Est. 1841 | YORK ST JOHN UNIVERSITY

Olawade, James O., Ebo, Titus Oloruntoba, Alabi, John Oluwatosin, Makanjuola, Babajide David, Egbon, Eghosasere and Olawade, David (2026) Digital twin technology in forensic mental health. Journal of Forensic and Legal Medicine, 120. p. 103137.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/14654/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.1016/j.jflm.2026.103137

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

Journal of Forensic and Legal Medicine 120 (2026) 103137

Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

journal homepage: www.elsevier.com/locate/yjflm

Review

Digital twin technology in forensic mental health

James O. Olawade [a], Titus Oloruntoba Ebo [b], John Oluwatosin Alabi [c], Babajide David Makanjuola [d], Eghosasere Egbon [e], David B. Olawade [f,g,h,*]

[a] Department of Guidance and Counselling, Adekunle Ajasin University, Akungba-Akoko, Nigeria
[b] Forensic Mental Health Unit, Nottinghamshire Healthcare NHS Foundation Trust, United Kingdom
[c] Department of Business and Management, University of Sussex Business School, University of Sussex, Falmer, Brighton, BN1 9RH, United Kingdom
[d] Sheffield Business School, Department of Accounting, Banking and Finance, Sheffield Hallam University, Howard St, Sheffield City Centre, Sheffield, S1 1WB, United Kingdom
[e] Department of Tissue Engineering and Regenerative Medicine, Faculty of Life Science Engineering, FH Technikum, Vienna, Austria
[f] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[g] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham, ME7 5NY, United Kingdom
[h] Department of Public Health, York St John University, London, United Kingdom

ARTICLE INFO

Keywords:
Digital twins
Forensic mental health
Risk assessment
Digital phenotyping
Precision psychiatry

ABSTRACT

Forensic mental health services face significant challenges in managing violence and self-harm risks, optimizing therapeutic security, and planning pathways for individuals with serious mental disorders within criminal justice systems. Traditional risk assessment tools provide static snapshots that degrade over time and offer limited personalization. Digital twin technology, which creates dynamic, data-driven computational replicas of real-world entities, presents a transformative opportunity to enhance decision-making in this complex field. This narrative review synthesizes emerging concepts, opportunities, and risks surrounding the use of digital twin technology in forensic mental health, examining how this innovation could augment clinical practice while addressing critical ethical and legal considerations. We conducted a narrative review of recent literature on digital twins in healthcare, digital psychiatry, risk management in forensic mental health, and related ethical frameworks, synthesizing findings from peer-reviewed journals, consensus statements, and policy documents to map plausible applications, technical constraints, and governance requirements specific to forensic mental health contexts. Digital twins could enhance violence and self-harm risk management through continuous updating, personalize care pathways across prisons, courts, and secure hospitals, optimize ward staffing and security protocols, and support rights-respecting care planning. However, deployment requires robust attention to data provenance, algorithmic fairness, transparency, clinical validity, and human rights safeguards. We identify a staged translational pathway with essential guardrails for safe implementation. While digital twin technology holds considerable promise for forensic mental health, realizing these potential demands rigorous validation, strong governance frameworks, and sustained co-design with service users, clinicians, and legal stakeholders to ensure safety and rights protection.

1. Introduction

Forensic mental health services operate at one of the most challenging intersections in healthcare, where psychiatry meets law and public protection imperatives.[1] These services support individuals with serious mental disorders who have offended or pose significant risks, delivering care within prisons, secure hospitals, courts, and community settings.[2] The decisions made within this system have profound consequences not only for public safety but also for the liberty, dignity, and recovery prospects of vulnerable individuals. Clinicians must constantly balance therapeutic goals against risk management, navigating complex legal frameworks while upholding human rights and ethical principles.[3] This delicate equilibrium demands sophisticated tools that can support nuanced, individualized decision-making.

Current approaches to risk assessment and care planning in forensic mental health rely heavily on structured professional judgement instruments such as the Historical Clinical Risk Management 20 (HCR-20), the Violence Risk Appraisal Guide (VRAG), and the Psychopathy

* Corresponding author.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.jflm.2026.103137
Received 9 November 2025; Received in revised form 17 December 2025; Accepted 15 April 2026
Available online 16 April 2026
1752-928X/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Checklist Revised (PCL-R).[4,5] These tools provide systematic frameworks for evaluating static historical factors, dynamic clinical variables, and risk management considerations.[6] While they represent significant advances over unstructured clinical judgement, meta-analytic research reveals important limitations. Their predictive validity varies considerably across settings and populations, deteriorates over time, and provides limited guidance for individualized interventions.[5] Moreover, these instruments typically produce single point-in-time assessments that cannot capture the rapid fluctuations in risk states that characterize many forensic patients' trajectories.[7] These limitations persist even with optimal implementation, suggesting that non-technological improvements alone, such as enhanced training, increased staffing, or improved multi-disciplinary collaboration, while valuable, may be insufficient to address the fundamental constraint of static assessment approaches in dynamic clinical environments.

The broader landscape of healthcare is witnessing a digital transformation that promises to revolutionize how we monitor, predict, and personalize care.[8] Digital twin technology stands at the forefront of this revolution.[9] Originally developed in aerospace and manufacturing to create virtual replicas of physical systems, digital twins have recently gained traction in medicine.[10] These sophisticated computational models integrate multiple data streams, update continuously with real-world observations, and enable scenario testing before implementing changes in practice.[11] Digital twins are particularly relevant for modelling complex psychiatric phenomena in forensic contexts because they can simultaneously integrate biological markers (sleep, activity, heart rate variability), clinical observations (symptom fluctuations, medication responses), behavioural patterns (social interactions, ward incidents), and contextual factors (environmental stressors, interpersonal dynamics) into unified, dynamic representations that capture the multi-factorial nature of mental health risk trajectories. Early applications in cardiology, oncology, and intensive care demonstrate the potential for digital twins to improve outcomes through precision medicine approaches.[12] The extension of this technology to mental health, and specifically to forensic mental health, represents both an exciting frontier and a critical test of our ability to deploy advanced technologies ethically in high-stakes environments.[13]

Digital phenotyping, the moment-by-moment quantification of the individual-level human phenotype using data from personal digital devices, provides a complementary innovation that could fuel forensic mental health digital twins.[14] Smartphones, wearable sensors, and environmental monitors can now capture continuous streams of behavioural, physiological, and contextual information such as sleep patterns, physical activity, social interactions, and ambient conditions.[15,16] In general psychiatry, research demonstrates that these digital phenotypes correlate meaningfully with symptom fluctuations, medication adherence, and relapse risk.[17] When combined with traditional clinical data and sophisticated analytical approaches, these streams could enable the kind of continuous, contextualized risk assessment that forensic settings urgently need.[18] However, the coercive nature of many forensic environments raises profound questions about voluntariness, privacy, and the potential for technology to entrench rather than reduce restrictiveness.[19]

Despite the promise of digital twin technology, its application to forensic mental health remains largely conceptual and speculative, with no robust clinical implementations yet documented in the literature.[20] This represents a critical limitation that must be acknowledged from the outset: the evidence base consists primarily of theoretical frameworks, adjacent applications in other healthcare domains, and digital phenotyping research in non-forensic psychiatric populations. The technology is immature for forensic contexts, and many of the applications discussed in this review should be understood as aspirational rather than currently achievable. This creates both an opportunity and a responsibility to chart a careful course forward. The rationale for developing digital twin technology in forensic mental health, rather than pursuing exclusively non-technological improvements, rests on three considerations: first, the documented limitations of current risk assessment approaches suggest that incremental refinements to existing methods may reach a ceiling of predictive validity; second, the maturation of enabling technologies in digital health demonstrates technical feasibility even if forensic implementations remain absent; and third, the unique opportunity to address system-level optimization (ward environments, pathways, resource allocation) alongside individual-level care represents a capability that traditional quality improvement methods cannot readily achieve. The novelty of this review lies in its systematic mapping of digital twin concepts specifically to the unique requirements and constraints of forensic mental health, bridging literature that have remained largely separate.

This narrative review addresses the gap between emerging technological capabilities and forensic mental health needs by synthesizing dispersed knowledge across healthcare technology, digital psychiatry, risk assessment research, and medical ethics. Narrative reviews are appropriate for exploring emerging topics where evidence remains preliminary and conceptual frameworks are still developing, allowing flexible integration of diverse sources to map the landscape of possibilities, constraints, and priorities. Our synthesis draws on literature from peer-reviewed journals, expert consensus statements, policy documents, and related fields to construct a coherent understanding of digital twin applications, technical requirements, ethical implications, and governance needs specific to forensic mental health.

Our aim is to provide a comprehensive synthesis that can guide researchers, clinicians, policymakers, and technology developers towards safe, effective, and ethical implementation while acknowledging substantial uncertainties and the speculative nature of many proposed applications. The specific objectives are to: (1) define digital twin technology in the forensic mental health context; (2) identify and critically appraise potential applications across person, ward, and system levels; (3) evaluate the current evidence base and readiness for deployment; (4) analyze ethical, legal, and social implications with particular attention to coercion, surveillance, institutional power dynamics, and structural biases embedded in forensic data; (5) propose a staged translational pathway with governance safeguards; and (6) establish research priorities for the field.

2. Conceptual framework: digital twins across levels

2.1. Person-level digital twins

Person-level digital twins represent individual patients by synthesizing electronic health records (diagnoses, medications, assessments, progress notes), incident logs (aggression, self-harm, rule violations), legal documentation (index offences, court reports, tribunal decisions), continuous monitoring from wearables (sleep, activity, heart rate variability, circadian rhythms), proximity sensors (movement patterns, social interactions), and environmental data (noise, crowding, temperature). These twins perform dynamic risk state estimation providing short-term forecasts (hours to days) of violence, self-harm, or absconding; risk trait profiling identifying enduring patterns and vulnerabilities; counterfactual scenario testing simulating intervention impacts (medication adjustments, observation changes, graded leave, environmental modifications) before implementation; and generation of explainable narratives highlighting key risk contributors and modifiable targets.[21] Outputs support multiple high-stakes decisions including care planning (therapeutic priorities, intervention intensity, resource allocation), observation level refinement based on continuously updated estimates rather than periodic assessments, leave planning through simulation of progression pathways testing readiness and conditions, and tribunal preparations providing structured, data-informed narratives for transparent scrutiny of detention and risk judgements.[22]

However, critical ethical considerations accompany person-level applications. The intensive individual monitoring required raises concerns about surveillance, loss of privacy, and the potential for digital

twins to function as tools of institutional control rather than therapeutic support. In forensic settings characterized by involuntary detention and significant power imbalances between staff and patients, the implementation of person-level digital twins risks exacerbating existing coercive dynamics. Patients may feel pressured to accept monitoring or may modify their behaviour not for therapeutic benefit but to appear compliant and influence release decisions. Furthermore, the data feeding person-level twins inevitably reflects historical patterns of institutional bias, for instance, if Black patients have historically been subject to more restrictive interventions or more frequent incident reporting, algorithmic models trained on such data may perpetuate these disparities by flagging similar patients as higher risk. These ethical challenges must be addressed through robust governance, transparent consent processes where genuinely possible, and continuous bias monitoring.

