatry*, 71(1), 36–43. <https://doi.org/10.1016/j.biopsych.2011.07.033>
- Reyes, A. C., & Amador, A. A. (2009). Qualitative and quantitative EEG abnormalities in violent offenders with antisocial personality disorder. *Journal of Forensic and Legal Medicine*, 16(2), 59–63. <https://doi.org/10.1016/j.jflm.2008.08.001>
- Rosenbacke, R., Melhus, Å., McKee, M., & Stuckler, D. (2024). How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: Systematic review. *JMIR AI*, 3, Article e53207. <https://doi.org/10.2196/53207>
- Semrau, M., Gronholm, P. C., Eaton, J., Maulik, P. K., Ayele, B., Bakolis, I., Mendon, G. B., Bhattarai, K., Brohan, E., Cherian, A. V., Daniel, M., Girma, E., Gurung, D., Hailemariam, A., Hanlon, C., Healey, A., Kallakuri, S., Li, J., Loganathan, S., Ma, N., Ma, Y., Metsahel, A., Ouali, U., Yaziji, N., Zgueb, Y., Zhang, W., Zhang, X., Thornicroft, G., & Votruba, N. (2023). Reducing stigma and improving access to care for people with mental health conditions in the community: Protocol for a multi-site feasibility intervention study (Indigo-Local). *Research Square*. <https://doi.org/10.21203/rs.3.rs-3237562/v1>
- Siep, N., Tonnaer, F., van de Ven, V., Arntz, A., Raine, A., & Cima, M. (2019). Anger provocation increases limbic and decreases medial prefrontal cortex connectivity with the left amygdala in reactive aggressive violent offenders. *Brain Imaging and Behavior*, 13(5), 1311–1323. <https://doi.org/10.1007/s11682-018-9945-6>

- Singh, A. S., Borgohain, L., & Singh, B. (2023). Aggression and its correlation with plasma C-reactive protein levels in patients with schizophrenia: A hospital-based cross-sectional study. *Indian Journal of Psychiatry*, 65(6), 667–670. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_427_22
- Soyka, M. (2011). Neurobiology of aggression and violence in schizophrenia. *Schizophrenia Bulletin*, 37(5), 913–920. <https://doi.org/10.1093/schbul/sbr103>
- Subías-Beltrán, P., Pitarch, C., Migliorelli, C., Marte, L., Galofré, M., & Orte, S. (2024). The role of transparency in AI-driven technologies: Targeting healthcare. In *AI - Ethical and Legal Challenges [Working Title]*. IntechOpen. <https://doi.org/10.5772/intechopen.1007444>
- Suchting, R., Gowin, J. L., Green, C. E., Walss-Bass, C., & Lane, S. D. (2018). Genetic and psychosocial predictors of aggression: Variable selection and model building with component-wise gradient boosting. *Frontiers in Behavioral Neuroscience*, 12, Article 89. <https://doi.org/10.3389/fnbeh.2018.00089>
- Takahashi, A. (2024). Associations of the immune system in aggression traits and the role of microglia as mediators. *Neuropharmacology*, 256, 110021. <https://doi.org/10.1016/j.neuropharm.2024.110021>
- Tavory, T. (2024). Regulating AI in mental health: Ethics of care perspective. *JMIR Mental Health*, 11, Article e58493. <https://doi.org/10.2196/58493>
- Tay, J. L., Li, Z., & Sim, K. (2022). Effectiveness of artificial intelligence methods in personalized aggression risk prediction within inpatient psychiatric treatment settings: A systematic review. *Journal of Personalized Medicine*, 12(9), Article 1470. <https://doi.org/10.3390/jpm12091470>
- Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. *Frontiers in Digital Health*, 6, Article 1280235. <https://doi.org/10.3389/fdgth.2024.1280235>
- 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. (2023). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology*, 42(1), 3–15. <https://doi.org/10.1007/s11604-023-01474-3>
- Wagels, L., Habel, U., Raine, A., & Clemens, B. (2022). Neuroimaging, hormonal, and genetic biomarkers for pathological aggression—Success or failure? *Current Opinion in Behavioral Sciences*, 43, 101–110. <https://doi.org/10.1016/j.cobeha.2021.08.007>
- Waller, R., Dotterer, H. L., Murray, L., Maxwell, A. M., & Hyde, L. W. (2017). White-matter tract abnormalities and antisocial behavior: A systematic review of diffusion tensor imaging studies across development. *NeuroImage: Clinical*, 14, 201–215. <https://doi.org/10.1016/j.nicl.2017.01.014>
- Wang, Y., Becker, B., Wang, J., Wang, Y., Zhang, L., Mei, Y., Li, H., & Lei, Y. (2024). Exaggerated sensitivity to threat and reduced medial prefrontal engagement during threat generalization in reactive aggressive adolescents. *NeuroImage*, 294, 120645. <https://doi.org/10.1016/j.neuroimage.2024.120645>
- Watts, D., Mamak, M., Moulden, H., Upfold, C., de Azevedo Cardoso, T., Kapczinski, F., & Chaimowitz, G. (2023). The HARM models: Predicting longitudinal physical aggression in patients with schizophrenia at an individual level. *Journal of Psychiatric Research*, 161, 91–98. <https://doi.org/10.1016/j.jpsychires.2023.02.030>
- Weltens, I., Bak, M., Verhagen, S., Vandenberk, E., Domen, P., van Amelsvoort, T., & Drukker, M. (2021). Aggression on the psychiatric ward: Prevalence and risk factors. A systematic review of the literature. *PLoS ONE*, 16(10), e0258346. <https://doi.org/10.1371/journal.pone.0258346>

- Yang, Y., & Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: A meta-analysis. *Psychiatry Research: Neuroimaging*, 174(2), 81–88. <https://doi.org/10.1016/j.psychresns.2009.03.012>
- Yousefi, M., Akhbari, M., Mohamadi, Z., Karami, S., Dasoomi, H., Atabi, A., Sarkeshikian, S. A., Abdoullahi Dehaki, M., Bayati, H., Mashayekhi, N., Varmazyar, S., Rahimian, Z., Asadi Anar, M., Shafiei, D., & Mohebbi, A. (2024). Machine learning-based algorithms for virtual early detection and screening of neurodegenerative and neurocognitive disorders: A systematic review. *Frontiers in Neurology*, 15, Article 1413071. <https://doi.org/10.3389/fneur.2024.1413071>