Sci

MDPI

Review

Governing Healthcare AI in the Real World: How Fairness, Transparency, and Human Oversight Can Coexist: A Narrative Review

Paolo Bailo [1], Giulio Nittari [2], Giuliano Pesel [1], Emerenziana Basello [3], Tommaso Spasari [4] and Giovanna Ricci [1,*]

[1] Section of Legal Medicine, School of Law, University of Camerino, 62032 Camerino, Italy; paolo.bailo@unicam.it (P.B.); dr.giuliano.pesel@gmail.com (G.P.)
[2] Telemedicine and Telepharmacy Centre, School of Medicinal and Health Products Sciences, University of Camerino, 62032 Camerino, Italy; giulio.nittari@unicam.it
[3] Nursing Degree Course, Department of Medicine, Surgery and Health Sciences, University of Trieste, 34137 Trieste, Italy; emybasello@mac.com
[4] Section of Occupational and Legal Medicine and BioLaw, Niccolò Cusano University, 00166 Rome, Italy
* Correspondence: giovanna.ricci@unicam.it; Tel.: +39-0737-402435

Abstract

Artificial intelligence (AI) is rapidly shifting from experimental pilots to mainstream clinical infrastructure, redefining how evidence, accountability, and ethics intersect in healthcare. This narrative review integrates insights from peer-reviewed studies and policy frameworks to examine seven cross-cutting aspects: bias and fairness, explainability, safety and quality, privacy and data protection, accountability and liability, human oversight, and procurement and deployment. Findings reveal persistent inequities driven by dataset bias and opaque design; the need for explainability tools that are validated, task-specific, and usable by clinicians; and the centrality of post-market surveillance for sustaining patient safety. Privacy-preserving methods such as federated learning and differential privacy show promise but demand rigorous validation and regulatory coherence. Emerging liability models advocate shared enterprise responsibility, while governance-by-design—embedding transparency, auditability, and equity across the AI lifecycle—appears most effective in balancing innovation with public trust. Ethical, legal, and technical safeguards must evolve together to ensure that AI augments, rather than replaces, clinical judgment and institutional accountability.

Keywords: artificial intelligence; bioethics; patient safety; health equity; liability; legal

Academic Editors: Claus Jacob and Huosheng Hu

Received: 9 December 2025
Revised: 29 January 2026
Accepted: 2 February 2026
Published: 6 February 2026

1. Introduction

Since around 2018, the uptake of artificial intelligence (AI) in healthcare has shifted from proof of concept to routine deployment in imaging, triage, documentation, and operational logistics. With that shift came a familiar tension: systems that promise earlier detection, workflow relief, and population-level insight also import hard problems—distributional bias, opacity, safety under dataset shift, privacy leakage, and the diffusion of responsibility across institutions and vendors [1,2]. What distinguishes the present moment is not only model performance, but whether governance instruments—documentation, evaluation, audit, and oversight—travel with algorithms into real clinical contexts, where incentives, staffing, data pipelines, and accountability structures determine how "safe" or "fair" a tool remains after go-live.

Despite the rapidly expanding literature on ethical principles for healthcare AI, a persistent gap remains between high-level commitments and practice-facing governance. Much of the discourse still treats fairness, transparency, and accountability as abstract values rather than as operational properties produced by concrete mechanisms: data governance rules, pre-specified validation and monitoring plans, audit trails, change control for model updates, incident reporting pathways, and procurement clauses that allocate duties across vendors and deployers. In clinical practice, this gap is consequential. Governance failures tend to surface not as philosophical dilemmas but as miscalibrated alerts, subgroup performance cliffs, untracked model updates, opaque interfaces, and uncertainty about who can pause, rollback, or retire a system when harms emerge.

The ethical themes most consistently identified in recent syntheses can now be treated with sufficient coherence to support implementation: fairness as measured and monitored across subgroups; transparency calibrated to clinicians and patients rather than engineers; safety and quality as lifecycle properties; privacy and data protection that remain robust under reuse and repurposing; clear lines of accountability and liability; meaningful human oversight; and procurement and deployment practices that encode obligations rather than aspirations [3]. These domains are not after-the-fact constraints. They shape upstream data collection and labeling, midstream validation and interface design, and downstream model updating, drift detection, and incident response. The decision to focus on these seven domains is therefore practice-driven: together, they capture the principal governance levers that determine whether AI can be deployed and maintained lawfully, safely, and equitably across the clinical lifecycle.

Methodological guardrails have matured in parallel with deployment. Reporting frameworks such as the Consolidated Standards of Reporting Trials-AI (CONSORT-AI) and Standard Protocol Items: Recommendations for Interventional Trials-AI (SPIRIT-AI) extensions require disclosure of data provenance, human–AI interaction, failure modes, and update plans, making external appraisal and replication more realistic for AI-enabled interventions [4,5]. Their core message is straightforward but operationally demanding: evaluation must track the whole sociotechnical system, not only the algorithmic core. In other words, a defensible assessment of clinical AI requires attention to workflows, user behavior, interface design, escalation pathways, and the institutional capacity to monitor and correct performance once the tool is embedded in routine care.

Legal and regulatory systems are simultaneously attempting to keep pace. In the European Union, the 2024 AI Act overlays a horizontal, risk-based regime onto the General Data Protection Regulation (GDPR) and medical device obligations, translating governance expectations into enforceable duties for risk management, data governance, transparency, human oversight, and post-market monitoring [6]. At the same time, the EU-centric approach sits within a wider global landscape in which other jurisdictions pursue partially convergent, partially divergent strategies—an issue of direct relevance to multinational vendors, cross-border data infrastructures, and healthcare institutions adopting tools developed elsewhere.

In the United States, the Food and Drug Administration (FDA) has articulated a regulatory approach for AI-enabled medical devices that explicitly addresses post-market adaptation through predetermined change control plans (PCCPs) and related transparency expectations for machine learning-enabled devices [7,8]. In the United Kingdom, the Medicines and Healthcare products Regulatory Agency's (MHRA) Software and AI as a Medical Device Change Programme similarly frames regulatory reform across the software lifecycle, and practice-facing resources within the National Health Service (NHS) have operationalized procurement and information-governance questions for adopting AI in care settings [9–11]. In Canada, Health Canada has issued pre-market guidance for machine

learning-enabled medical devices, including expectations that connect risk management, evidence, and equity considerations across the lifecycle [12]. Australia's Therapeutic Goods Administration (TGA) has also published guidance addressing the regulation and evidence requirements for medical-device software using AI, reinforcing the practical importance of documentation and ongoing performance assurance [13].

In Asia, Singapore has advanced practical AI governance frameworks emphasizing internal accountability, transparency, and risk management alongside a mature privacy regime, and its Health Sciences Authority (HSA) has issued lifecycle-oriented guidance for software medical devices, including AI-enabled tools [14,15]. Japan has developed dedicated regulatory workstreams for AI-based Software as a Medical Device (SaMD) through its regulator (PMDA), explicitly addressing post-market learning and safety standardization challenges [16]. South Korea's Ministry of Food and Drug Safety (MFDS) has published guidance addressing machine learning–enabled medical devices used to diagnose, manage, or predict disease [17]. China's National Medical Products Administration (NMPA) has also issued guidance on the classification of AI-based medical software products as part of strengthening supervision of the sector [18]. These Asian developments, while not identical in legal form or enforceability, reinforce a common direction of travel: lifecycle governance, documentation, and post-deployment oversight are increasingly treated as core regulatory expectations rather than optional "best practices". Alongside these jurisdiction-specific instruments, cross-cutting governmental and intergovernmental frameworks such as the National Institute of Standards and Technology (NIST) AI Risk Management Framework and WHO guidance on AI ethics and governance further underline the shift from high-level principles to auditable governance mechanisms across the AI lifecycle [19,20].

Meanwhile, the rise in generative models and large language models has intensified familiar concerns—hallucinations, provenance, prompt sensitivity—and introduced new ones around disclosure, scope of automation, and the boundary between clinical decision support and administrative augmentation [21]. In parallel, the most visible clinical value cases remain those tied to risk prediction and early identification, where performance claims and governance duties intersect directly at the bedside. For example, recent work on machine learning-based disease risk prediction systems highlights both the clinical promise of individualized probability estimates and the governance need for usable, task-specific explanations of influential factors to support appropriate reliance by clinicians [22]. This kind of applied literature illustrates why "explainability" cannot be treated as a generic virtue: its clinical function (usability, appropriate reliance, patient communication) and its legal-regulatory function (documentation, contestability, auditability) may overlap but are not interchangeable.

Further work in regulatory science advocates a lifecycle approach to AI-based medical device evaluation, including managed updates and continuous real-world monitoring, so that safety and accountability keep pace with iterative releases [23]. Legal scholarship likewise probes where responsibility should sit among developers, deployers, clinicians, and institutions; proposals range from adapted malpractice standards to product-liability and no-fault schemes that preserve innovation while ensuring redress [24]. Against this background, a narrative review remains appropriate because the relevant evidence base is heterogeneous by design—spanning clinical implementation research, bioethics, health law, regulatory guidance, and institutional governance—yet the central question is practical: how can governance-by-design embed fairness, transparency, and human oversight into real-world deployment without collapsing into slogans or compliance minimalism?

