

Benefits and ethical considerations surrounding the integration of artificial intelligence (AI) into risk assessment in forensic psychology and psychiatry

Evangelou K¹, Pavlidou L¹, Belimezakis N¹, Zervas V¹, Hadjiphilippou M², Louka C³

¹ Department of Medicine, National and Kapodistrian University of Athens, Athens, Greece

² Department of Nursing, National and Kapodistrian University of Athens, Athens, Greece

³ Department of Chemistry, University of Cyprus, Nicosia, Cyprus

INTRODUCTION

Artificial intelligence (AI) and machine-learning (ML) models are increasingly proposed for violence and recidivism risk assessment in forensic psychiatry and psychology, where risk judgements guide detention, supervision and conditional release.

Existing structured instruments show only moderate discrimination and are time-consuming, while unaided clinical judgement is less reliable. Recent studies using ML on electronic health records report area-under-the-curve (AUC) values in the mid-to-high 0.70s, comparable with or better than many structured instruments, but generalisability and methodological quality remain mixed.

Ethical debate has intensified over whether such tools will improve fairness and consistency or instead amplify bias and opacity in high-stakes forensic decision-making.

AIM

To summarise key benefits and ethical concerns of integrating AI-based tools into forensic risk assessment and to outline safeguards for responsible clinical and legal use.

METHODS

Narrative review and ethical-legal analysis of three bodies of literature:

- empirical studies of ML-based violence and recidivism prediction in forensic or high-risk psychiatric populations (including natural language processing models using clinical text or threatening communications);
- reports of AI-enabled early-warning systems in mental-health services; and scholarship and policy documents on algorithmic risk assessment.

Evidence was organised into domains of performance and clinical utility, fairness, transparency, privacy and data governance, and responsibility.

RESULTS

Across forensic and other psychiatric samples, ML models generally show discrimination comparable to or better than actuarial instruments. For example, a JAMA Network Open study using routine clinical notes to predict inpatient violence reported internal AUCs of 0.76-0.80, although cross-site validation dropped to the mid 0.60s, underscoring the need for local calibration.

Other work with nursing electronic medical records found a violence rate of about 20% and demonstrated that aggressive incidents can be predicted with acceptable accuracy (AUC 0.68) and potentially embedded into routine ward workflows.

Natural language processing models applied to electronic health records and incident reports can automatically identify dynamic behavioural and linguistic markers (eg, agitation, threats, escalating conflict) that precede violence and workplace assaults, providing earlier warning for targeted de-escalation and staff protection.

In criminal-justice settings, algorithmic tools standardise criminogenic factors and offer scalable triage for bail, sentencing and community supervision, making explicit how prior offending, substance misuse and social instability shape risk categorisation.

Emerging explainable-AI techniques, including feature-importance plots, case-based reasoning and interpretable models such as decision trees and Bayesian networks, have been shown to increase interpretability in forensic psychiatry and digital forensics,

partially opening the “black box” to clinicians, courts and defendants.

Ethical concerns centre on the risk that models trained on historical criminal-justice and mental-health data reproduce or amplify structural disparities, as illustrated by analyses of COMPAS and related recidivism tools showing higher false-positive rates for minority groups despite similar reoffending rates.

The opacity of many models conflicts with requirements for forensic transparency and due process, limiting the ability of evaluatees and courts to understand or contest scores.

Large-scale reuse of sensitive psychiatric and criminological data for AI development raises privacy, consent and governance challenges, particularly when data are repurposed beyond their original context or shared across agencies.

Finally, responsibility for erroneous high-stakes predictions is diffuse, and automation bias may lead clinicians and judges to overweight numerical risk estimates relative to nuanced formulation and the person’s own account.

CONCLUSION

AI-based risk assessment tools in forensic psychiatry and psychology offer promising but incremental gains in predictive accuracy, particularly via NLP and dynamic early-warning applications.

These benefits are inseparable from ethical, legal and social risks. Ethical integration requires treating AI outputs as one strand within a transparent, clinically led and rights-respecting assessment; local validation and bias auditing; strong data-protection and governance; allocation of accountability; and training of clinicians, legal actors and service users in their strengths and limitations. Under such safeguards, AI may enhance rather than erode the fairness and legitimacy of high-stakes forensic decisions.

CORRESPONDENCE

Kyriacos Evangelou; National and Kapodistrian University of Athens; NKUA

evangeloukyriacos@gmail.com