2.2. Ward and service-level digital twins

Ward-level digital twins shift focus from individuals to the operational environment, modelling patient mix (numbers, acuity, diagnosis distribution, risk scores, incident rates, medication complexity), staff characteristics (skill mix, qualified nurses, support workers, specialists, shift patterns, leave coverage, supervision ratios), physical layout, observation policies, security protocols, room occupancy, and environmental conditions. The primary function is "what if" scenario testing to optimize safety, therapeutic climate, and resource efficiency by simulating impacts of admitting high-acuity patients, evaluating de-escalation training effects on staff confidence and incident prevention, optimizing staffing rosters against predicted patient needs to ensure adequate coverage without excessive costs or burnout, and supporting physical environmental redesign including changes to layout, sensory rooms, outdoor access, or activity spaces.[23] This capability addresses longstanding concerns about institutional environments that, while designed for security, may inadvertently increase distress and aggression through overcrowding, under-stimulation, or lack of privacy, enabling services to test interventions virtually and make evidence-informed investments in environmental improvements that reduce restrictive interventions while maintaining safety.

Ward-level applications present distinct ethical considerations related to collective surveillance and potential unintended consequences. Aggregated monitoring might identify patterns useful for violence prevention but could also enable management practices that prioritize institutional efficiency over patient dignity and autonomy. For example, optimization algorithms might recommend staffing patterns or admission decisions that reduce costs or incident rates but inadvertently compromise therapeutic relationships or patient choice. Additionally, ward-level data collection may normalize comprehensive surveillance as an institutional expectation, creating environments where privacy becomes exceptional rather than normative. The political dimensions of forensic decision-making further complicate matters: recommendations from ward-level digital twins may conflict with professional judgement, union agreements, or institutional cultures, potentially generating resistance or selective adoption that undermines effectiveness. Safe implementation requires transparency about optimization criteria, stakeholder engagement in defining acceptable trade-offs, and mechanisms to challenge recommendations that conflict with patient-centered values.

2.3. Pathway and population-level digital twins

Pathway-level digital twins model patients flow through the forensic mental health system, from initial contact with criminal justice agencies through various secure settings and ultimately to conditional discharge and community integration.[20] These system-level models would incorporate transition probabilities between settings (e.g., from prison to medium secure hospital, from medium to low secure, from low secure to community), lengths of stay in each setting, readmission rates, and outcomes including community tenure, violent recidivism, quality of life, and costs.[24]

Population-level digital twins could evaluate policy changes before implementation.[25] For instance, policymakers considering alternatives to custody for certain offences could simulate the likely impact on secure hospital demand, community service capacity, and public safety outcomes.[26] Changes to conditional discharge criteria, such as relaxing residence requirements or reducing supervision intensity, could be modelled to estimate effects on recall rates and resource requirements [27–29]. Such models would need to incorporate equity considerations, examining whether policy changes differentially affect subgroups defined by ethnicity, gender, diagnosis, or socioeconomic factors.

These higher-level digital twins complement person and ward-level applications by addressing system design questions that individual-level data alone cannot answer. They also provide context for interpreting person-level twin outputs: an individual's predicted trajectory exists within a system characterized by resource constraints, wait times, and policies that shape both opportunities and barriers to recovery. However, pathway and population-level models risk concealing structural injustices within aggregate statistics. If current forensic mental health systems disproportionately detain ethnic minorities or disadvantage women due to services designed primarily for men, models that treat these patterns as baseline assumptions rather than targets for transformation may inadvertently legitimize inequitable structures. Furthermore, population-level optimization focused narrowly on metrics like bed occupancy rates or recidivism may obscure outcomes that matter most to service users, community integration, meaningful relationships, employment, quality of life, if these are not explicitly prioritized in modelling objectives. Ethical deployment of population-level digital twins demands critical examination of whose interests are served by optimization, whose voices inform the definition of desirable outcomes, and how models can be designed to challenge rather than replicate structural inequalities.

The conceptual framework for forensic mental health digital twins operates across three interconnected architectural levels, as illustrated in Fig. 1, each serving distinct but complementary functions in supporting clinical decision-making and system optimization. Table 1 summarizes the key components and data streams for digital twins across these three levels.

3. Enabling technologies: digital phenotyping and continuous monitoring

Digital phenotyping provides continuous data streams that could fuel person-level digital twins.[30] The concept refers to the quantification of human behaviour using data from personal digital devices, particularly smartphones and wearables.[31] In mental health research, studies have demonstrated that passive smartphone sensors can detect patterns associated with depression, anxiety, psychosis, and bipolar disorder.[32,33] Sleep regularity derived from accelerometer data correlates with mood stability.[34] Global positioning system mobility patterns reflect behavioural activation.[35] Communication metadata (call and message frequency, timing patterns) provide proxies for social functioning.[36] Screen time and app usage patterns may indicate engagement or withdrawal.[37] However, it is crucial to note that these associations derive primarily from research in general psychiatric populations and community settings rather than forensic environments. The extent to which digital phenotyping patterns generalize to individuals in secure settings, where environmental constraints, institutional routines, and the stress of detention may fundamentally alter baseline patterns, remains unknown and requires dedicated validation research.

Wearable devices offer additional physiological signals including heart rate, heart rate variability, skin conductance, and body temperature.[38] Heart rate variability shows promise as a marker of autonomic nervous system regulation and stress response, with reduced variability

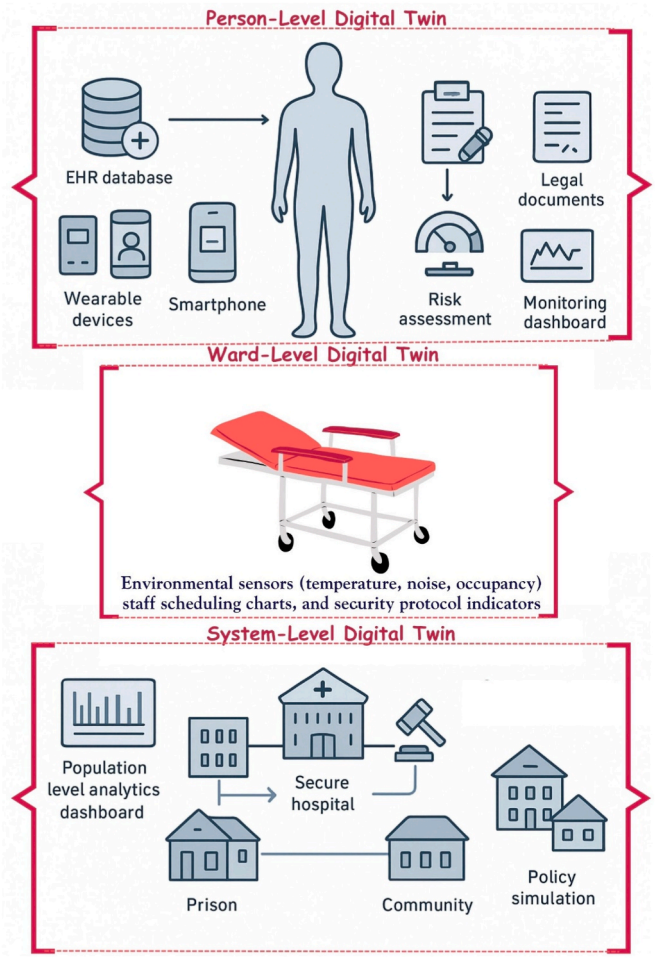

Fig. 1. Three-level digital twin (DT) architecture for forensic mental health. The framework integrates person-level DTs (patient data, risk assessment), **ward-level DTs** (milieu optimization, restrictive practice reduction), and system-level DTs (patient pathways, resource allocation) with bidirectional data flow for continuous learning and evidence-informed decision-making.

associated with various psychiatric conditions and predictive of aggression in some studies [39–41]. Actigraphy provides detailed rest-activity patterns that capture sleep quality, circadian rhythms, and daytime activity levels, all relevant to risk and recovery in severe mental illness.[42] Again, the evidence base comes predominantly from non-forensic populations, and whether these physiological markers maintain their predictive validity under conditions of involuntary detention, restricted freedom of movement, and institutional stress requires empirical verification.

Environmental sensors deployed in secure settings could complement personal devices by capturing ward-level contextual factors.[43] Noise level monitors might identify periods of high ambient sound that increase arousal and tension.[44] Occupancy sensors could detect crowding in common areas that may precipitate conflict.[45] Light sensors could inform understanding of whether patients receive adequate exposure to natural daylight, which affects circadian rhythms and mood.[46] Temperature and air quality sensors might reveal environmental stressors affecting comfort and wellbeing.[47]

The integration of these streams with traditional clinical data creates opportunities for rich, contextualized risk assessment. For example, a digital twin might detect that an individual's violence risk escalates following nights of disrupted sleep combined with high ward noise levels and recent medication non-adherence, a pattern that could be addressed through environmental interventions, sleep hygiene support, and enhanced medication monitoring.[48] This represents an aspirational capability rather than a demonstrated one: the technical integration of multimodal data streams, the development of validated algorithms linking patterns to outcomes, and the establishment of causal pathways (rather than mere correlations) between factors and risk all remain substantial research challenges. This kind of individualized, mechanistic understanding contrasts with the actuarial approach of traditional risk tools.

However, the deployment of continuous monitoring in forensic settings raises profound ethical questions. Many individuals in secure care are detained involuntarily and may not be able to provide free, informed consent to monitoring.[49] The power imbalance between staff and patients, combined with patients' understandable desire to present themselves favourable to influence discharge decisions, creates risks of coercion.[50] Transparency about what is monitored, how data are used, who has access, and how monitoring relates to clinical care versus security becomes essential.[51] Governance frameworks must establish clear

Table 1
Components and data streams for forensic mental health digital twins.

Level	Primary Focus	Key Data Streams	Core Functions	Decision Support Applications	Key Ethical Considerations
Person-Level (Clinical DT)[13]	Individual patient trajectories and risk states	Electronic health records; medications; psychological assessments; incident logs; legal status; wearable sensors (sleep, activity, heart rate variability); proximity beacons; environmental sensors; patient-reported outcomes	Dynamic risk estimation; risk trait profiling; counterfactual scenario testing; explainable narratives; uncertainty quantification	Care planning; observation level adjustment; medication optimization; leave progression; conditional discharge preparation; tribunal reports	Intensive surveillance; consent under coercion; privacy erosion; potential for behavior modification to influence release decisions; perpetuation of historical biases in incident reporting
Ward-Level (Operational DT)[20]	Secure unit functioning and therapeutic milieu	Patient census and acuity; staff skill mix and shifts; observation policies; security protocols; room occupancy; incident rates; environmental conditions (noise, temperature); activity schedules	Scenario testing for admissions; staffing optimization; environmental redesign simulation; seclusion/restraint reduction; de-escalation training evaluation	Admission decisions; staffing allocation; environmental modifications; policy testing; resource planning; safety protocol optimization	Normalization of collective surveillance; optimization prioritizing institutional efficiency over dignity; potential conflicts with professional judgment and stakeholder values; risk of depersonalization
Pathway-Level (Population DT)[13]	System flows and population outcomes	Transition rates between settings; lengths of stay; readmission rates; recidivism data; service capacity; wait times; costs; community outcomes; equity metrics across demographics	Policy simulation; capacity planning; pathway optimization; equity impact assessment; cost-effectiveness analysis	Service commissioning; policy evaluation; resource allocation; system redesign; alternatives to secure care; community service planning	Risk of legitimizing structural inequalities; concealment of systemic injustices within aggregate statistics; potential prioritization of institutional metrics over service user-defined outcomes; equity implications of policy recommendations

boundaries, ensuring that data serve therapeutic purposes and reduce rather than entrench restrictiveness. Moreover, the very act of continuous monitoring may alter behavior in ways that undermine both therapeutic relationships and the validity of data collected. Patients aware of being monitored may engage in strategic self-presentation, concealing genuine distress or modifying conduct to appear low-risk, potentially defeating the intended purpose of early risk detection while simultaneously imposing psychological burdens associated with persistent surveillance.