This thematic narrative review therefore analyzes the literature from 2018 to 2025 and integrates peer-reviewed scholarship with governmental and regulator-issued sources to examine seven domains—bias and fairness, explainability and transparency, safety and

quality, privacy and data protection, accountability and liability, human oversight, and procurement and deployment—in order to distill the ethical and legal implications of healthcare AI and to identify practice-facing directions for lifecycle governance.

2. Materials and Methods

We conducted a narrative review using a concept-driven approach to synthesize ethical, legal, and governance issues in healthcare AI. Searches were performed in Scopus, Web of Science Core Collection, PubMed/MEDLINE, Embase, and IEEE Xplore (journal articles only), covering 1 January 2018 to 9 November 2025 and limited to English-language, peer-reviewed journals (reviews, legal-policy analyses, and empirical studies addressing healthcare AI). In addition, given the governance focus of this manuscript, we purposively incorporated governmental and regulator-issued sources (e.g., statutory texts, implementing guidance, supervisory statements, and public procurement guidance) to support the comparative analysis across jurisdictions, including official regulator and governmental guidance from the United States (FDA), the United Kingdom (MHRA; NHS England; NICE), Canada (Health Canada), Australia (TGA), and selected Asian jurisdictions. We excluded conference proceedings, preprints, editorials/letters, and the non-official grey literature. A prototypical query combined controlled terms and keywords: ("artificial intelligence" OR AI OR "machine learning" OR "deep learning" OR "large language model*" OR "generative AI") AND (health* OR clinic* OR medicine) AND (ethic* OR governance OR oversight OR accountability OR liabilit* OR privacy OR GDPR OR "AI Act"). Two reviewers independently screened titles/abstracts and full texts; disagreements were resolved by consensus. Inclusion required explicit analysis of ethical and/or legal implications of AI in healthcare; studies focused solely on informatics or performance without ethics-legal content were excluded. We applied backward/forward citation chasing and hand-searched key journals (e.g., Lancet Digital Health, NPJ Digital Medicine, BMJ, JAMA). Data extraction captured study type, clinical domain, AI class (discriminative vs. generative/LLM—Large Language Model), ethical/legal foci, jurisdictional context, governance mechanisms, evaluation standards, and recommendations. No meta-analysis or formal risk-of-bias tool was applied; instead, we prioritized higher-level syntheses where available and triangulated findings across domains to support an integrative narrative interpretation.

3. Bias and Fairness

Bias in healthcare AI is multifactorial and rarely reducible to a single technical cause. Data underrepresentation interacts with modeling choices and institutional workflows to yield performance disparities and allocative harms that persist across settings [25–28]. Underdiagnosis and subgroup error asymmetries are now well documented in imaging and clinical prediction, with effects that are most visible for marginalized populations, including people with disabilities [29] and minoritized racial/ethnic groups [30–32]. These findings underscore that "fairness" is not an exogenous constraint but a design property that must be engineered and governed throughout the lifecycle—from data capture and labeling to validation, deployment, and post-market monitoring [27,33–36].

Sources of bias span dataset composition (sampling, measurement error), algorithmic objectives (proxy labels, cost-based targets), and sociotechnical embedding (triage rules, workflow incentives). For example, optimizing for healthcare costs as a proxy for need reproduced racial disparities in access to care in a widely deployed tool [31]. In medical imaging, models can infer sensitive attributes such as race from pixels even when clinicians cannot, raising both fairness and privacy concerns [33]. Conversely, diversified datasets and multisite training tend to reduce, though not eliminate, subgroup disparities [28,29,32]. Notably, parallel governance developments outside the EU increas-

ingly treat non-discrimination and bias assessment as practical requirements rather than aspirational principles, including in Asian frameworks that foreground human-centric, trustworthy, and accountable AI [14,37–39].

Mitigation evidence remains mixed. Pre-, in-, and post-processing strategies (reweighting, adversarial debiasing, calibrated thresholds) can improve selected metrics but often trade off against accuracy or shift error to unsupervised subgroups [28,29]. Recent reviews caution that many techniques are not interoperable and that fairness goals vary across contexts, requiring explicit value choices and stakeholder engagement to prioritize harms and benefits [26]. For clinical deployment, this implies that "fairness" should be specified in operational terms—target populations, error asymmetries that matter clinically, acceptable trade-offs, and escalation pathways when performance drifts—rather than asserted as a general virtue. In practice, governance mechanisms—model documentation, audit trails, impact assessments, and clear accountability for monitoring—are essential complements to technical fixes [25,35,40,41].

Legal-regulatory frames (GDPR principles; EU AI Act risk-based controls) increasingly shape design requirements around transparency, data lineage, and non-discrimination, but operationalization in clinics depends on institutional capacity: who is responsible for drift detection, subgroup audits, and corrective action [25,35,40,41]? Participatory and value-sensitive approaches help align model objectives with patient and community priorities, yet they face real constraints of time, representation, and authority [35]. The practical lesson is that fairness governance cannot be delegated to model developers alone; it requires organizational mandates, resourcing, and enforceable responsibilities across the vendor–provider interface.

Two empirical touchstones set the current agenda. First, rigorous error analysis by subgroup—ideally pre-specified and powered—should be routine, and adverse events linked to automation bias or brittle generalization must be reportable like any other safety signal [42–45]. Second, fairness goals should "level up," improving performance for disadvantaged groups without degrading care for others; this typically requires better targets, richer data, and careful thresholding rather than purely formal parity constraints [36,46,47]. Taken together, bias and fairness in healthcare AI demand integrated technical, ethical, and legal interventions that continue after go-live, with clear lines of responsibility and resourcing for ongoing surveillance and remediation [26,27,41].

4. Explainability and Transparency

The concepts of explainability and transparency are required by both ethical and legal principles: they enable contestability, support informed consent, and facilitate the attribution of responsibility when AI informs or automates care. In practice, explainability serves clinical usability and appropriate reliance, whereas transparency supports documentation, traceability, and audit, including change logs for model updates and interface revisions.

In Europe, the EU AI Act and the GDPR require disclosure of model logic, data provenance, and human intervention pathways, while the scope of a "right to explanation" remains debated [33,46,48]. These instruments elevate transparency from an ethical aspiration to an operational duty embedded in risk management and oversight. Similar signals appear in Asian regulatory and governmental guidance, relevant to cross-border procurement and multinational vendors, which increasingly treats documentation and lifecycle controls as conditions for trustworthy deployment [14,16–18].

Clinically, explainability is tied to trust and autonomy. Where AI shapes diagnosis or triage, clinicians and patients need intelligible accounts of system function, limits, and plausible failure modes [49]. Explanations should be calibrated to tasks, distinguishing model explainability (why an output is produced) from decision transparency (how outputs

should be acted upon) [50,51]. Regulatory transparency, by contrast, prioritizes demonstrability through stable documentation and the ability to contest and audit decisions.

The technical literature converges on two persistent difficulties. First, widely used post hoc methods (e.g., LIME—Local Interpretable Model-agnostic Explanations, SHAP—SHapley Additive exPlanations) can aid sense-making but may suffer from instability, locality, and user-interface burdens; systematic reviews in medicine and healthcare report improvements in perceived transparency alongside gaps in dataset diversity and alignment with clinicians' mental models [52,53]. Comparative studies show that different explanation techniques can yield divergent attributions on the same case, challenging reliability claims and complicating audit trails [54]. Explanation dashboards can help, but most tooling is not designed around clinical workflows or explicit safety cases [55].

Second, there are trade-offs and alternatives. Empirical work notes that interpretability via simpler models may reduce accuracy for some tasks, whereas other scholars argue that inherently interpretable models should replace black boxes in high-stakes settings when feasible [56–58]. A recurring critique warns that post hoc explanations can offer a "false hope" of safety if they are neither faithful to model internals nor actionable at the bedside [54]. The practical implication is to specify the purpose of explanation and validate it prospectively, including under plausible distribution shifts and real workflow constraints.

At deployment, transparency must be bound to intended use and interface design. A survey of FDA-cleared imaging AI finds heterogeneous outputs and often limited explainability vis-à-vis the labeled clinical role, signaling the need for standardized, task-appropriate disclosures and usability testing [59]. Meta-analytic evidence shows efficiency gains in imaging workflows, but few studies report whether explanations improve safety or decision quality—underscoring that "transparent enough for uptake" is not the same as "transparent enough for accountability" [60]. User-centred approaches propose heuristics (what, to whom, and when to explain), but institutions still require governance for versioning explanation assets and monitoring real-world utility [61]. Methodological syntheses recommend anchoring explainability in measurable documentation: specify explanation targets, test stability across shifts, and evidence clinical usefulness—not only aesthetic plausibility—for generative and LLM-based systems [62–65].