4. Potential applications and use cases

4.1. Dynamic violence risk assessment

Violence risk assessment represents perhaps the most obvious application for forensic mental health digital twins.[52] Current structured professional judgement tools provide valuable frameworks but produce static assessments that become outdated quickly.[53] A digital twin approach would augment baseline structured judgement with streaming data to produce horizon-specific risk updates.[54] Clinicians might receive alerts when short-term risk escalates based on real-time indicators, allowing preventive interventions before incidents occur.[55]

This approach moves beyond simple risk scores to actionable intelligence about modifiable drivers. Importantly, the digital twin would quantify uncertainty in its estimates, acknowledging areas where evidence is weak or conflicting. Clinicians would retain ultimate decision-making authority, using twin outputs as one input alongside their own observations, therapeutic relationships, and professional judgement. However, the feasibility of implementing such systems currently remains limited: algorithms capable of reliably distinguishing imminent risk from baseline variation in forensic populations have not been validated, false positive rates may be unacceptably high in low-base-rate environments (where most patients do not engage in violence), and the mechanisms by which early warning systems would trigger clinically meaningful responses without generating alarm fatigue or defensive practices require empirical investigation.

4.2. Medication and treatment optimization

Forensic mental health patients often receive complex medication regimens including antipsychotics, mood stabilizers, antidepressants, and medications to manage side effects.[56] Balancing efficacy against side effects such as weight gain, sedation, sexual dysfunction, and metabolic disturbance remains challenging, particularly because side effects may contribute to non-adherence that in turn increases relapse risk.[57,58] A digital twin could model individual responses to different medication combinations, incorporating pharmacogenetic data where available, prior response history, current symptoms, and side effect profiles.[59] This represents an aspirational application: the mechanistic understanding of how psychiatric medications work at the individual level remains incomplete, inter-individual variability in response is substantial and poorly predicted by available biomarkers, and the ability to accurately simulate medication effects prospectively has not been demonstrated in psychiatric populations.

Scenario testing would allow clinicians and patients to explore alternatives: "If we reduce the antipsychotic dose to minimize sedation, how might this affect symptom stability and risk over the next three months?[60] If we switch to a different mood stabilizer to address weight gain, what is the probability of destabilization during the transition?" Such simulations would support shared decision-making, helping patients understand trade-offs and participate meaningfully in treatment planning.[61] This transparency becomes particularly important in forensic contexts where medication can be administered compulsorily but where engaging patients' collaboration improves outcomes and respects autonomy to the greatest extent possible. Yet the validity of such scenario testing depends entirely on the accuracy of underlying models, and overstated certainty in predictions based on immature algorithms could mislead both clinicians and patients, potentially leading to suboptimal decisions disguised as evidence-based medicine.

4.3. Seclusion and restraint reduction

Reducing restrictive practices remains a major priority for forensic services, driven by both rights concerns and evidence that such interventions can traumatize patients and damage therapeutic relationships.[62] Ward-level digital twins could identify conditions that predict seclusion or restraint use, enabling preventive interventions.[63] Analysis might reveal incidents requiring restraint cluster at shift changes, suggesting improved handover processes or additional staffing at transition times. Or twins might show that combinations of patient acuity predict high-risk periods, informing bed management decisions.[64] These insights, while potentially valuable, are also achievable through traditional quality improvement methodologies including incident analysis, staff interviews, and observational studies, raising questions about whether digital twin technology offers sufficient added value to justify its costs and risks in this domain.

Scenario testing could evaluate interventions before implementation.[65] A service considering enhanced sensory modulation rooms might simulate their impact on incident rates by modelling patient usage patterns and hypothesized effects on arousal regulation.[66] Training programmes focused on trauma-informed de-escalation could be evaluated by estimating their likely effect on staff responses to early warning signs. Such modelling would support evidence-informed quality improvement while minimizing the risk of investing in interventions that prove ineffective.[67] However, the accuracy of such simulations depends on having validated models of complex multi-causal events (seclusion and restraint incidents), understanding of how novel interventions would alter staff behavior and patient responses, and baseline data of sufficient quality, conditions that may not obtain in many forensic services, limiting the reliability of scenario testing outputs.

4.4. Leave planning and conditional discharge

Decisions about graded community exposure and eventual discharge rank among the most consequential in forensic mental health. Current practice relies heavily on clinician judgement informed by risk tools, care team discussions, trial leaves, and multi-agency input.[68] Digital twins could augment this process by simulating different leave pathways and discharge scenarios. For an individual approaching conditional discharge, the twin might model trajectories under various supervision intensities, testing whether weekly versus fortnightly community psychiatric nurse visits would adequately support stability.[69,70] Simulations could incorporate contingency plans, examining how quickly risk might escalate if circumstances deteriorate and what monitoring would enable early detection.[71] These applications face the greatest uncertainty: long-term outcome prediction in community contexts is inherently limited by unknown future events, changes in social circumstances, and the difficulty of modeling complex interactions between individual vulnerabilities and community factors. Overreliance on algorithmic predictions in discharge decisions could lead to unjustified detention (if predictions are overly pessimistic) or inadequate support (if predictions are overly optimistic), with profound consequences for liberty and safety.

Importantly, such modelling would need to extend beyond recidivism outcomes to capture quality of life, community integration, employment, relationships, and patient-reported recovery outcomes.[72] A discharge plan that minimizes re-offending risk through maximum restriction may not represent optimal care if it unnecessarily limits freedom and impedes social recovery.[73] Digital twins should therefore incorporate multiple outcomes weighted according to patient values and legal requirements, supporting decisions that balance legitimate concerns. The challenge of defining and measuring such diverse outcomes,

eliciting patient values under conditions of involuntary detention where preferences may be strategically expressed, and weighting competing considerations algorithmically without inappropriate value judgments remains substantial and may ultimately prove intractable without preserved human judgment at the center of decision-making.

Table 2 outlines potential use cases and anticipated benefits across these domains, integrating consideration of both opportunities and constraints including evidence-based versus speculative claims.

5. Evidence base and clinical readiness

5.1. Current state of evidence

The evidence base for digital twins in forensic mental health remains conceptual and emerging, with no published studies describing fully implemented, validated systems in routine clinical practice, though constituent elements draw on mature research streams. Digital phenotyping research demonstrates associations between passively collected smartphone and wearable data and mental health outcomes (accelerometer-derived activity correlating with depression severity, sleep regularity predicting bipolar mood episodes, communication patterns predicting psychosis relapses), though most studies remain observational and cross-sectional with forensic populations underrepresented, limiting direct applicability.[74] Furthermore, the majority of digital phenotyping research examines associations rather than establishing causation, relies on convenience samples that may not represent forensic populations, and has not adequately addressed how findings generalize across cultural contexts, diagnostic categories, or institutional environments. The gap between demonstrating correlations in research settings and deploying predictive algorithms in high-stakes forensic decisions is substantial.

Machine learning applied to forensic risk assessment produces mixed results, with some studies reporting improved accuracy using large datasets while others find modest gains disappearing upon rigorous external validation or revealing unacceptable racial, gender, or socioeconomic bias, emphasizing the consistent need for algorithm transparency, multi-site external validation, and continuous monitoring for drift and bias rather than one-time validation.[75] Critical examinations of machine learning in criminal justice contexts more broadly reveal troubling patterns of bias amplification, lack of transparency, poor generalization across jurisdictions, and absence of meaningful accountability when predictions prove erroneous, concerns that may be even more acute in forensic mental health where detention decisions affect fundamental rights and where affected individuals often lack resources to challenge algorithmic recommendations effectively.

5.2. Readiness assessment

Assessing readiness for forensic mental health digital twins across technical, clinical, organizational, and ethical dimensions reveal variable preparedness.[13] Technical readiness appears moderate, with computing infrastructure, data storage, and analytical methods existing and improving, and many forensic services maintaining electronic health records albeit with considerable variation in data quality, completeness, and standardization, while integration of continuous monitoring from wearables and sensors remains rare in routine practice but technically feasible.[76] Clinical readiness faces greater challenges given limited clinician awareness and confidence with digital twin concepts, training curricula typically excluding digital health technologies or data science, and organizational cultures in secure settings prioritizing stability and risk aversion, though growing frustration with current tool limitations and increasing technological familiarity in younger clinician cohorts may support adoption.[77] Organizational readiness varies widely, with some forward-thinking secure hospitals and prison mental health teams embracing quality improvement methodologies, electronic record systems, and research partnerships that

Table 2
Potential use cases and anticipated benefits of digital twins in forensic mental health.

Use Case	Current Limitations	Digital Twin Contribution	Anticipated Benefits	Key Challenges	Evidence Status
Violence Risk Assessment[20]	Static snapshots; degraded temporal validity; limited individualization; moderate predictive accuracy	Continuous updating from multimodal data; personalized risk trajectories; explainable drivers; horizon-specific forecasts	Earlier detection of risk escalation; targeted preventive interventions; reduced incidents; more nuanced understanding of modifiable factors	Data quality and completeness; model calibration across contexts; false alarm management; clinician trust	Speculative: No validated algorithms exist for real-time forensic violence risk prediction; digital phenotyping associations derive from non-forensic populations
Medication Optimization	Trial-and-error approach; poor side effect prediction; limited personalization; adherence challenges	Simulation of medication responses; side effect profiling; pharmacogenetic integration; adherence pattern analysis	Enhanced efficacy; reduced side effects; improved adherence; shared decision-making; tribunal transparency	Limited mechanistic understanding of psychiatric medications; individual variability; placebo effects; non-pharmacological confounders	Speculative: Prospective medication response modeling not demonstrated in psychiatric populations; mechanistic understanding insufficient for reliable simulation
Seclusion/ Restraint Reduction	Reactive responses; limited understanding of precipitants; insufficient staff training; environmental factors underestimated	Ward-level scenario testing; incident pattern analysis; early warning systems; environmental modification simulation; de-escalation protocol optimization	Reduced restrictive interventions; improved staff confidence; better therapeutic climate; rights-respecting care; staff safety	Complex multi-causal incidents; difficulty isolating intervention effects; resistance to practice change; resource constraints	Partially feasible: Incident pattern analysis achievable with existing methods; prospective simulation of novel interventions remains unvalidated
Leave Progression	Risk aversion; one-size-fits-all protocols; limited evidence for optimal pacing; inadequate trial leave data	Simulated leave pathways; risk-benefit analysis across scenarios; incorporation of protective factors; personalized progression plans	Optimized leave progression; reduced unnecessary delays; evidence-informed decisions; improved patient and family confidence	Difficulty simulating community contexts; unknown variables; low base rates complicate validation; tribunal acceptance	Speculative: Long-term outcome prediction in community settings inherently uncertain; simulation accuracy undemonstrated
Conditional Discharge	High uncertainty; limited long-term prediction; poor understanding of community risk dynamics; variable support quality	Long-term trajectory simulation; supervision intensity modelling; contingency planning; equity-sensitive recommendations	Reduced unnecessary detention; lower recall rates; improved community integration; resource-efficient supervision; rights protection	Long-term prediction inherently uncertain; community data often unavailable; ethical concerns about false positives *vs* false negatives	Highly speculative: Longest prediction horizon with greatest uncertainty; risk of inappropriate algorithmic influence on liberty-depriving decisions

could facilitate pilot implementations, while others struggle with resource constraints, staff shortages, and outdated infrastructure, compounded by system fragmentation between prisons, hospitals, and community services creating data sharing barriers.[78] The political dimensions of forensic services, including tensions between therapeutic and custodial priorities, union concerns about staffing algorithms, historical distrust between services and oversight bodies, and competing demands from commissioners, regulators, and service users, create additional barriers to innovation that purely technical or clinical readiness assessments may overlook. Successful implementation will require navigating these political dynamics thoughtfully, building coalitions of support, and addressing legitimate stakeholder concerns rather than treating organizational resistance as irrational obstruction.

Ethical and regulatory readiness remains the most critical gap, as governance frameworks for artificial intelligence in healthcare evolve without specificity to forensic mental health's unique legal and ethical landscape, leaving unresolved questions about consent, data protection, algorithmic fairness, clinical accountability, and human rights compatibility, further complicated by mental health tribunal involvement requiring members to understand and trust digital twin outputs while legal precedents establish how such tools fit within existing detention and discharge frameworks. International variation in legal frameworks governing forensic mental health adds further complexity: systems in England and Wales operate under substantially different statutory frameworks than those in Scotland, other European jurisdictions, North America, or Australia, affecting consent requirements, detention criteria, discharge processes, and oversight mechanisms. Any digital twin implementations will require jurisdiction-specific adaptation rather than one-size-fits-all approaches, and guidance for such adaptation remains largely absent from current literature.

6. Ethical, legal, and social implications

6.1. Human rights and least restrictive practice

Forensic mental health operates under intense human rights scrutiny from international instruments (United Nations Convention on the Rights of Persons with Disabilities, European Convention on Human Rights) and domestic legislation (UK Mental Health Act) establishing principles of least restrictive practice, proportionality, and necessity, requiring any deployed technology to demonstrably reduce rather than entrench restrictiveness.[79] Digital twins pose both opportunities and risks: continuously updated risk assessment could enable earlier relaxation of restrictions reducing unnecessary detention, transparent recommendations could support patient advocacy by making decision rationales accessible to patients, families, and legal representatives, and ward-level optimization could improve conditions affecting dignity and wellbeing; conversely, risks include rationalizing excessive surveillance, reifying stigmatizing risk categories, or shifting decision-making towards algorithmic outputs lacking contextual judgement and compassion essential to humane care.[52,80] The history of risk assessment tools in forensic mental health reveals a troubling pattern: innovations promised to enable more individualized, less restrictive care but in practice were often deployed to justify continued detention and expand surveillance. Any digital twin implementation must be assessed not only against stated intentions but also against demonstrated impacts on actual restrictiveness of care, with mechanisms to identify and correct tendency towards net-widening or rights erosion.

Ensuring rights compatibility requires embedding human rights principles into digital twin design from inception by prioritizing least restrictive outcomes in objective functions, incorporating patient preferences and values into decision criteria, enabling transparent contestation of recommendations, maintaining human oversight such that algorithmic outputs inform but never determine decisions, and establishing independent oversight potentially including patient and advocacy representatives to review implementations for rights impacts.[22] Furthermore, digital twins should be designed with "opt-out" capabilities where clinically safe to do so, ensuring that patients who object to monitoring are not automatically subject to more restrictive conditions as a consequence, a principle that requires careful balancing against safety imperatives but remains essential to respecting autonomy to the maximum extent compatible with public protection mandates.

6.2. Privacy and data protection

The data streams feeding forensic mental health digital twins would be extraordinarily sensitive.[13] Psychiatric diagnoses, incident histories, index offences, medication details, sensor-derived behavioural patterns, and contextual information about relationships and social networks all carry substantial privacy implications.[81] Data protection law, including the UK General Data Protection Regulation and Data Protection Act, establishes requirements for lawful basis, purpose limitation, data minimization, security, and accountability.[82,83]

Processing health data about individuals in criminal justice contexts creates special category considerations. While consent provides the preferred lawful basis for data processing, forensic settings raise concerns about whether detained individuals can provide free, informed consent.[49] Alternative lawful bases such as public interest, particularly provision of health care and public health, may apply but require careful justification and proportionality assessment.[84] Purpose limitation principles demand clarity about whether data serve clinical care, security, research, or other purposes, with strict separation where incompatible purposes exist.[85] The risk of "function creep", where data collected for therapeutic purposes gradually become available for security, administrative, or research uses without adequate governance, is particularly acute in forensic settings where clinical and custodial functions overlap. Robust technical and policy safeguards against unauthorized access and purpose drift become essential, potentially including data segregation architectures that physically separate clinical from security data systems.

Recent investigations into mental health apps have revealed serious privacy failures including sharing of sensitive data with third parties, inadequate security leading to breaches, and lack of transparency about data practices.[86,87] Digital twins must exceed these low standards, implementing robust security including encryption, access controls, audit logging, and penetration testing.[88] Data sharing between organizational boundaries (e.g., between prison and hospital, or between health service requires) requires explicit governance including data sharing agreements, minimal necessary disclosure, and patient notification.[89] Moreover, forensic mental health data may be subject to additional disclosure requirements under criminal justice legislation, creating tensions with therapeutic confidentiality that digital twin implementations must navigate carefully, ideally with legal advice and clear protocols for responding to data access requests from law enforcement, courts, or other agencies.

6.3. Algorithmic fairness and bias

Forensic mental health exhibits substantial inequalities, with ethnic minorities (particularly Black individuals) over-represented in secure services relative to general population prevalence even after accounting for known risk factors, women receiving care in settings designed primarily for men, and individuals with intellectual disabilities facing communication barriers and detention for behaviours that might not trigger forensic involvement in neurotypical individuals.[90] Digital twins trained on historical data risk perpetuating or amplifying these inequalities if past clinician decisions reflected implicit bias (lower thresholds for labelling Black patients as dangerous, higher readiness to use restrictive interventions), if monitoring data are collected unequally (more intensive surveillance of certain groups), or if outcome data reflects differential policing and criminal justice processing.[91] The structural biases embedded in forensic data are not incidental noise but

reflect deep-seated inequalities in how mental health and criminal justice systems operate. Algorithms trained on such data will learn to replicate, and potentially amplify through feedback loops, these patterns unless explicit counter-measures are implemented. Yet even with fairness-aware algorithms, fundamental questions remain about whether it is possible to build equitable systems atop unjust foundations or whether more radical transformation of forensic mental health itself is required before algorithmic tools can be deployed ethically.

Addressing fairness requires proactive bias audits at multiple stages: during data collection examining whether certain groups are over- or under-represented and whether data quality differs systematically, during model development employing fairness-aware algorithms that explicitly test for and mitigate disparate impact across protected characteristics, and during deployment continuously monitoring outcomes disaggregated by ethnicity, gender, age, diagnosis, and other relevant categories while investigating and correcting emergent disparities, with counterfactual fairness testing examining whether recommendations would differ if only a person's protected characteristic changed.[92,93] Importantly, fairness extends beyond statistical parity to encompass substantive justice, meaning even models producing similar accuracy across groups may perpetuate inequality if disadvantaging historically marginalized groups, necessitating engagement with affected communities, patient advocates, and equality specialists to inform fairness criteria and acceptability thresholds. Meaningful engagement with affected communities requires more than token consultation: it demands genuine power-sharing in defining what fairness means, establishing accountability mechanisms when algorithms produce discriminatory outcomes, and ensuring that concerns raised by marginalized groups can halt or modify implementations rather than being noted and overridden by technical or administrative priorities.

6.4. Transparency and explainability

Legal processes in forensic mental health demand transparency.[94] Mental health tribunals scrutinize the evidence basis for detention and discharge decisions, with clinicians required to justify their assessments and recommendations.[95] Patients and legal representatives must be able to challenge opinions and present contrary evidence.[96] Introducing algorithmic decision support could undermine this transparency if recommendations emerge from "black box" models that clinicians cannot explain, and tribunals cannot interrogate.[97]

Explainable artificial intelligence techniques offer partial solutions. Methods such as attention mechanisms, saliency maps, local interpretable model-agnostic explanations, and Shapley additive explanations can highlight which input features most influenced a particular prediction.[98] However, these techniques provide post-hoc rationalization rather than guaranteeing model decisions are inherently interpretable. Furthermore, post-hoc explanations may be misleading or unstable: small changes in inputs can produce dramatically different explanations, techniques may highlight features that are correlates rather than causes of outcomes, and explanations can appear plausible while failing to capture actual model logic. For forensic applications where liberty interests are at stake, post-hoc explainability may be insufficient, consideration should be given to restricting acceptable algorithms to inherently interpretable models (such as decision trees, rule-based systems, or generalized additive models) even if these sacrifice some predictive performance compared to black-box alternatives. For forensic applications, model cards documenting design choices, training data, validation performance, limitations, and intended uses should accompany any deployment.[99] Clinicians should receive training enabling them to understand and communicate model logic, uncertainty, and constraints.[100]

Explanations must be tailored to audiences. Clinicians require technical details sufficient to judge reliability and applicability to specific cases. Patients need accessible explanations that empower them to understand and question recommendations. Tribunal members need summaries focused on legal criteria such as necessity and proportionality. Legal representatives need sufficient documentation to mount challenges. Developing multi-level explanation approaches that serve these diverse needs represents a substantial design challenge. Minimum standards for legally adequate explanations in forensic contexts should specify identification of key factors contributing to predictions with quantified importance weights, disclosure of uncertainty ranges and confidence intervals, documentation of model limitations including known failure modes and circumstances where predictions are unreliable, clear statements of what the model does and does not assess (for instance, a violence risk model does not assess treatment needs, discharge readiness, or quality of life), comparisons to relevant baseline rates and alternative assessment methods, and information about how to challenge or override algorithmic recommendations. These standards should be developed collaboratively with tribunal members, legal practitioners, and patient advocates to ensure they meet the needs of legal processes.

6.5. Clinical accountability and governance

Digital twins are decision assistants, not autonomous actors. Clinical and legal accountability must remain with human decision-makers.[21] However, the introduction of sophisticated algorithmic tools can create accountability ambiguity: if a clinician follows a digital twin recommendation that proves mistaken, who bears responsibility? If a clinician overrides a twin recommendation and an adverse outcome occurs, does this create increased liability? These questions have not been resolved in existing case law, and the absence of clear legal frameworks creates risks for both clinicians (who may face blame regardless of whether they follow or ignore algorithmic recommendations) and patients (who may receive suboptimal care as clinicians defer excessively to algorithms to avoid perceived liability for overrides). Establishing accountability frameworks before deploying digital twins, rather than allowing them to emerge reactively through litigation, should be a governance priority.

Clarifying accountability requires explicit governance structures.[101] Digital twin developers bear responsibility for rigorous design using appropriate methods and validated algorithms, transparent documentation of system capabilities, limitations, and intended uses, ongoing monitoring of deployed systems for performance degradation or safety signals, and prompt notification of known issues or necessary updates to implementers and users.[102] Implementers (healthcare organizations adopting digital twins) bear responsibility for appropriate deployment contexts matching system specifications, staff training ensuring users understand system capabilities and limitations, integration with clinical workflows that preserves rather than replaces professional judgment, local monitoring of outcomes and safety, and maintaining clear policies about when and how algorithmic recommendations may be overridden. Clinicians bear responsibility for assessing whether twin recommendations suit individual cases, recognizing model limitations and uncertainty, exercising professional judgment integrating algorithmic outputs with other information sources, and clearly documenting clinical reasoning whether following or overriding recommendations.[103]

Organizational governance, including designated accountable officers, clinical safety oversight committees, and incident reporting systems, should monitor digital twin performance and adverse events.[104] Shared accountability models recognizing that safe deployment requires contributions from developers, implementers, clinicians, and oversight bodies, rather than locating responsibility solely with any single party, may better reflect the reality of sociotechnical systems while avoiding diffusion of responsibility that leaves no one accountable when harms occur.