5. Safety and Quality

The safety and quality of clinical AI cannot be inferred from development-phase performance alone. In healthcare, "safe enough" is a moving target shaped by population mix, clinical workflows, data pipelines, and the frequency with which models and interfaces are updated. Cross-sector syntheses and domain-specific evidence converge on recurrent risks that can undermine real-world validity after go-live: distorted or low-fidelity data pipelines, unanticipated distribution shift, opacity that frustrates oversight, and weak organizational capacity to learn from errors—within a global market characterized by uneven regulatory supervision across jurisdictions [66–68].

Professional surveys and practitioner-facing commentary add a workforce lens that is frequently underspecified in governance discussions. Clinicians commonly acknowledge potential efficiency gains, yet report concerns about deskilling, job displacement, workflow disruption, and the redistribution of scarce expertise toward monitoring and exception handling rather than patient-facing work [69,70]. These concerns are not peripheral: they affect how systems are used, how errors are detected, and whether escalation pathways are realistically implemented under routine pressures. Accordingly, safety should be treated not as a static performance metric (e.g., Area Under the Receiver Operating Characteristic curve—AUC) but as a continuous, socio-technical property that can be produced—or

eroded—by training, staffing, interface design, and human–system integration across the lifecycle [67,68].

Standards and regulatory science are increasingly attempting to codify quality signals in operational terms. In China, the YY/T (Chinese medical device industry standard code prefix) series and related guidance are translating dataset governance and risk management into more concrete expectations for medical AI development, validation, and surveillance [71,72]. In the United States, the FDA has formalised a lifecycle approach for AI-enabled devices through final guidance on PCCPs, explicitly linking update pathways, documentation, and post-market performance assurance to continued safety and effectiveness [7]. In addition, regulatory convergence is visible in joint Good Machine Learning Practice (GMLP) principles and in complementary national guidance from the United Kingdom (MHRA SaMD Change Programme), Canada (Health Canada pre-market guidance for ML-enabled medical devices), and Australia (TGA guidance on AI and medical device software), each reinforcing auditable change control and lifecycle monitoring as core quality expectations [8,9,12,13].

Parallel work emphasizes that adaptive and continuously learning systems raise unresolved questions about proportionality, evidentiary thresholds, and acceptable update pathways, particularly when performance changes are iterative rather than episodic [73,74]. Within the EU context, these developments interact with conformity assessment logics and post-market obligations, including how "notified bodies" and institutional governance structures can credibly audit adaptive models and version changes over time [75].

Questions of responsibility are correspondingly becoming more prominent, because safety failures in practice are rarely attributable to a single actor. Legal scholarship and regulatory debates increasingly recognise that accountability is distributed—often unevenly—across developers, vendors, deployers, and clinicians, and that liability concepts developed for static products fit imperfectly with systems that can change after deployment [24,76]. In operational terms, this diffusion becomes clinically consequential when adverse performance emerges, and it is unclear who has authority and duty to investigate, pause, rollback, or retire a system, particularly where procurement and contractual terms place monitoring obligations on institutions without granting sufficient access to model documentation and update information. This constellation produces a predictable governance weakness—often described as an "accountability gap"—unless responsibilities for drift detection, subgroup audits, incident response, and corrective action are explicitly allocated and resourced [41].

Practical experience indicates that it is not possible to anticipate all clinically relevant risks and failure modes prior to launch; accordingly, post-deployment governance should be designed to surface weak signals early and trigger defined corrective actions (e.g., recalibration, threshold adjustment, workflow redesign, retraining, or retirement) rather than relying on informal escalation [77]. For instance, external validation of the widely deployed Epic Sepsis Model, which showed poor discrimination and calibration, illustrates how proprietary systems can underperform at the bedside despite broad adoption [77]. Methodological contributions therefore advocate explicit safety cases, structured post-market surveillance, and stress testing under plausible shifts—including changes in protocols, coding practices, and user behaviour induced by system presence—so monitoring does not degrade into sporadic audits [78,79]. Evidence from the FDA's Manufacturer and User Facility Device Experience (MAUDE) database further underscores the need for standardized benefit–risk documentation and incident taxonomies, as reports of safety events involving ML-enabled devices are often too sparse to support root-cause analysis [78,79]. Real-world studies using Natural Language Processing (NLP) on electronic health records quantify adverse events and resource utilization, but also show that measurement infrastructures—not only algorithms—shape what we can detect and fix [80]. A resilience agenda for clinical AI

therefore emphasizes explicit safety cases, stress testing under shift, and governance that treats monitoring as a first-class requirement, not a bolt-on [81].

Overall, safety and quality are socio-technical accomplishments rather than intrinsic attributes of an algorithm. Human-centred design, organizational readiness, and governance capacity determine whether explanations, alerts, and hand-off protocols are usable and whether clinicians can maintain appropriate reliance under time constraints. Without credible documentation, resourced post-market surveillance, and workforce capability to interpret and act on safety signals, even high-performing models remain vulnerable to silent failure in real-world care [59,67,70,82].

6. Privacy and Data Protection

When analysing privacy and data protection in clinical and healthcare settings, the focus extends well beyond confidentiality and cybersecurity. It includes consent validity, proportionality, fairness in data-driven decision-making, and the legitimacy of secondary uses and cross-border data flows. Across recent scholarship, the most recurrent challenges converge on four risks: excessive or weakly governed data circulation, fragile consent pathways, inadequate security controls, and distributive harms that can fall disproportionately on already vulnerable patients [83–85]. Within the EU, GDPR duties—lawfulness, purpose limitation, data minimisation, integrity/confidentiality, and "data protection by design and by default"—must be read together with sectoral rules and the risk-based obligations of the (now adopted) AI Act [86]. Alignment is necessary to prevent governance gaps between care and research uses and to clarify when "meaningful human involvement" is required to avoid solely automated decisions [87]. Institutions operating across borders must contend with partially convergent privacy regimes outside Europe: Asian frameworks similarly emphasize purpose limitation, minimisation, and security safeguards while differing in enforcement models and transfer rules [88–91], and comparable constraints shape deployment in the United States, where Health Insurance Portability and Accountability Act (HIPAA) compliance and strengthening cybersecurity expectations for electronic protected health information (ePHI) frame organizational safeguards for AI systems processing clinical data [92]. Operationally, UK implementation guidance underscores procurement-time information-governance checks, clarity on medical-device status, and the requirement that outputs remain reviewable by staff, illustrating how privacy duties translate into deployment controls in routine care [11]. In Canada and Australia, privacy regulators have likewise issued AI-specific guidance emphasising accountability, vendor due diligence, and safeguards for organisations adopting commercially available AI products, with provincial health-sector guidance further operationalising these expectations for hospitals and their vendors [93,94].

Generative AI introduces additional model- and data-level threats that complicate traditional assumptions about clinical confidentiality. Model inversion and extraction, memorisation and leakage, opaque training provenance, and unauthorised scraping are not hypothetical risks; such phenomena have already been documented in clinical domains, including ophthalmology [95]. The issues raised by generative systems cannot be treated as a static compliance target, because the threat surface evolves with deployment modalities, fine-tuning practices, and downstream reuse of outputs. As a result, privacy governance for generative AI must be positioned as a continuous programme of monitoring, updating, and improvement rather than a one-time "sign-off" exercise.

Numerous technical approaches can strengthen privacy protections in concrete terms, but none function as a standalone solution. Federated learning (FL) allows institutions to train models without exchanging raw data; when coupled with differential privacy (DP) and secure aggregation, FL can reduce re-identification risks while maintaining

competitive performance in multi-institutional settings [96–99]. However, DP budgets, gradient sparsity, and distribution shift can impose measurable utility costs; accuracy-aware and adaptive DP schemes may mitigate these losses but do not eliminate them, particularly for minority subgroups and rare outcomes where the performance margin is clinically consequential [100,101]. The practical implication is that Privacy-Enhancing Technologies (PETs) should be selected with explicit reference to clinical tasks and validated against clinically meaningful endpoints, rather than being judged solely on surrogate privacy metrics that may not map to real-world harm reduction.

Beyond FL and DP, blockchain-backed audit trails and homomorphic encryption are frequently presented as enabling integrity, verifiability, and controlled disclosure in health data ecosystems. The evidentiary base, however, indicates that most blockchain applications remain at early technology-readiness levels (TRL 3–5), and that governance constraints—consent verification, on/off-chain partitioning, interoperability, and performance—continue to limit real-world uptake in healthcare contexts. Where deployed, hybrid architectures (e.g., permissioned ledgers for consent and audit, combined with off-chain encrypted storage and PETs for analytics) appear more pragmatic than "full-stack" blockchain designs and better aligned with clinical and institutional constraints [102,103].

From a legal perspective, GDPR provides robust guardrails for special-category data, international transfers, and data subject rights; yet hospitals and vendors frequently struggle with heterogeneous local implementations, legacy infrastructures, and unclear accountability across joint controllers and processors [85,87,104]. These frictions are especially visible in IoT-enabled and app-mediated care, where privacy policies are inconsistent, security controls are uneven, and purpose creep is common [105,106]. A proportionality-by-design approach—minimising data and permissions ex ante, embedding contextual safeguards, and maintaining tight linkage between data collection and clinical purpose—has been proposed for mHealth and mental-health apps and remains broadly applicable to clinical AI, including systems embedded in routine pathways rather than discrete research protocols [107].