Safety cases, borrowed from high-reliability industries such as aviation and nuclear energy, provide a potential governance model. A safety case comprises structured argument and evidence demonstrating that a system is acceptably safe for specific purposes within defined constraints.[105] For a forensic mental health digital twin, the safety case

would articulate the intended application, the evidence supporting its validity and reliability, the residual uncertainties and risks, the controls and safeguards in place, and the monitoring and review processes. Regulators, ethics committees, and service providers could scrutinize safety cases before authorizing deployment, with ongoing reviews to maintain authorization. Safety case components specific to forensic mental health digital twins should include clear specification of intended use cases with explicit exclusions (for instance, intended for short-term violence risk forecasting in inpatient settings but not validated for long-term predictions, community settings, or discharge decisions), validation evidence from forensic populations demonstrating performance across demographic subgroups and clinical presentations, bias audit results with plans for continuous fairness monitoring, analysis of failure modes including circumstances where predictions would be unreliable, safeguards against misuse or over-reliance, interfaces and training designed to support appropriate clinical judgment, incident reporting and learning systems, and sunset clauses requiring periodic revalidation or system retirement.

Table 3 summarizes key governance considerations and proposed safeguards.

7. Limitations of this review

7.1. Evidence base limitations

This review faces inherent constraints arising from the nascent state of the field. No published studies describe fully implemented, validated digital twins in routine forensic mental health practice. Consequently, much of our analysis extrapolates from adjacent domains including digital twins in other healthcare specialties, digital phenotyping in general psychiatry, and machine learning applications to forensic risk assessment. While these extrapolations rest on reasonable assumptions, they remain speculative until empirical evidence specific to forensic mental health digital twins accumulates. Readers should interpret the applications and benefits discussed in this review as potential rather than demonstrated capabilities, with substantial research and development required before clinical deployment becomes appropriate.

The heterogeneity of forensic mental health systems internationally limits generalizability. Services in England and Wales differ substantially from those in Scotland, which differ from European, North American, and Australian systems in legal frameworks, organizational structures, funding models, and populations served. Findings from one jurisdiction may not translate elsewhere. Our synthesis attempts to identify principles that transcend jurisdictional specifics, but practical implementation will require localization. Specific jurisdictional considerations affecting digital twin implementations include: (1) consent and capacity frameworks (varying legal standards for when detained individuals can provide valid consent); (2) detention criteria and review processes (affecting which decisions algorithmic tools might inform); (3) data protection regimes (affecting lawful bases for processing and cross-border data sharing); (4) forensic service structures (affecting feasibility of system-level modeling); (5) cultural contexts including public attitudes toward surveillance, technology in criminal justice, and mental health; and (6) resource availability for implementation and governance. Adaptation guidance for different jurisdictions should be developed collaboratively with local stakeholders, legal experts, and patient representatives rather than assuming universal applicability.

7.2. Methodological limitations

As a narrative review, this work does not employ systematic search, quality assessment, and meta-analytic synthesis methods characteristic of systematic reviews. Our literature coverage, while extensive, may have missed relevant sources. Our interpretation reflects our perspectives and expertise but may not represent consensus across diverse stakeholders. The lack of formal quality assessment means we have weighted sources based on judgement rather than standardized criteria.

The interdisciplinary nature of this topic spans computer science, psychiatry, law, ethics, and health services research. While we have attempted to synthesize across these domains, depth in any single area necessarily remains limited. Specialists in each field might identify gaps or nuances our generalist approach misses.

7.3. Temporal limitations

The technologies, regulations, and practices relevant to digital twins evolve rapidly. Findings current at the time of writing may become outdated as new evidence emerges, technologies advance, or policies change. Digital twin definitions themselves remain contested, with ongoing efforts towards standardization. Our synthesis represents a snapshot of understanding in 2025 but should be updated regularly as the field matures.

7.4. Scope limitations

We have necessarily omitted or treated superficially several important considerations. Detailed technical specifications for data architecture, interoperability standards, and algorithm design exceed our scope but will prove critical for implementation. Cost-effectiveness analysis and health economic modelling would inform commissioning decisions but require empirical data not yet available. Change management, training curricula, and workforce development merit dedicated attention beyond what we provide. International perspectives, particularly from low- and middle-income countries, remain under-represented in both our review and the underlying literature. Additionally, this review has not adequately addressed the patient and carer perspective on digital twin technology: what are service users' attitudes toward continuous monitoring, algorithmic risk assessment, and data-driven care planning? What concerns do they prioritize, and what safeguards would they require for acceptable deployment? Addressing these questions through qualitative research, patient consultation, and participatory design represents an essential prerequisite for ethical implementation that this review identifies but does not fulfill.

Realizing the potential of digital twins in forensic mental health requires a carefully staged implementation approach with robust risk mitigation at each phase, as outlined in Fig. 2, ensuring that innovation proceeds safely within appropriate ethical and clinical safeguards. This pathway should be understood as a roadmap requiring adaptation based on emerging evidence and stakeholder feedback rather than a fixed prescription.

8. Research and policy recommendations

Given the substantial uncertainties and ethical complexities identified throughout this review, we propose the following concrete recommendations to guide responsible development of digital twin technology in forensic mental health:

8.1. Research priorities

(1) Foundational validation studies: Conduct prospective observational studies in forensic settings to establish whether digital phenotyping patterns observed in community psychiatric populations generalize to secure environments, examining whether sensor-derived markers maintain predictive validity under conditions of involuntary detention and institutional routine.
(2) Algorithm development and validation: Develop and rigorously validate algorithms using multi-site forensic datasets with external validation, temporal validation (testing whether models trained on historical data perform adequately on future cohorts), and subgroup validation across demographic and clinical

Table 3
Governance framework and safeguards for forensic mental health digital twins.

Governance Domain	Key Risks	Essential Safeguards	Oversight Mechanisms	Responsibility	Practical Implementation Challenges
Human Rights Compliance[13]	Excessive restriction; loss of dignity; procedural injustice; inadequate voice for patients	Human rights impact assessment pre-deployment; least-restrictive principle embedded in algorithms; patient preference incorporation; contestable recommendations; independent advocacy involvement	Human rights committee review; patient advocate scrutiny; tribunal examination of twin role	Service providers; clinical teams; ethics committees; patient advocates	Operationalizing "least restrictive" in algorithms difficult; balancing patient voice with safety mandates; ensuring advocacy representatives have technical literacy to meaningfully review implementations
Data Protection and Privacy[106]	Unauthorized access; data breaches; purpose creep; inadequate consent; third-party sharing	Encryption; access controls; audit trails; purpose limitation; data minimization; transparent privacy notices; lawful basis documentation; regular security testing	Data protection officer oversight; Information Commissioner's Office regulation; security audits; breach notification protocols	Data controllers; information governance teams; IT security	Tension between comprehensive monitoring for clinical benefit and data minimization principles; managing data sharing across organizational boundaries; genuine consent difficult in coercive settings
Algorithmic Fairness[107]	Amplified existing biases; disparate impact on protected groups; inequitable outcomes; discriminatory surveillance	Bias audit in development; protected characteristic stratification; counterfactual fairness testing; continuous disaggregated monitoring; corrective recalibration; equity expertise in governance	Fairness audit committee; equality impact assessments; community consultation; whistleblower protections	Algorithm developers; clinical safety teams; equality specialists	Defining fairness metrics in contexts with existing structural inequalities; obtaining representative data for subgroup validation; genuine community engagement resource-intensive; resistance to external oversight
Transparency and Explainability[108]	Black box opacity; inability to contest; clinician confusion; tribunal rejection; patient disempowerment	Inherently interpretable models preferred where feasible; post-hoc explanation methods for complex models; model cards; documentation standards; training for clinicians; multi-level explanations; uncertainty communication; open validation data where feasible	Ethics committee review; legal scrutiny; patient information testing; tribunal feedback	Developers; implementation teams; medical education; legal departments	Trade-offs between interpretability and performance; explanations may be incomplete or misleading; training curricula need development; legal standards for adequacy undefined; resource constraints for comprehensive documentation
Clinical Safety[13]	Incorrect predictions; adverse events; over-reliance; deskilling; accountability gaps	Rigorous validation (internal, external, temporal); safety case documentation; performance monitoring; incident reporting; drift detection; recalibration protocols; human-in-the-loop design	Clinical safety officer; pharmacovigilance-style monitoring; serious incident reviews; regulatory inspections	Clinical governance; medical directors; commissioners; regulators	Validation requires large datasets often unavailable in forensic settings; defining safety thresholds difficult with low base rates; incident attribution challenging in complex systems; regulatory frameworks immature
Lifecycle Management[109]	Model degradation; scope creep; obsolescence; inadequate updating; loss of institutional knowledge	Version control; change management; retirement policies; documentation archiving; sunset clauses; handover protocols	Technical governance board; periodic reviews; commissioning oversight	IT teams; clinical leadership; service continuity planning	Maintaining expertise as staff turnover; funding for ongoing monitoring and updates uncertain; resistance to retiring deployed systems; lack of standards for when recalibration vs replacement required

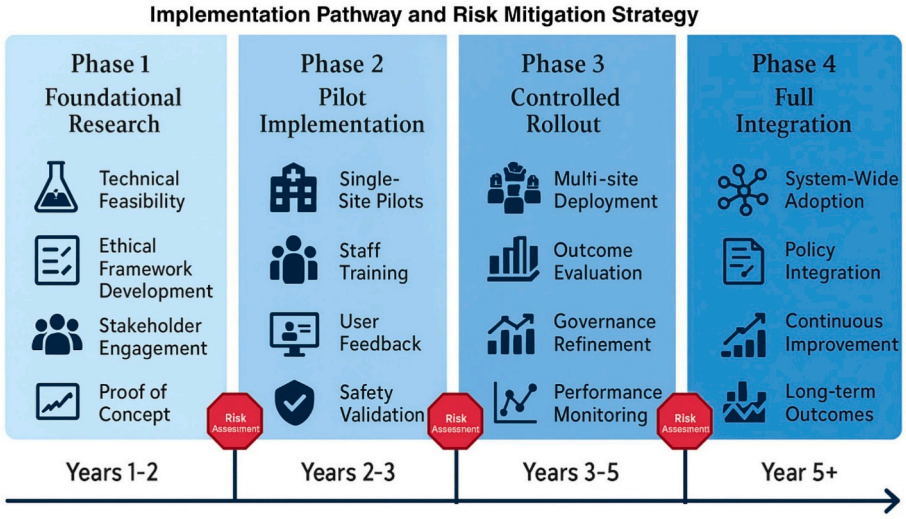

Fig. 2. Staged implementation pathway for digital twin technology in forensic mental health with integrated risk mitigation. The translational pathway progresses through four sequential phases over a 5+ year timeline: foundational research establishes technical feasibility and ethical frameworks; pilot implementation tests single-site deployments with intensive monitoring; controlled rollout expands to multiple sites with outcome evaluation; full integration achieves system-wide adoption with policy integration. Risk mitigation checkpoints between phases ensure safety validation, while parallel governance activities (ethical oversight, continuous validation, stakeholder engagement) provide ongoing safeguards.

characteristics. Validation should prioritize transparency, with data and code made openly available where ethically permissible.