Critically, "release-and-forget" anonymisation is no longer a safe assumption in modern, high-dimensional clinical datasets. Face-recognition studies have re-identified participants from Magnetic Resonance Imaging/Computed Tomography (MRI/CT) even after defacing pipelines, and population-uniqueness analyses show that "de-identified" data can often be re-linked with few attributes, challenging legal adequacy under modern standards [108,109]. Synthetic data can support privacy-preserving sharing and development, but privacy and utility guarantees are method- and context-dependent; rigorous, standardized evaluation and attack testing are essential before clinical adoption [110–112]. This implies that the system, algorithm, or device must undergo thorough testing according to recognised and reproducible methodologies to ensure reliability, robustness, accuracy, and regulatory compliance. Overall, privacy protection in healthcare AI requires a layered strategy: purpose-specific governance (e.g., DPIAs—Data Protection Impact Assessments, registries, access logs), PETs validated against clinical utility, and continuous monitoring for leakage and drift across the lifecycle.

7. Accountability and Liability

Accountability and liability are the institutional "hinges" on which trustworthy clinical AI must turn. Normative principles—autonomy, beneficence, justice, transparency, and accountability—require translation into operational duties for developers, healthcare organisations, and clinicians, supported by auditable records across the model lifecycle rather than post hoc assurances [25,67,85]. In practice, tort doctrines and product-liability rules are strained by adaptive and opaque systems and by distributed agency: clinically

salient outputs often emerge from the interaction of data pipelines, local configuration choices, workflow constraints, and iterative updates rather than a single negligent act. Accordingly, the literature converges on a persistent "responsibility gap" unless governance and liability are co-designed, including explicit role allocation and enforceable monitoring duties [24,76,113,114]. Recent cross-specialty analyses and radiology-focused syntheses emphasize that safety cases, performance surveillance, and clearly defined implementation roles (e.g., model owner, clinical champion, risk officer, information governance lead) are preconditions for meaningful accountability—not add-ons once deployment has occurred [67,115–117]. These pressures are not unique to Europe: Asian governance and regulatory guidance increasingly foreground accountable entities, lifecycle documentation, and post-deployment control as necessary conditions for trustworthy adoption in health systems, with implications for multinational vendors and cross-border procurement [14,16–18].

Traditional notions of concurrent responsibility in clinical practice are imperfect fits for the activities of clinical AI. Models of enterprise liability and shared responsibility better reflect the multi-actor reality than frameworks that focus exclusively on individual clinician negligence: they acknowledge overlapping tasks among physicians (appropriate use, supervision, disclosure), institutions (procurement, governance, monitoring, training), and manufacturers (design controls, quality management, post-market obligations) [24,114–116]. In this view, the safety and reliability of clinical AI depend on a chain of interconnected responsibilities in which each actor has a specific, non-isolated role: clinicians remain responsible for context-sensitive use and oversight; healthcare organisations for adoption decisions, resourcing, and governance capacity; and manufacturers for technical quality, transparency, and update practices.

Liability for the use of AI in healthcare, therefore, shifts attention from a fault-based model centred on individual negligence towards more systemic and adaptive approaches. No-fault schemes aim to provide timely compensation without requiring proof of individual fault, potentially reducing litigation burden while enabling the adoption of beneficial technologies. In parallel, insurance pools or compensation funds distribute risk across hospitals, manufacturers, and insurers, seeking a more resilient arrangement that preserves patient redress while avoiding innovation-chilling uncertainty [24,118]. Comparative EU analyses indicate that the AI Act and related liability reforms move governance forward—particularly regarding provider/user duties and documentation—yet important gaps persist for black-box models, causation proofs, and continuously learning updates [76,119–121].

Accountability also has a practical "how" that is realised through local governance mechanics. Hospital-level frameworks (e.g., ABCDS lifecycle—Algorithm-Based Clinical Decision Support) embed pre-deployment "silent" evaluation, effectiveness checks, role assignment, and continuous quality assurance, making responsibilities inspectable and enforceable rather than implicit [116,117]. Enterprise risk management (ERM) approaches can align clinical risk, cybersecurity, and regulatory obligations, clarifying escalation pathways and decision rights when AI systems deviate from expected performance [117]. In radiology, governance questions ("who decides and how?") arise at each stage—selection, credentialing, workflow integration, monitoring, and de-implementation—requiring explicit decision authorities and contemporaneous documentation capable of withstanding scrutiny by courts and regulators [115].

Informed consent and transparency remain pillars of legal accountability, but they require adaptation as AI shapes diagnosis, triage, and treatment pathways. Tiered disclosures, plain-language explanations of AI roles and risks, and clinician education have been proposed, alongside explainability tools that support patient-facing dialogue in a clinically

usable manner [49,122,123]. Explainable Artificial Intelligence (XAI), however, does not by itself close the liability loop. What matters legally is whether implementers met a defensible standard of care: appropriate selection and calibration, monitoring for drift, documentation of known limitations, and disclosure of material information to patients when AI meaningfully affects clinical decision-making [25,49,122,124]. Recent peer-reviewed commentary also links intended use, explainability obligations, and regulator-recognised PCCPs to auditable accountability, particularly for systems subject to frequent updates [125,126].

Global perspectives in guidelines, regulations, and recommendations highlight diverse—and at times conflicting—approaches, characterised by non-uniform technical standards, different evidentiary expectations, and variable governance models. EU doctrine tends to couple rights protection with harmonised compliance duties, while other jurisdictions more heavily rely on judge-made negligence principles to adjust implementer obligations for AI-enabled care [119–121]. Sectoral syntheses (notably in radiology) converge on a pragmatic evidentiary pathway: performance claims anchored to clinically meaningful endpoints, post-market studies, and transparent governance records that courts can parse when harms occur [67,115,125]. The research frontier remains predominantly empirical: there is a sustained need for analyses of real-world liability disputes and prospective evaluations of ERM and governance models to demonstrate that they reduce preventable harm without stifling innovation [24,116,117].

8. Human Oversight

In healthcare settings where advanced algorithms or AI systems are deployed, ethical principles continue to require a concrete and meaningful human presence and intervention. In the clinical domain, ethically acceptable and safe AI technologies are expected to support, rather than replace, human decision-making, thereby safeguarding patient autonomy and beneficence as core bioethical principles. Recent reviews emphasize that human oversight should be mandated throughout the entire clinical pathway of the AI life cycle—including procurement, validation, implementation, monitoring, and de-implementation—rather than being limited to pre-market control [25,67,127].

Regulators increasingly articulate this lifecycle view. The FDA's 2025 perspective frames oversight around intended use, evidence standards, and post-market adaptation for learning systems—placing provider institutions alongside manufacturers in maintaining performance and documenting change [128]. In parallel, governance implementations in oncology and multi-stakeholder roadmaps show how local committees, role assignment, and transparent decision rights can operationalize oversight across clinical, operational, and research programmes, while ecosystem approaches address commercial actors and conflicts of interest [129–131]. Comparable emphases on accountable governance, lifecycle documentation, and post-deployment control also appear in Asian governmental and regulatory guidance, reinforcing the practical relevance of human oversight for cross-border procurement and multinational deployment [14–18].

Oversight also requires tools. Human-centred monitoring ("algorithmovigilance") focuses on detecting drift, subgroup performance cliffs, and incident patterns after go-live, with design work demonstrating how dashboards and workflows can support clinicians and risk officers [132]. Conceptual bridges from pharmacovigilance outline taxonomies and routines for safety signal detection tailored to AI [133]. Together, these approaches move oversight from principle to daily practice.

Traditional ethics committees or institutional review boards (IRBs) are no longer sufficient to address the challenges posed by emerging technologies such as clinical AI, connected devices, and health big data. These bodies will need to adapt to complex, dynamic, and multi-actor contexts. Current studies question whether traditional IRBs are

adequately equipped to deal with the sociotechnical risks of AI, and recommend additional expertise, digital tools, and explicit procedures for the assessment of automated or semi-automated interventions [134,135]. Ethics committees will therefore need to broaden their composition to include AI technology experts alongside clinicians, bioethicists, and legal scholars, and to involve representatives of patient organisations and the wider social community [136].

Human oversight is regarded as essential in decision-making: clinician training and consent processes, user-centred alerting, and governance records that link model intent to interface behaviour. Recommendations for AI-enabled clinical decision support call for pre-specified monitoring plans, clear delineation of roles, and transparent communication to patients and staff; participatory methods further align oversight criteria with local values and workflows [137]. At the policy level, EU AI Act–oriented analyses underscore that "human oversight" is not symbolic—it must prevent or minimise health and rights risks through concrete controls and documentation that regulators (and courts) can inspect [138].