(3) Bias audits and fairness research: Conduct comprehensive bias audits examining whether digital twin recommendations perpetuate or amplify existing inequalities in forensic mental health, testing multiple fairness definitions (demographic parity, equalized odds, predictive parity) and engaging affected communities in defining acceptable fairness criteria.

(4) Implementation science: Study organizational factors affecting successful adoption, including clinician attitudes, workflow integration, training needs, and barriers to appropriate use. Examine how digital twins affect clinical decision-making processes, therapeutic relationships, and organizational cultures in forensic settings.

(5) Participatory design research: Engage service users, carers, and advocacy representatives in co-designing digital twin systems, interfaces, and governance structures. Qualitative research should explore patient perspectives on monitoring, algorithmic assessment, and acceptable trade-offs between privacy and purported safety or therapeutic benefits.

(6) Long-term outcomes research: Evaluate whether digital twin implementations achieve intended benefits (reduced restrictive practices, improved safety, enhanced discharge planning) and monitor for unintended consequences (surveillance creep, deskilling, erosion of therapeutic relationships, rights impacts).

(7) Economic evaluation: Conduct cost-effectiveness and cost-benefit analyses comparing digital twin approaches to enhanced conventional methods, accounting for implementation costs, maintenance, governance overhead, and opportunity costs of alternative investments.

8.2. Policy and governance recommendations

(1) Regulatory clarity: Policymakers should establish clear regulatory frameworks for algorithmic decision support in forensic mental health, specifying approval processes, safety standards, ongoing monitoring requirements, and enforcement mechanisms. This should include guidance on when algorithmic tools require regulatory approval versus falling under existing medical device or software regulations.

(2) Legal reform: Legislatures should consider whether existing mental health and data protection legislation adequately addresses algorithmic tools in forensic contexts, potentially requiring amendments to clarify consent frameworks, establish rights to explanation, and define accountability for algorithmic recommendations.

(3) Standards development: Professional bodies, standards organizations, and multi-stakeholder groups should develop technical and ethical standards for forensic mental health digital twins, including data quality requirements, validation protocols, explainability standards, fairness metrics, and governance structures.

(4) Ethics guidance: Ethics committees, institutional review boards, and research ethics organizations should develop specialized guidance for reviewing digital twin research and implementations in forensic mental health, addressing challenges unique to this domain including consent in coercive settings, justice considerations, and long-term data retention.

(5) Pilot programs with robust evaluation: Jurisdictions considering digital twin implementations should begin with carefully designed pilot programs in single sites with extensive monitoring, independent evaluation, patient and staff feedback mechanisms, and clearly defined stopping rules if safety or rights concerns emerge. Pilots should be time-limited with decisions about continuation based on demonstrated evidence of benefit and acceptable risk.

(6) Investment in non-technological improvements: Policymakers should ensure that enthusiasm for digital innovation does not divert attention and resources from proven non-technological improvements including adequate staffing, staff training in trauma-informed care, environmental design promoting dignity and comfort, and expansion of community forensic services offering less restrictive alternatives.

(7) Moratorium on specific applications: Given current evidence limitations and ethical concerns, we recommend that certain applications should not be pursued until fundamental questions are resolved: • Long-term outcome prediction (>6 months) for discharge decisions, given inherent uncertainty and undemonstrated validity

- Fully automated decision-making without human override, given accountability concerns and legal requirements for human judgment
- Deployment in jurisdictions lacking independent tribunal review or legal advocacy for detained patients, given inadequate safeguards against misuse
- Use of digital twins primarily for security or custodial purposes rather than therapeutic benefit, given incompatibility with healthcare ethics and potential for coercion

8.3. Stakeholder engagement recommendations

(1) Service user involvement: Ensure meaningful participation of current and former forensic mental health service users in all stages of digital twin development, implementation, and governance. This should include paid advisory roles, representation on oversight committees, involvement in design decisions, and power to halt implementations that violate agreed principles.
(2) Workforce engagement: Involve clinicians, nurses, psychologists, social workers, and other frontline staff in defining needs, design requirements, and acceptable implementations. Address workforce concerns about deskilling, liability, and impacts on therapeutic relationships proactively rather than treating resistance as obstruction.
(3) Public dialogue: Given forensic mental health's intersection with public safety, conduct transparent public dialogue about digital twin technology including its promises, limitations, risks, and governance. Public understanding and trust are prerequisites for sustainable implementation.
(4) International collaboration: Foster international research collaborations and knowledge sharing while recognizing jurisdictional differences requiring local adaptation. Learn from implementations across different systems while avoiding uncritical transfer of approaches that may not suit different legal, cultural, or organizational contexts.

9. Conclusion

Digital twin technology represents a potentially transformative innovation for forensic mental health, offering pathways to move beyond episodic, static assessments towards continuously updated, personalized, explainable decision support. The vision encompasses person-level twins that refine risk assessment and optimize treatments; ward-level twins that improve therapeutic environments and reduce restrictive practices; and system-level twins that guide policy and resource allocation. Enabling technologies including digital phenotyping, wearable sensors, and advanced analytics have matured to a point where implementation appears technically feasible.

However, the translation of digital twins from concept to clinical reality in forensic mental health must proceed with exceptional care and humility about current limitations. The populations served are among the most vulnerable in healthcare, often detained involuntarily under conditions that restrict liberty and autonomy. The decisions supported by digital twins carry profound consequences for both public safety and individual rights. The history of technology in criminal justice includes cautionary tales of systems that entrenched bias, eroded due process, and prioritized efficiency over humanity. We must learn from these failures.

Current evidence does not support widespread deployment of digital twins in forensic mental health. The technology remains immature, with no validated implementations, substantial gaps in foundational research, unresolved ethical and legal questions, and uncertain effects on the outcomes that matter most to service users. Enthusiasm for innovation must be tempered by rigorous skepticism, especially when the stakes involve liberty, dignity, and fundamental rights.

The pathway forward demands: rigorous validation establishing that digital twins improve outcomes across multiple dimensions including safety, rights protection, equity, and recovery; robust governmental frameworks that ensure transparency, contestability, and accountability; sustained co-design with service users ensuring that implementations respect dignity and support rather than constrain recovery; regulatory oversight that keeps pace with technological change while protecting fundamental rights; and workforce development that equips clinicians to use sophisticated tools wisely, retaining professional judgement and therapeutic relationships that technology cannot replace.

Critically, there are contexts in which digital twins should not be deployed regardless of technical sophistication: jurisdictions lacking independent legal review of detention decisions; settings where patients have no effective advocacy or legal representation; implementations prioritizing institutional efficiency or security over therapeutic benefit; systems failing to demonstrate measurable reduction in restrictiveness; and environments where governance structures are inadequate to prevent misuse, bias, or coercion. Acknowledging these boundaries is as important as articulating the technology's promise.

If appropriate conditions are met, digital twins could help forensic mental health services deliver care that is simultaneously safer and more humane. They could enable earlier detection of deterioration, more targeted interventions, reduced unnecessary restrictions, and transparent decision-making that enhances rather than undermines trust. They could support the challenging work of clinicians navigating the interface of health, justice, and public protection. Most importantly, they could improve lives, both by reducing preventable harm and by promoting recovery, reintegration, and respect for human dignity.

However, they could equally entrench surveillance, rationalize continued detention, amplify existing inequalities, undermine therapeutic relationships, and create an illusion of evidence-based decision-making while obscuring value judgments and political choices behind a veneer of algorithmic objectivity. Which outcome prevails will depend not on the technology itself but on the governance, values, and power structures within which it is embedded.

The task ahead is substantial but worthwhile. As forensic mental health services face increasing demands, complexity, and scrutiny, innovations that enhance quality while upholding values become essential. Digital twins, developed and deployed with wisdom, care, and humility, and potentially not at all if foundational questions cannot be satisfactorily resolved, may prove to be such an innovation. The current moment, with emerging technical capabilities and growing recognition of digital health's potential, offers an opportunity to shape this future deliberately and well. We must seize it thoughtfully, with clear-eyed recognition of both opportunities and dangers, and with unwavering commitment to the principle that technological innovation in forensic mental health must serve human flourishing and human rights rather than merely institutional efficiency or risk minimization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Meynen G. Walls and laws: structural barriers to forensic psychiatric research. *Eur Psychiatry*. 2017;44:208–209. https://doi.org/10.1016/J.EURPSY.2017.04.010.
2. Rutherford M, Duggan S. Forensic mental health services: facts and figures on current provision. *Br J Forensic Pract*. 2008;10:4–10. https://doi.org/10.1108/14636646200800020.
3. Pedersen SH, Radovic S, Nilsson T, Eriksson L. Dual-roles and beyond: values, ethics, and practices in forensic mental health decision-making. *Med Health Care Philos*. 2025;28:199–211. https://doi.org/10.1007/S11019-024-10247-2.
4. Douglas KS, Yeomans M, Boer DP. Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Crim Justice Behav*. 2005;32:479–510. https://doi.org/10.1177/0093854805278411.

5. Coid JW, Yang M, Ullrich S, et al. Most items in structured risk assessment instruments do not predict violence. *J Forensic Psychiatr Psychol*. 2011;22:3–21. https://doi.org/10.1080/14789949.2010.495990.
6. Hogan NR, Olver ME. A prospective examination of the predictive validity of five structured instruments for inpatient violence in a secure forensic hospital. *Int J Forensic Ment Health*. 2018;17:122–132. https://doi.org/10.1080/14999013.2018.1431339.
7. Hogan NR, Olver ME. Static and dynamic assessment of violence risk among discharged forensic patients. *Crim Justice Behav*. 2019;46:923–938. https://doi.org/10.1177/0093854819846526.
8. Rangavittal PB. Evolving role of AI in enhancing patient care within digital health platforms. *J Artif Intell Cloud Comput*. 2022:1–6, 10.47363/JAICC/2022(1)241.
9. Erol T, Mendi AF, Dogan D. *The digital twin revolution in healthcare. 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. 2020. https://doi.org/10.1109/ISMSIT50672.2020.9255249.
10. Iliuță ME, Moisescu MA, Pop E, Ionita AD, Caramihai SI, Mitulescu TC. Digital twin—A review of the evolution from concept to technology and its analytical perspectives on applications in various fields. *Appl Sci*. 2024;14. https://doi.org/10.3390/APP14135454.
11. Cellina M, Cè M, Alì M, et al. Digital twins: the new frontier for personalized medicine? *Appl Sci*. 2023;13. https://doi.org/10.3390/APP13137940.
12. Chen S, Zhu P. Digital twin technology and applications in the medical field: bridging the virtual for precision medicine. In: *2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI)*. 2024:221–226. https://doi.org/10.1109/DTPI61353.2024.10778770.
13. Alimour SA, Alrabeei M. A novel model for digital twins in mental health: the biopsychosocial AI-Driven Digital Twin (BADT) framework. *Swiss Conference on Data Science*. 2024:6–10. https://doi.org/10.1109/SDS64317.2024.10883917.
14. Bernardos AM, Pires M, Ollé D, Casar JR. *Digital phenotyping as a tool for personalized mental healthcare. International Conference on Pervasive Computing Technologies for Healthcare*. 2019:403–408. https://doi.org/10.1145/3329189.3329240.
15. Sheikh M, Qassem M, Kyriacou PA. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Front Digit Health*. 2021;3. https://doi.org/10.3389/FDGTH.2021.662811.
16. Yang M, Ngai ECH, Hu X, et al. Digital phenotyping and feature extraction on smartphone data for depression detection. *Proc IEEE*. 2024;112:1773–1798. https://doi.org/10.1109/JPROC.2025.3542324.
17. Merchant R, Torous J, Rodriguez-Villa E, Naslund JA. Digital technology for management of severe mental disorders in low-income and middle-income countries. *Curr Opin Psychiatr*. 2020;33:501–507. https://doi.org/10.1097/YCO.0000000000000626.
18. Marsch LA. Opportunities and needs in digital phenotyping. *Neuropsychopharmacology*. 2018;43:1637–1638. https://doi.org/10.1038/S41386-018-0051-7.
19. Mulder T, Jagesar RR, Klingenberg AM, P, Mifsud Bonnici J, Kas MJ. New European privacy regulation: assessing the impact for digital medicine innovations. *Eur Psychiatry*. 2018;54:57–58. https://doi.org/10.1016/J.EURPSY.2018.07.003.
20. Spitzer M, Dattner I, Zilcha-Mano S. Digital twins and the future of precision mental health. *Front Psychiatr*. 2023;14. https://doi.org/10.3389/FPSYT.2023.1082598/PDF.
21. De Kerckhove D. The personal digital twin, ethical considerations. *Phil Trans Roy Soc A*. 2021;379. https://doi.org/10.1098/RSTA.2020.0367.
22. Mandischer N, Atanasyan A, Schluse M, Robmann J, Mikelsons L. Perspectives-observer-transparency – a novel paradigm for modelling the human in human-to-anything interaction based on a structured review of the human digital twin. *IEEE Int Conf Syst Man Cybern*. 2024:215–222. https://doi.org/10.1109/SMC54092.2024.10831941.
23. Sieve R, Kobialka P, Slaughter L, Schlatte R, Johnsen EB, Tarifa SLT. BedreFlyt: improving patient flows through hospital wards with digital twins. *Electron Proc Theoret Comput Sci EPTCS*. 2025;418:1–15. https://doi.org/10.4204/EPTCS.418.1.
24. Slade K, Samele C, Valmaggia L, Forrester A. Pathways through the criminal justice system for prisoners with acute and serious mental illness. *J Forensic Leg Med*. 2016;44:162–168. https://doi.org/10.1016/J.JFLM.2016.10.007.
25. Almirall E, Callegaro D, Bruins P, Santamaría M, Martínez P, Cortés U. Deep air – a smart city AI synthetic ta digital twin living the alability ta oblems. *Front Artif Intell Appl*. 2022;356:83–86. https://doi.org/10.3233/FAIA220319.
26. Smirnov AV. "Digital Twin" of the arctic population in demographic research and territorial development management. *Arctic and North*. 2023:260–272. https://doi.org/10.37482/ISSN2221-2698.2023.53.260.
27. Riordan S, Haque S, Humphreys M. Possible predictors of outcome for conditionally discharged patients – a preliminary study. *Med Sci Law*. 2006;46:31–36. https://doi.org/10.1258/RSMMSL.46.1.31.
28. Jewell A, Cocks C, Cullen AE, Fahy T, Dean K. Predicting time to recall in patients conditionally released from a secure forensic hospital: a survival analysis. *Eur Psychiatry*. 2018;49:1–8. https://doi.org/10.1016/J.EURPSY.2017.11.005.
29. Pichot G, Rocheteau J, Attiogbé C. Model-driven development of digital twins for supervision and simulation of sensor-and-actuator networks (extended abstract). *STAF Workshops*. 2022.
30. Fagherazzi G. Deep digital phenotyping and digital twins for precision health: time to dig deeper. *J Med Internet Res*. 2020;22. https://doi.org/10.2196/16770.
31. De Boer C, Ghomrawi H, Zeineddin S, Linton S, Kwon S, Abdullah F. A call to expand the scope of digital phenotyping. *J Med Internet Res*. 2022;25. https://doi.org/10.2196/39546.
32. Torous J, Powell AC. Current research and trends in the use of smartphone applications for mood disorders. *Internet Interv*. 2015;2:169–173. https://doi.org/10.1016/J.INVENT.2015.03.002.
33. Braund TA, Zin MT, Boonstra TW, et al. Smartphone sensor data for identifying and monitoring symptoms of mood disorders: a longitudinal observational study. *JMIR Ment Health*. 2021;9. https://doi.org/10.2196/35549.
34. Lyall LM, Wyse CA, Graham N, et al. Association of disrupted circadian rhythmicity with mood disorders, subjective wellbeing, and cognitive function: a cross-sectional study of 91 105 participants from the UK biobank. *Lancet Psychiatry*. 2018;5:507–514. https://doi.org/10.1016/S2215-0366(18)30139-1.
35. Lee TY, Chen CH, Chen IM, et al. Dynamic bidirectional associations between global positioning system mobility and ecological momentary assessment of mood symptoms in mood disorders: prospective cohort study. *J Med Internet Res*. 2023;26. https://doi.org/10.2196/55635.
36. Fudolig MID, Monsivais D, Bhattacharya K, Jo HH, Kaski K. Different patterns of social closeness observed in mobile phone communication. *J Comput Soc Sci*. 2018;3. https://doi.org/10.1007/S42001-019-00054-8.
37. Gray L, Marcynikola N, Barnett I, Torous J. The potential for digital phenotyping in understanding mindfulness app engagement patterns: a pilot study. *J Integr Compl Med*. 2024;30:1108–1115. https://doi.org/10.1089/JICM.2023.0698.
38. Birenboim A, Dijst M, Scheepers FE, Poelman MP, Helbich M. Wearables and location tracking technologies for mental-state sensing in outdoor environments. *Prof Geogr*. 2019;71:449–461. https://doi.org/10.1080/00330124.2018.1547978.
39. De Vries-Bouw M, Popma A, Vermeiren R, Doreleijers TAH, Van De Ven PM, Jansen LMC. The predictive value of low heart rate and heart rate variability during stress for reoffending in delinquent male adolescents. *Psychophysiology*. 2011;48(11):1597–1604. https://doi.org/10.1111/J.1469-8986.2011.01233.X.
40. Jang EH, Kim AY, Yu HY. Relationships of psychological factors to stress and heart rate variability as stress responses induced by cognitive stressors. *Kor Soc Emotion Sensibility*. 2018;21:71–82. https://doi.org/10.14695/KJSOS.2018.21.1.71.
41. Clamor A, Ludwig L, Lincoln TM. Heart rate variability as an index of emotion (dys)regulation in psychosis? *Int J Psychophysiol*. 2020;158:310–317. https://doi.org/10.1016/J.IJPSYCHO.2020.08.016.
42. Ho FYY, Poon CY, Wong VWH, et al. Actigraphic monitoring of sleep and circadian rest-activity rhythm in individuals with major depressive disorder or depressive symptoms: a meta-analysis. *J Affect Disord*. 2024;361:224–244. https://doi.org/10.1016/J.JAD.2024.05.155.
43. McLaughlin P, Nikkhah A, Umer Waqar M, Kennedy HG, Davoren M. Change in quality of life after moving a national forensic mental health service: a dundrum forensic redevelopment evaluation study (D-FOREST). *BJPsych Open*. 2024;10. https://doi.org/10.1192/BJO.2024.206. S62–S62.
44. Meyer E. *Cabrillo National Monument: Acoustic Monitoring Report*. 2021. https://doi.org/10.36967/2303446.
45. Medley DB, Morris JE, Stone CK, Song J, Delmas T, Thakrar K. An association between occupancy rates in the emergency department and rates of violence toward staff. *J Emerg Med*. 2012;43:736–744. https://doi.org/10.1016/J.JEMERMED.2011.06.131.
46. Gubin DG, Borisenkov MF, Kolomeichuk SN, et al. Evaluating circadian light hygiene: methodology and health implications. *Russ Open Med J*. 2024;13. https://doi.org/10.15275/RUSOMJ.2024.0415.
47. Dai H, Imani S, Choi JH. Correlating indoor environmental quality parameters with human physiological responses for adaptive comfort control in commercial buildings. *Energies (Basel)*. 2025;18. https://doi.org/10.3390/EN18092280.
48. Cuzzocrea A, Benlaredj I. An innovative big data framework for supporting multidimensional risk analysis and prediction over digital twins. *BigData Congress [Services Society]*. 2024:6214–6222. https://doi.org/10.1109/BIGDATA62323.2024.10825260.
49. Munthe C, Radovic S, AnckarsÄter H. Ethical issues in forensic psychiatric research on mentally disordered offenders. *Bioethics*. 2010;24:35–44. https://doi.org/10.1111/J.1467-8519.2009.01773.X.
50. Verbeke E, Vanheule S, Cauwe J, Truijens F, Froyen B. Coercion and power in psychiatry: a qualitative study with ex-patients. *Soc Sci Med*. 2019;223:89–96. https://doi.org/10.1016/J.SOCSCIMED.2019.01.031 (1967).
51. Seneviratne O, Kagal L. *Enabling privacy through transparency. 2014 Twelfth Annual International Conference on Privacy, Security and Trust*. 2014:121–128. https://doi.org/10.1109/PST.2014.6890931.
52. Zio E, Miqueles L. Digital twins in safety analysis, risk assessment and emergency management. *Reliab Eng Syst Saf*. 2024;246. https://doi.org/10.1016/J.RESS.2024.110040.
53. Falzer PR. Valuing structured professional judgment: predictive validity, decision-making, and the clinical-actuarial conflict. *Behav Sci Law*. 2013;31(1):40–54. https://doi.org/10.1002/BSL.2043.
54. Hilton S, Langton J, Conroy P, Stecki C. *Digital availability twin – targeted risk mitigation from design to operation. 2023 Annual Reliability and Maintainability Symposium (RAMS) 2023*. 2023-January:1–6. https://doi.org/10.1109/RAMS51473.2023.10088191.
55. Klenz B. How to use streaming analytics to create a real-time Digital Twin. Proceedings of the SAS Global Forum 2018 Conference. Cary, NC: SAS Institute Inc; 2004–2018.
56. Kim SK, Lee DB. Psychopharmacological treatment patterns in patients with schizophrenia and schizoaffective disorder in forensic inpatient settings. *Korean J Leg Med*. 2017;41:115–121. https://doi.org/10.7580/KJLM.2017.41.4.115.
57. Young SL, Taylor M, Lawrie SM. "First do no harm." A systematic review of the prevalence and management of antipsychotic adverse effects. *J Psychopharmacol*. 2015;29:353–362. https://doi.org/10.1177/0269881114562090.