9. Procurement and Deployment

Procurement and implementation of clinical artificial intelligence are most successful when governance, engineering, and clinical operations are centralized and coordinated, and when all involved persons and systems operate within a single process rather than duplicative or parallel pathways. Recurrent themes across reviews include bias mitigation and dataset suitability checks at the point of purchase, transparency obligations embedded in contracts and model documentation, variation in regulatory expectations across jurisdictions, and the need for multidisciplinary decision rights spanning Information Technology (IT), legal, quality, and frontline services [25,67,127]. In operational terms, this jurisdictional variability spans EU and North American expectations and increasingly incorporates Asian lifecycle-oriented guidance and regulator-issued requirements that shape vendor documentation, update practices, and post-deployment control, with direct implications for cross-border procurement and multinational deployment [15–18]. Practice-facing instruments illustrate how these requirements can be operationalized at the purchasing stage: in the UK, the NHS "buyer's guide" structures due diligence through standardized questions on safety, effectiveness, monitoring, documentation, and deployment readiness, supported by an assessment template for procurement decisions [139], and NICE's updated Evidence Standards Framework explicitly aligns evidence tiers for AI and data-driven technologies (including adaptive algorithms) with deployment and lifecycle expectations [10]. Beyond the UK, procurement due diligence is increasingly anchored to regulator-issued lifecycle controls for AI-enabled software: in the United States, FDA guidance on PCCPs clarifies how update pathways, documentation, and post-market assurance should be specified for AI-enabled device software functions and translated into contractual obligations at purchase [7], while Canada and Australia provide complementary signals through Health Canada's pre-market guidance for machine learning–enabled medical devices and Australia's TGA guidance on AI and medical device software, reinforcing evidentiary expectations, documentation duties, and lifecycle controls that inform procurement specifications and deployment readiness checks [12,13]. This convergence supports governance arrangements that deliberately integrate diverse expertise—computer scientists, legal professionals, quality specialists, and clinicians in direct contact with patients—so that adoption decisions remain defensible and implementable in routine care.

Scoping work also shows that sector-specific contexts (e.g., dental and primary care) surface distinct integration risks, reinforcing the value of local stakeholder input during pre-procurement market scans and pilot selection [127]. A recent synthesis of implementation

frameworks underscores that health systems should evaluate tools not only for accuracy but also for maintainability, monitorability, and equity impact before and after go-live [140].

Organizational readiness is decisive. Institutions that plan for training, role assignment, and escalation paths report smoother adoption and fewer adverse surprises during early deployment. Commentaries point to persistent capability gaps—particularly around updating, de-implementation, and staff education—which procurement teams should address via explicit vendor obligations (e.g., change control plans, audit logs, and support for post-market studies) [70,141].

Lifecycle-oriented governance frameworks offer practical scaffolding. The Health Equity Across the AI Lifecycle (HEAAL) framework embeds equity checks across procurement and deployment stages; multi-stakeholder governance models outline how to balance speed, scope, and capability; and provincial programs (e.g., British Columbia) illustrate coordinated intake, risk assessment, and evaluation at system scale [130,142,143]. Together, these approaches make responsibilities inspectable and enable course-correction when performance drifts or local workflows change.

Operationally, health services need to balance local flexibility with standardization. A useful sequencing is "flexibility first, then standardize": configure early pilots to fit local practice, then codify reusable standards for data quality, monitoring, and documentation [144]. Recent work reframes this tension as "responsible scaling," arguing that standardization should support—not flatten—local sociotechnical configurations [145]. Reporting and oversight tools (e.g., DECIDE-AI—Developmental and Exploratory Clinical Investigations of Decision-support systems driven by Artificial Intelligence items for early clinical evaluation, and ABCDS oversight for local governance) help translate that balance into concrete procurement and deployment checklists [116,146].

Deployment pipelines should anticipate frequent updates. Continuous integration and deployment (CI/CD) patterns—versioned models, rollback plans, signed artifacts, and release notes mapped to intended use—reduce downtime and compliance risk, provided they are paired with post-market monitoring and clear decision rights for pausing, updating, or retiring models [147,148]. In sum, effective procurement and deployment combine equity-aware selection, contractual transparency, lifecycle monitoring, and agile change control to keep AI safe, fair, and adaptable in real care environments.

Table 1 below summarizes the governance themes for clinical artificial intelligence.

Table 1. Governance themes for clinical artificial intelligence.

Governance Theme	Key Risk	Priority Response (Operational Mechanisms)
Bias and Fairness	Systematically unequal performance and allocative harms due to biased data, proxies, and context-dependent deployment.	Fairness-by-design with pre-specified subgroup metrics and thresholds; datasheets/model cards; multi-site external validation; routine subgroup audits post-deployment; documented escalation and remediation (recalibration, retraining, de-implementation); stakeholder/patient review for impact and redress.
Explainability and Transparency	Clinically and legally insufficient intelligibility, contestability, and traceability of outputs and updates.	Separate clinical usability explainability from regulatory transparency; task-bound explanation validation (stability, usefulness); standardized documentation (model cards, intended use, limitations); audit logs and versioning; release notes for updates; change control plan linking model/interface changes to evidence.

Table 1. *Cont.*

Governance Theme	Key Risk	Priority Response (Operational Mechanisms)
Safety and Quality	Silent performance degradation under dataset shift, workflow change, or updates, with delayed detection of harm.	Pre-deployment silent trials and prospective validation on local workflows; post-market surveillance with drift detection; incident reporting and safety signal review; periodic re-certification/reevaluation; predefined rollback/kill-switch authority; integration into institutional quality management and risk registers.
Privacy and Data Protection	Unlawful secondary use, re-identification, leakage/memorization (incl. generative AI), and weak accountability across controllers/processors.	DPIA and purpose limitation; data minimisation and access controls; encryption, logging, retention rules; privacy-enhancing techniques where appropriate (e.g., federated learning + differential privacy) with clinically meaningful utility testing; documented data-sharing/transfer governance and breach response.
Accountability and Liability	Responsibility gaps in multi-actor systems (developer–vendor–provider–clinician), especially under frequent updates and opaque services.	Explicit Responsible, Accountable, Consulted, Informed (RACI)-style allocation of duties; contractual clauses on documentation, monitoring, audit rights, and update notification; preserved evidence trails (logs, validation reports, monitoring outputs); defined incident investigation pathway; alignment with enterprise risk management; suitable compensation/insurance arrangements where appropriate.
Human Oversight	Nominal "human-in-the-loop" without real authority, skills, or time to intervene, leading to automation bias and unmanaged risk.	Defined decision rights (override, pause, retire); training and competency checks; escalation pathways and accountability for monitoring actions; human-factors testing of interfaces/alerts; governance committee with documented meeting outputs; periodic review of reliance patterns and override events.
Procurement and Deployment	Technology-led purchasing that omits governance requirements, long-term maintainability, monitoring capacity, and de-implementation.	Multidisciplinary procurement with mandatory governance artefacts (model card, monitoring plan, DPIA, change control plan); performance Service Level Agreements (SLAs) and audit clauses; interoperability and data-quality prerequisites; implementation and training plan; obligations for post-market studies and update transparency; explicit de-implementation/exit provisions.

10. Discussion

A review of the literature indicates that, in healthcare, the real-world performance of artificial intelligence depends less on numerical metrics and technical measurements and far more on how a set of interconnected dimensions is jointly managed, including:

- Equity—ensuring that all patients are treated fairly;
- Transparency—enabling an understanding of how AI systems reach their decisions;
- Safety—preventing harmful errors;
- Privacy—protecting patients' data;
- Responsibility—clarifying who is accountable in the event of problems;
- Oversight—maintaining meaningful human control;
- Organizational practices—determining how the hospital or institution integrates and uses AI in routine care.

First, bias is not a static property of datasets but a moving target across the lifecycle. Bias should therefore be treated as a dynamic phenomenon that requires ongoing monitoring and correction throughout the entire life cycle of the system or model, rather than only at the outset. Evidence from the implementation and methods literature shows that subgroup error asymmetries persist unless diversity is engineered into sampling frames, labels, and external validation, and unless monitoring detects performance cliffs after go-live [28]. Participatory approaches add a pragmatic lens—if patients, frontline clinicians, and communities help define problem scope, success metrics, and redress pathways, models are more likely to meet equity objectives and withstand distribution shift. These insights align with earlier demonstrations that proxy targets can entrench disparities (e.g., using costs to proxy need) and that fairness must be specified as a design constraint, not an ex post patch [31,32].

Explainability and transparency connect model behavior to clinical judgment and patient autonomy. Regulatory analyses clarify that transparency is now an operational duty—tied to documentation, the availability of human intervention, and risk management—rather than an aspirational ethic [149]. Yet integration at the bedside remains uneven: post hoc techniques can aid sense-making but may be unstable or misaligned with clinical mental models, so their purpose must be explicit (debugging vs. audit vs. communication) and prospectively validated [54]. Task-appropriate explainability for risk tools—what will be explained, to whom, and at what decision point—has been proposed, together with interface and workflow testing to confirm that explanations are intelligible and actionable for clinicians under time pressure [64]. In short, transparency that is "good enough for uptake" may still be insufficient for accountability unless it is linked to intended use, documentation, and oversight.

Safety and quality hinge on post-market realities. Adverse-event signals derived from electronic health records and natural-language processing demonstrate that AI-mediated care can introduce measurable burdens and harms alongside efficiencies, underscoring the need for robust surveillance infrastructures [80]. Organizational ethnographies describe a 'responsibility vacuum' that arises when no actor is clearly tasked with ongoing monitoring, subgroup auditing or corrective action, resulting in fragile implementations despite strong pre-market performance [41]. These findings echo external validation failures of widely implemented proprietary models, reminding us that safety is a continuous property produced by monitoring, feedback loops, and the capacity to pause, update, or retire systems [59,146,150]. Accordingly, early-stage evaluation guidance and local governance checklists should be treated not as publication formalia but as procurement and deployment requirements that bind vendors and implementers to measurable safety work over time [146,151].

Privacy and data protection are equally dynamic. Privacy-enhancing technologies—federated learning with differential privacy, secure aggregation, and cryptographic supports—can reduce re-identification risk while enabling multi-institutional learning, but they impose non-trivial utility costs and new attack surfaces that must be quantified on clinically meaningful endpoints [96,97]. Systematic reviews of blockchain-based data sharing report promising integrity and auditability features but also governance and interoperability gaps that limit production use in hospitals [103]. In regulated settings, these technical measures must be coupled to GDPR-conformant governance (data minimization, purpose limitation, and accountability) and to clear documentation of roles and transfers; otherwise, privacy duties drift as models are repurposed, fine-tuned, or embedded in new workflows.

Accountability and liability supply the incentives that keep these duties real. Physician-centric malpractice theories map poorly onto multi-actor, learning systems; enterprise or

shared-responsibility models better reflect overlapping duties across developers (design and quality), deployers (procurement, monitoring, de-implementation), and clinicians (selection, disclosure, oversight) [114,118]. These approaches do not absolve professionals; rather, they distribute obligations and create auditable records that courts and regulators can parse. Practical implications follow: procurement should require change-control plans, model registries, and incident reporting; deployment should assign decision rights for pausing or rolling back updates; and documentation should connect intended use, evidence claims, and explanation assets to the interfaces clinicians actually see [116,125,152].

Human oversight operationalizes these expectations. Regulators now frame oversight as lifecycle management with explicit duties for users and providers, especially for learning systems whose risks evolve with context [128]. For example, governance programs in oncology demonstrate how multidisciplinary committees, role assignment, and transparent decision pathways can make oversight inspectable, while "algorithmovigilance" prototypes illustrate monitoring dashboards and escalation routines for drift, subgroup harms, and incident clustering [129,132,133]. Ethics review must also adapt: IRBs need domain-specific expertise, digital tooling, and procedures for semi-automated interventions, together with inclusive membership to sustain legitimacy for research that blurs clinical and data-science boundaries [134–136].

Procurement and deployment are where these strands are braided into day-to-day operations. Multi-stakeholder frameworks emphasize equity checks and monitoring plans from the outset, and provincial models show how coordinated intake and risk assessment can scale governance across health systems [130,142,143]. Organizations face a perennial tension between local flexibility and standardization; a sensible sequence is to permit configuration for early pilots and then lock down reusable standards for data quality, telemetry, and documentation—"flexibility first, then standardize" [144,145]. To support frequent model changes without safety debt, CI/CD practices (versioning, signed artifacts, rollback plans) should be paired with post-market surveillance, DECIDE-AI-style reporting, and clear authority to halt models when evidence degrades [146,147,153]. Throughout, workforce capability—training, time, and support—remains a critical constraint; without it, even well-designed governance cannot be enacted at the bedside [70].

This narrative review has limitations that should be made explicit. The literature search and source selection were purposive and concept-driven rather than exhaustive, and we did not apply formal risk-of-bias instruments; accordingly, reproducibility is lower than in systematic approaches, and interpretive subjectivity cannot be fully eliminated. In addition, the rapid evolution of generative and adaptive systems means that technical practices, documentation norms, and supervisory expectations can change faster than peer-reviewed consolidation, creating an inherent lag between governance needs and settled evidence.

Taken together, the literature points to a pragmatic agenda and a set of practice-facing research directions: specify fairness as a design and monitoring target; treat explainability as a task-bound, validated clinical asset; build surveillance and de-implementation capacity to make safety continuous; align privacy PETs with legal accountability; distribute responsibility across the enterprise; embed oversight in daily operations; and procure for equity, transparency, and change. Future work should prioritise implementation studies that evaluate governance mechanisms in situ (monitoring dashboards, escalation pathways, change-control plans, and procurement clauses), clarify which stakeholders benefit and under what conditions (patients, clinicians, risk managers, procurement offices, and regulators), and generate empirically grounded evidence on how accountability and liability models operate in real disputes and adverse-event investigations.

11. Conclusions

Trustworthy clinical AI is not delivered by algorithms alone, but by institutions that procure, deploy, and govern them with equity, safety, and accountability in view. Across the EU and emerging Asian governance trajectories, the consistent practical requirement is the same: translate principles into auditable mechanisms—role assignment and decision rights, dataset and subgroup checks, documentation and versioning, change control for updates, and resourced post-deployment surveillance with incident reporting and corrective action. Sustained clinical value also depends on workforce capability and modern infrastructure. Future research should test, in real care settings, which governance and procurement mechanisms most effectively reduce preventable harm while preserving clinical utility and operational feasibility.

Author Contributions: Conceptualization, P.B., G.R. and E.B.; methodology, G.N.; validation, G.N. and T.S.; formal analysis, P.B.; investigation, E.B. and G.N.; data curation, G.P. and E.B.; writing—original draft preparation, P.B. and G.P.; writing—review and editing, P.B. and T.S.; supervision, P.B. and G.R.; project administration, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript/study, the author(s) used DeepL and ChatGPT 5.1 for the purposes of enhance fluency, syntax, and grammar (text editing). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABCDS	Algorithm-Based Clinical Decision Support (lifecycle/oversight framework for clinical prediction models)
AI	Artificial Intelligence
AUC	Area Under the Receiver Operating Characteristic Curve
CI/CD	Continuous Integration and Continuous Deployment (or Delivery) pipelines
CONSORT-AI	Consolidated Standards of Reporting Trials—Artificial Intelligence (CONSORT extension for AI interventions)
DECIDE-AI	Developmental and Exploratory Clinical Investigations of Decision-Support Systems Driven by Artificial Intelligence
DP	Differential Privacy
DPIA/DPIAs	Data Protection Impact Assessment/Data Protection Impact Assessments
ePHI	Electronic protected health information
ERM	Enterprise Risk Management
FDA	United States Food and Drug Administration
FL	Federated Learning
GDPR	General Data Protection Regulation
GMLP	Good Machine Learning Practice (joint principles)
HEAAL	Health Equity Across the AI Lifecycle (framework for assessing how AI affects health equity)

HIPAA	Health Insurance Portability and Accountability Act
HSA	Health Sciences Authority
IRB/IRBs	Institutional Review Board/Institutional Review Boards
IT	Information Technology
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MAUDE	Manufacturer and User Facility Device Experience (FDA adverse event database for medical devices)
MFDS	Ministry of Food and Drug Safety
MHRA	Medicines and Healthcare products Regulatory Agency
ML	Machine Learning
MRI/CT	Magnetic Resonance Imaging/Computed Tomography
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NMPA	National Medical Products Administration
PCCP/PCCPs	Predetermined Change Control Plan/Predetermined Change Control Plans (for regulated AI/ML-enabled medical devices)
PET/PETs	Privacy-Enhancing Technology/Privacy-Enhancing Technologies
PMDA	Pharmaceuticals and Medical Devices Agency
RACI	Responsible, Accountable, Consulted, Informed
SaMD	Software as a Medical Device
SLA/SLAs	Service Level Agreement/Service Level Agreements
SHAP	SHapley Additive exPlanations
SPIRIT-AI	Standard Protocol Items: Recommendations for Interventional Trials—Artificial Intelligence (SPIRIT extension for AI interventions)
TGA	Therapeutic Goods Administration
TRL/TRLs	Technology Readiness Level/Technology Readiness Levels
XAI	Explainable Artificial Intelligence
YY	Chinese medical device industry standard code prefix (YY/YYT series for sectoral specifications, e.g., dataset quality standards)

References

1. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef]
2. He, J.; Baxter, S.L.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **2019**, *25*, 30–36. [CrossRef]
3. Morley, J.; Machado, C.C.V.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **2020**, *260*, 113172. [CrossRef]
4. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef]
5. Cruz Rivera, S.; Liu, X.; Chan, A.-W.; Denniston, A.K.; Calvert, M.J. The SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat. Med.* **2020**, *26*, 1351–1363. [CrossRef] [PubMed]
6. Busch, F.; Kather, J.N.; Johner, C.; Moser, M.; Truhn, D.; Adams, L.C.; Bressem, K.K. Navigating the European Union Artificial Intelligence Act for healthcare. *npj Digit. Med.* **2024**, *7*, 210. [CrossRef] [PubMed]
7. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. 18 August 2025. Available online: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence (accessed on 14 January 2026).
8. U.S. Food and Drug Administration. Artificial Intelligence in Software as a Medical Device (SaMD). 25 March 2025. Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device (accessed on 14 January 2026).