58. Gurrera RJ, Gearin PF, Love J, et al. Recognition and management of clozapine adverse effects: a systematic review and qualitative synthesis. *Acta Psychiatr. Scand.* 2022 May;145(5):423–441. https://doi.org/10.1111/acps.13406.
59. Kirby M, Asan O. Digital twin implementation in the pharmaceutical industry. *Proc Int Symp Human Factors Ergonom Health Care*. 2024;13:151–156. https://doi.org/10.1177/2327857924131002.
60. Gatner DT, Moulden HM, Mamak M, Chaimowitz GA. At risk of what? Understanding forensic psychiatric inpatient aggression through a violence risk scenario planning lens. *Int J Forensic Ment Health*. 2021;20:398–407. https://doi.org/10.1080/14999013.2021.1899343.
61. Graves J, Garbett S, Zhou Z, Schildcrout JS, Peterson J. Comparison of decision modeling approaches for health technology and policy evaluation. *Med Decis Mak*. 2021;41:453. https://doi.org/10.1177/0272989X21995805.
62. Scott HV, Gillespie M. Restrictive measures in forensic mental health and their role in recovery: a narrative literature review. *British J Ment Health Nurs*. 2023;12:1–8. https://doi.org/10.12968/BJMH.2022.0016.
63. Milne-Ives M, Fraser LK, Khan A, et al. Life course digital twins–intelligent monitoring for early and continuous intervention and prevention (LifeTIME): Proposal for a retrospective cohort study. *JMIR Res Protoc*. 2021;11. https://doi.org/10.2196/35738.
64. Goulet MH, Lessard-Deschênes C. The model of prevention of seclusion and restraint use in mental health: an integrative review. *Santè mentale au Quèbec*. 2022 Jan 1;47(1):151–180.
65. Schimpf C, Barbrook-Johnson P, Castellani B. Cased-based modelling and scenario simulation for ex-post evaluation. *Evaluation*. 2021;27:116–137. https://doi.org/10.1177/1356389020978490.
66. Barbic SP, Chan N, Rangi A, et al. Health provider and service-user experiences of sensory modulation rooms in an acute inpatient psychiatry setting. *PLoS One*. 2019;14. https://doi.org/10.1371/JOURNAL.PONE.0225238.
67. Grundy AC, Papastravrou Brooks C, Johnston I, Cree L, Callaghan P, Price O. Evaluation of a novel co-designed and co-delivered training package to de-escalate violence and aggression in UK acute inpatient, PICU and forensic mental health settings. *J Psychiatr Ment Health Nurs*. 2024;31:1145–1154. https://doi.org/10.1111/JPM.13074.
68. Muheizen JL. The effect of risk assessment data and colleagues' consensus on clinical decisions to discharge insanity acquittees -Escholarship. https://escholarship.org/uc/item/83m9n3hx.
69. Hou GY, Lal A, Schulte PJ, et al. Informing intensive care unit digital twins: dynamic assessment of cardiorespiratory failure trajectories in patients with sepsis. *Shock*. 2025;63:573–578. https://doi.org/10.1097/SHK.0000000000002536.
70. Coffey M. A risk worth taking? Value differences and alternative risk constructions in accounts given by patients and their community workers following conditional discharge from forensic mental health services. *Health Risk Soc*. 2012;14:465–482. https://doi.org/10.1080/13698575.2012.682976.
71. Dou M, Chen J, Chen D, et al. Modeling and simulation for natural disaster contingency planning driven by high-resolution remote sensing images. *Future Gener Comput Syst*. 2014;37:367–377. https://doi.org/10.1016/J.FUTURE.2013.12.018.
72. Ainsworth SA, Taxman FS. Creating simulation parameter inputs with existing data sources: estimating offender risks, needs, and recidivism. *Simulation Strategies to Reduce Recidivism*. 2013:115–142. https://doi.org/10.1007/978-1-4614-6188-3_5.
73. Sheridan MS. Limits of discharge-planning. *Soc Work*. 1981;26. https://doi.org/10.1093/SW/26.2.179-A, 179–179.
74. Yang Z, Heaukulani C, Sim A, et al. Utility of digital phenotyping based on wrist wearables and smartphones in psychosis: observational study. *JMIR mHealth uHealth*. 2025;13. https://doi.org/10.2196/56185.
75. Spivak BL, Shepherd SM. Machine learning and forensic risk assessment: new frontiers. *J Forensic Psychiatr Psychol*. 2020;31:571–581. https://doi.org/10.1080/14789949.2020.1779783.
76. Sachowski J. Accomplishing forensic readiness. *Implement Digital Forensic Readiness*. 2016:151–153. https://doi.org/10.1016/B978-0-12-804454-4.00015-0.
77. Winter PD, Chico TJA. Using the non-adoption, abandonment, scale-up, spread, and sustainability (NASSS) framework to identify barriers and facilitators for the implementation of digital twins in cardiovascular medicine. *Sensors (Basel)*. 2023;23. https://doi.org/10.3390/S23146333.
78. Lehman WEK, Greener JM, Rowan-Szal GA, Flynn PM. Organizational readiness for change in correctional and community substance abuse programs. *J Offender Rehabil*. 2012;51:114–196. https://doi.org/10.1080/10509674.2012.633022.
79. Kennedy HG, Mullaney R, McKenna P, et al. A tool to evaluate proportionality and necessity in the use of restrictive practices in forensic mental health settings: the DRILL tool (Dundrum restriction, intrusion and liberty ladders). *BMC Psychiatr*. 2020 Oct 23;20(1):515. https://doi.org/10.1186/s12888-020-02912-6.
80. Ermolina LV, Zinovyev AM, Melnikova DA. Digital twins as a method of risk management transformation. *Digital Technologies in the New Socio-Economic Reality*. 2022;304:451–457. https://doi.org/10.1007/978-3-030-83175-2_56.
81. Appelbaum PS. Privacy in psychiatric treatment: threats and responses. *Focus*. 2003;1:396–406. https://doi.org/10.1176/FOC.1.4.396.
82. Carey P. *Data Protection: A Practical Guide to UK and EU Law*; 2004. https://dl.acm.org/doi/abs/10.5555/3265270.
83. Halinkina VS. Basic principles of personal data processing and protection. *Uzhhorod National University Herald Series: Law*. 2024;1:111–115. https://doi.org/10.24144/2307-3322.2024.81.1.17.
84. Taylor M, Kirkham RL. Health data, public interest, and surveillance for non-health-related purposes. *Ethical Issues in Covert, Security Surveillance Research*. 2021:93–118. https://doi.org/10.1108/S2398-601820210000008008.
85. Verhenneman G. *Protecting the patient through purpose limitation. The patient. Data Protection and Changing Healthcare Models*. 2021:241–338. https://doi.org/10.1017/9781839701252.013.
86. Iwaya LH, Babar MA, Rashid A, Wijayarathna C. On the privacy of mental health apps: An empirical investigation and its implications for app development. *Empir. Softw. Eng.* 2023 Jan;28(1):2. https://doi.org/10.1007/s10664-022-10236-0.
87. Zagorski N. Mental health apps miss the mark on usability standards. *Study Shows Psychiatr News*. 2019;54. https://doi.org/10.1176/APPI.PN.2019.5A21.
88. Bell CJ, Celnik M, Devgun J, Hansen O, Osborn W, Faiz G. Providing assurance of digital twins. In: *SPE Offshore Europe Conference and Exhibition*. SPE; 2023 Sep 5. D021S006R001 https://doi.org/10.2118/215599-MS.
89. Allen C, Des Jardins TR, Heider A, et al. Data governance and data sharing agreements for community-wide health information exchange: lessons from the beacon communities. *EGEMs*. 2014;2:5. https://doi.org/10.13063/2327-9214.1057.
90. Coleman M. Reflections on systemic barriers for ethnic minorities in accessing community-based forensic services for people with intellectual disabilities and autism. *J Intellect Disabil Offending Behav*. 2021;13:12–19. https://doi.org/10.1108/JIDOB-08-2021-0012.
91. Weinberger N, Hery D, Mahr D, Adler SO, Stadlbauer J, Ahrens TD. Beyond the gender data gap: co-creating equitable digital patient twins. *Front Digit Health*. 2025;7. https://doi.org/10.3389/FDGTH.2025.1584415.
92. Saleiro P, Kuester B, Hinkson L, Aequitas, et al. A bias and fairness audit. *Toolkit*. 2019. https://doi.org/10.48550/arXiv.1811.05577.
93. Maughan K, Ngong IC, Near JP. Prediction sensitivity: continual audit of counterfactual fairness in deployed classifiers. *ArXivOrg*. 2022. https://doi.org/10.48550/arXiv.2202.04504.
94. Mustafa AB, Zafar U. Integrating forensic mental health in legal processes. *J Pak Psychiatr Soc*. 2024;21. https://doi.org/10.63050/JPPS.21.02.331.
95. Mwebe H. Giving evidence before the first-tier tribunal (Mental Health): the role of inpatient mental health nurses. *British J Ment Health Nurs*. 2023;12:1–4. https://doi.org/10.12968/BJMH.2023.0012.
96. Chasse K. Challenging electronic Systems' and Devices' ability to produce reliable evidence. *SSRN Electron J*. 2019. https://doi.org/10.2139/SSRN.3378077.
97. Castro YVC de, Morii KY, Santos IA, Destefani AC, Destefani VC. Beyond black boxes: interpretable ai for enhanced risk assessment and ethical decision-making in forensic psychiatry. *Revista Ibero-Americana de Humanidades, Ciências e Educação*. 2024;10:2475–2479. https://doi.org/10.51891/REASE.V10I8.15298.
98. Wang Y. A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Res Rep Comput Sci*. 2024;3(2):25–33. https://doi.org/10.37256/RRCS.3220244750.
99. Gursoy F, Kakadiaris I. System cards for AI-Based decision-making for public policy. ArXiv n.d.;abs/2203.04754. https://doi.org/10.48550/ARXIV.2203.04754.
100. Hunink MGM. In search of tools to aid logical thinking and communicating about medical decision making. *Med Decis Mak*. 2001;21:267–277. https://doi.org/10.1177/0272989X0102100402.
101. Kim T-W, Park S-J. Governance and accountability - a shift in CONCEPTUALISATION.1. *Public Adm Q*. 2019;31:1–31. https://doi.org/10.21888/KPAQ.2019.3.31.1.001.
102. Vanderhorn E, Valluri S. Guidance on the verification and validation of digital twins. In: *SNAME Offshore Symposium*. SNAME; 2024 Feb 20. D011S004R001 https://doi.org/10.5957/TOS-2024-004.
103. Cappon G, Pellizzari E, Cossu L, et al. *System architecture of TWIN: a new digital twin-based clinical decision support system for type 1 diabetes management in children. IEEE 19th International Conference on Body Sensor Networks (BSN) 2023*. 2023:1–4. https://doi.org/10.1109/BSN58485.2023.10331272.
104. Skrynnyk O. Towards organizational development in digital organizational twin. *SocioEconomic Challenges 2021*. 2021;5(3):126–133. https://doi.org/10.21272/SEC, 5.
105. Buhl MD, Sett G, Koessler L, Schuett J, Anderljung M. Safety cases for frontier AI. *ArXivOrg*. 2024. https://doi.org/10.48550/ARXIV.2410.21572.
106. Lomotey RK, Kumi S, Ray M, Deters R. *Synthetic data digital twins and data trusts control for privacy in health data sharing. Proceedings of the 2024 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*. 2024:1–10. https://doi.org/10.1145/3643650.3658605.
107. Ryan S, Doherty G. *Fairness Definitions for Digital Mental Health Applications*. 2021.
108. Vorras A, Mitrou L. Unboxing the black box of artificial intelligence: algorithmic transparency and/or a right to functional explainability. *EU Internet Law in the Digital Single Market*. 2021:247–264. https://doi.org/10.1007/978-3-030-69583-5_10.
109. Pileggi P, Lazovik E, Broekhuijsen J, Borth M, Verriet J. *Lifecycle governance for effective digital twins: a joint systems engineering and IT perspective. IEEE Systems Conference*. 2020. https://doi.org/10.1109/SYSCON47679.2020.9275662.