9. Medicines and Healthcare products Regulatory Agency (MHRA). Software and AI as a Medical Device Change Programme Roadmap. 14 June 2023. Available online: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap (accessed on 14 January 2026).
10. National Institute for Health and Care Excellence (NICE). Evidence Standards Framework for Digital Health Technologies (Updated to Include AI and Data-Driven Technologies with Adaptive Algorithms). Available online: https://www.nice.org.uk/corporate/ecd7 (accessed on 14 January 2026).
11. NHS England. Artificial Intelligence—Information Governance Guidance (Includes Procurement-Time Checks, Medical-Device Status, and Reviewability of Outputs). 30 April 2025. Available online: https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence/ (accessed on 14 January 2026).
12. Health Canada. Pre-Market Guidance for Machine Learning-Enabled Medical Devices. 5 February 2025. Available online: https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/pre-market-guidance-machine-learning-enabled-medical-devices.html (accessed on 14 January 2026).
13. Therapeutic Goods Administration (TGA). Artificial Intelligence (AI) and Medical Device Software. 4 September 2025. Available online: https://www.tga.gov.au/products/medical-devices/software-and-artificial-intelligence/manufacturing/artificial-intelligence-ai-and-medical-device-software (accessed on 14 January 2026).
14. Personal Data Protection Commission (PDPC). Model AI Governance Framework (Second Edition). Available online: https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework (accessed on 14 January 2026).
15. Health Sciences Authority (HSA). Regulatory Guidelines for Software Medical Devices—A Life Cycle Approach (Revision 2, 29 April 2022). Available online: https://www.hsa.gov.sg/docs/default-source/hprg-mdb/guidance-documents-for-medical-devices/regulatory-guidelines-for-software-medical-devices---a-life-cycle-approach_r2-%282022-apr%29-pub.pdf (accessed on 14 January 2026).
16. Pharmaceuticals and Medical Devices Agency (PMDA). Report on AI-Based Software as a Medical Device (SaMD). 28 August 2023. Available online: https://www.pmda.go.jp/files/000266100.pdf (accessed on 14 January 2026).
17. Ministry of Food and Drug Safety (MFDS). Guidance on the Review and Approval of Artificial Intelligence (AI)-Based Medical Devices. 20 July 2023. Available online: https://www.mfds.go.kr/eng/brd/m_40/view.do?seq=72627 (accessed on 14 January 2026).
18. National Medical Products Administration (NMPA). NMPA Announcement on Guidance for the Classification Defining of AI-Based Medical Software Products. 8 July 2021. Available online: https://english.nmpa.gov.cn/2021-07/08/c_660267.htm (accessed on 14 January 2026).
19. National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Available online: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf (accessed on 14 January 2026).
20. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. 28 June 2021. Available online: https://www.who.int/publications/i/item/9789240029200 (accessed on 14 January 2026).
21. Ning, Y.; Teixayavong, S.; Shang, Y.; Savulescu, J.; Nagaraj, V.; Miao, D.; Mertens, M.; Ting, D.S.W.; Ong, J.C.L.; Liu, M.; et al. Generative artificial intelligence and ethical considerations in health care: A scoping review and ethics checklist. *Lancet Digit. Health* **2024**, *6*, e848–e856. [CrossRef] [PubMed]
22. Du, J.; Tao, X.; Zhu, L.; Wang, H.; Qi, W.; Min, X.; Wei, S.; Zhang, X.; Liu, Q.; Du, Q. Development of a visualized risk prediction system for sarcopenia in older adults using machine learning: A cohort study based on CHARLS. *Front. Public Health* **2025**, *13*, 1544894. [CrossRef]
23. Weissman, G.E. Evaluation and regulation of artificial intelligence medical devices. *Annu. Rev. Biomed. Data Sci.* **2025**, *8*, 81–99. [CrossRef]
24. Maliha, G.; Gerke, S.; Cohen, I.G.; Parikh, R.B. Artificial intelligence and liability in medicine: Balancing safety and innovation. *Milbank Q.* **2021**, *99*, 629–647. [CrossRef]
25. Pham, T. Ethical and legal considerations in healthcare AI: Innovation and policy for safe and fair use. *R. Soc. Open Sci.* **2025**, *12*, 241873. [CrossRef]
26. Mahamadou, A.J.D.; Trotsyuk, A.A. Revisiting technical bias mitigation strategies. *Annu. Rev. Biomed. Data Sci.* **2025**, *8*, 287–303. [CrossRef]
27. Hanna, M.G.; Pantanowitz, L.; Jackson, B.; Palmer, O.; Visweswaran, S.; Pantanowitz, J.; Deebajah, M.; Rashidi, H.H. Ethical and bias considerations in artificial intelligence/machine learning. *Mod. Pathol.* **2025**, *38*, 100686. [CrossRef]
28. Dehghani, F.; Paiva, P.; Malik, N.; Lin, J.; Bayat, S.; Bento, M. Accuracy–fairness trade-off in ML for healthcare: A quantitative evaluation of bias mitigation strategies. *Inf. Softw. Technol.* **2025**, *188*, 107896. [CrossRef]
29. Stanley, E.A.M.; Wilms, M.; Forkert, N.D. Disproportionate subgroup impacts and other challenges of fairness in artificial intelligence for medical image analysis. In *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13755, pp. 14–25.
30. Vogt, Y. Disability and algorithmic fairness in healthcare: A narrative review. *J. Med. Artif. Intell.* **2025**, *8*, 56. [CrossRef]

31. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef] [PubMed]
32. Seyyed-Kalantari, L.; Zhang, H.; McDermott, M.B.A.; Chen, I.Y.; Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **2021**, *27*, 2176–2182. [CrossRef] [PubMed]
33. Gichoya, J.W.; Banerjee, I.; Bhimireddy, A.R.; Burns, J.L.; Celi, L.A.; Chen, L.-C.; Correa, R.; Dullerud, N.; Ghassemi, M.; Huang, S.-C.; et al. AI recognition of patient race in medical imaging: A modelling study. *Lancet Digit. Health* **2022**, *4*, e406–e414. [CrossRef]
34. Liu, M.; Ning, Y.; Teixayavong, S.; Mertens, M.; Xu, J.; Ting, D.S.W.; Cheng, L.T.-E.; Ong, J.C.L.; Teo, Z.L.; Tan, T.F.; et al. A translational perspective towards clinical AI fairness. *npj Digit. Med.* **2023**, *6*, 172. [CrossRef]
35. Long, Y.; Novak, L.; Walsh, C.G. Searching for value-sensitive design in applied health AI: A narrative review. *Yearb. Med. Inform.* **2024**, *33*, 75–82. [CrossRef] [PubMed]
36. Ferrara, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* **2024**, *6*, 3. [CrossRef]
37. Ministry of Economy, Trade and Industry (METI). Governance Guidelines for Implementation of AI Principles (Ver. 1.1). 28 January 2022. Available online: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf (accessed on 14 January 2026).
38. Ministry of Science and ICT (MSIT). MSIT Releases People-Centered "National AI Ethical Guidelines" Draft 27 November 2020. Available online: https://english.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=467&sCode=eng (accessed on 14 January 2026).
39. Ministry of Science and Technology of the People's Republic of China (MOST). Release of "Ethical Norms for New Generation Artificial Intelligence". 26 September 2021. Available online: https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html (accessed on 14 January 2026).
40. Radanliev, P. AI ethics: Integrating transparency, fairness, and privacy in AI development. *Appl. Artif. Intell.* **2025**, *39*, 2463722. [CrossRef]
41. Owens, K.; Griffen, Z.; Damaraju, L. Managing a "responsibility vacuum" in AI monitoring and governance in healthcare: A qualitative study. *BMC Health Serv. Res.* **2025**, *25*, 1043. [CrossRef]
42. Kale, A.U.; Hogg, H.D.J.; Pearson, R.; Glocker, B.; Golder, S.; Coombe, A.; Coombe, A.; Waring, J.; Liu, X.; Moore, D.J.; et al. Detecting algorithmic errors and patient harms for AI-enabled medical devices in randomized trials: Protocol. *JMIR Res. Protoc.* **2024**, *13*, e55707. [CrossRef] [PubMed]
43. Kumar, A.; Aelgani, V.; Vohra, R.; Gupta, S.K.; Bhagawati, M.; Paul, S.; Saba, L.; Suri, N.; Khanna, N.N.; Laird, J.R.; et al. Artificial intelligence bias in medical system designs: A systematic review. *Multimed. Tools Appl.* **2024**, *83*, 18005–18057. [CrossRef]
44. Zhang, J.; Zhang, Z.-M. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 135. [CrossRef] [PubMed]
45. Hasanzadeh, F.; Josephson, C.B.; Waters, G.; Adedinsewo, D.; Azizi, Z.; White, J.A. Bias recognition and mitigation strategies in healthcare AI. *npj Digit. Med.* **2025**, *8*, 12. [CrossRef]
46. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef]
47. Griffin, A.C.; Wang, K.H.; Leung, T.I.; Facelli, J.C. Recommendations to promote fairness and inclusion in biomedical AI. *J. Biomed. Inform.* **2024**, *152*, 104693. [CrossRef]
48. Nazer, L.H.; Zatarah, R.; Waldrip, S.; Ke, J.X.C.; Moukheiber, M.; Khanna, A.K.; Hicklen, R.S.; Moukheiber, L.; Moukheiber, D.; Ma, H.; et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit. Health* **2023**, *2*, e0000278. [CrossRef]
49. van den Heuvel, J.; Porter, A.; Kirkpatrick, E.; Verjans, J.; Reddy, S.; Freckelton, I. The silent partner: A narrative review of AI's impact on informed consent. *J. Law Med.* **2025**, *32*, 74–84.
50. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
51. Park, S.H.; Kim, Y.-H.; Lee, J.Y.; Yoo, S.; Kim, C.J. Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review. *Sci. Ed.* **2019**, *6*, 91–98. [CrossRef]
52. Alkhanbouli, R.; Almadhaani, H.M.A.; Alhosani, F.; Simsekler, M.C.E. The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 110. [CrossRef]
53. Ali, S.; Akhlaq, F.; Imran, A.S.; Kastrati, Z.; Daudpota, S.M.; Moosa, M. The enlightening role of explainable artificial intelligence in medical and healthcare domains: A systematic literature review. *Comput. Biol. Med.* **2023**, *166*, 107555. [CrossRef] [PubMed]
54. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable AI in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef]

55. Kapcia, M.; Eshkiki, H.; Duell, J.; Fan, X.; Zhou, S.; Mora, B. ExMed: An AI tool for experimenting explainable techniques on medical data analytics. In Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Virtual, 1–3 November 2021; pp. 841–845.
56. Falvo, F.R.; Cannataro, M. Explainability techniques for artificial intelligence models in medical diagnostic. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 6907–6913.
57. Phillips, V. A counterintuitive approach to explainable AI in healthcare: Balancing transparency, efficiency, and cost. *AI Soc.* **2025**, *40*, 5735–5741. [CrossRef]
58. Rudin, C. Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
59. McNamara, S.L.; Yi, P.H.; Lotter, W. The clinician–AI interface: Intended use and explainability in FDA-cleared AI devices for medical image interpretation. *npj Digit. Med.* **2024**, *7*, 80. [CrossRef]
60. Wenderott, K.; Krups, J.; Zaruchas, F.; Weigl, M. Effects of artificial intelligence implementation on efficiency in medical imaging: A systematic literature review and meta-analysis. *npj Digit. Med.* **2024**, *7*, 265. [CrossRef]
61. Doumard, E.; Aligon, J.; Escriva, E.; Excoffier, J.-B.; Monsarrat, P.; Soulé-Dupuy, C. A quantitative approach for the comparison of additive local explanation methods. *Inf. Syst.* **2023**, *114*, 102254. [CrossRef]
62. van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A.; Viergever, M.A. Explainable artificial intelligence in deep learning-based medical image analysis: A survey. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef]
63. Loh, H.W.; Ooi, C.P.; Seoni, S.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Int. J. Med. Inform.* **2022**, *157*, 104222. [CrossRef] [PubMed]
64. Mandava, R.; Vellela, S.S.; Malathi, N.; Haritha, K.; Gorintla, S.; Dalavai, L. Exploring the Role of XAI in Enhancing Predictive Model Transparency in Healthcare Risk Assessment 2025. In Proceedings of the International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 28–30 May 2025; pp. 1–5. [CrossRef]
65. Kaur, A.; Goyal, S. Explainable AI in Healthcare: Introduction. In *Explainable Artificial Intelligence in the Healthcare Industry*; Kumar, A., Ananth Kumar, T., Das, P., Sharma, C., Dubey, A.K., Eds.; Wiley-Scrivener: Hoboken, NJ, USA, 2025; pp. 307–323. Available online: https://www.wiley.com/en-us/Explainable+Artificial+Intelligence+in+the+Healthcare+Industry-p-9781394249268 (accessed on 14 January 2026).
66. Blahodelskyi, O. Systematic review: Innovative approaches in artificial intelligence development. *Nigerian J. Technol.* **2025**, *43*, 839–848. [CrossRef]
67. Aldhafeeri, F.M. Governing artificial intelligence in radiology: A systematic review of ethical, legal, and regulatory frameworks. *Diagnostics* **2025**, *15*, 2300. [CrossRef] [PubMed]
68. Nawawi, M.H.M.; Ishak, M.S.; Raes, R.F.A.; Razak, I.A.; Hasan, S.; Rahim, A.I.A. The intersection of quality improvement, artificial intelligence and patient safety in healthcare—Current applications, challenges and risks, and future directions: A scoping review. *J. Med. Artif. Intell.* **2025**, *8*, 57. [CrossRef]
69. Doolan, P.; Michopoulou, S.; Meades, R. IPEM topical report: Results of a 2024 UK survey of artificial intelligence in medical physics and clinical engineering. *Phys. Med. Biol.* **2025**, *70*, 14TR01. [CrossRef]
70. Hodges, B.D. Education and the adoption of AI in healthcare: "What is happening?". *Healthc. Pap.* **2025**, *22*, 39–43. [CrossRef]
71. Zhang, Y.; Li, J.; Meng, X.; Li, S.; Wang, H. Interpretation of sectoral standard AI medical device—Specific requirement for datasets: Color fundus images of diabetic retinopathy. *Med. J. Peking Union Med. Coll. Hosp.* **2025**, *16*, 916–924. [CrossRef]
72. Gornet, M.; Maxwell, W. The European approach to regulating AI through technical standards. *Internet Policy Rev.* **2024**, *13*, 1–27. [CrossRef]
73. Reddy, S. Global harmonization of AI-enabled software as a medical device regulation: Addressing challenges and unifying standards. *Mayo Clin. Proc. Digit. Health* **2024**, *3*, 100191. [CrossRef]
74. Chauhan, S.B.; Gaur, R.; Akram, A.; Singh, I. Artificial intelligence-driven insights for regulatory intelligence in medical devices: Evaluating EMA, FDA, and CDSCO frameworks. *Glob. Clin. Eng. J.* **2025**, *7*, 11–24. [CrossRef]
75. Vardas, E.P.; Marketou, M.; Vardas, P.E. Medicine, healthcare and the AI Act: Gaps, challenges and future implications. *Eur. Heart J. Digit. Health* **2025**, *6*, 833–839. [CrossRef]
76. Duffourc, M.N.; Gerke, S. The proposed EU directives for AI liability leave worrying gaps likely to impact medical AI. *npj Digit. Med.* **2023**, *6*, 77. [CrossRef]
77. Wong, A.; Otles, E.; Donnelly, J.P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penoza, C.; et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern. Med.* **2021**, *181*, 1065–1070. [CrossRef] [PubMed]
78. Lyell, D.; Wang, Y.; Coiera, E.; Magrabi, F. More than algorithms: Analysis of safety events involving ML-enabled medical devices reported to the FDA. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1227–1236. [PubMed]

79. Handley, J.L.; Krevat, S.A.; Fong, A.; Ratwani, R.M. Artificial intelligence-related safety issues associated with ML-enabled medical devices: An analysis of MAUDE reports. *npj Digit. Med.* **2024**, *7*, 351. [PubMed]
80. Abrisqueta-Costa, P.; García-Marco, J.A.; Gutiérrez, A.; Hernández-Rivas, J.Á.; Andreu-Lapiedra, R.; Arguello-Tomas, M.; Leiva-Farré, C.; López-Roda, M.D.; Callejo-Mellén, Á.; Álvarez-García, E.; et al. Real-world evidence on adverse events and healthcare resource utilization in patients with chronic lymphocytic leukaemia in Spain using natural language processing: The SRealCLL study. *Cancers* **2024**, *16*, 4004. [PubMed]
81. Sáez, C.; Ferri, P.; García-Gómez, J.M. Resilient artificial intelligence in health: Synthesis and research agenda toward next-generation trustworthy clinical decision support. *J. Med. Internet Res.* **2024**, *26*, e50295. [CrossRef]
82. Stogiannos, N.; Cuocolo, R.; D'Antonoli, A.T.; dos Santos, D.P.; Harvey, H.; Huisman, M.; Kocak, B.; Kotter, E.; Lekadir, K.; Shelmerdine, S.C.; et al. Recognising errors in AI implementation in radiology: A narrative review. *Eur. J. Radiol.* **2025**, *191*, 112311. [CrossRef]
83. Federico, C.A.; Trotsyuk, A.A. Biomedical data science, artificial intelligence, and ethics: Navigating challenges in the face of explosive growth. *Annu. Rev. Biomed. Data Sci.* **2024**, *7*, 1–14. [CrossRef] [PubMed]
84. Williamson, S.M.; Prybutok, V. Balancing privacy and progress: A review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. *Appl. Sci.* **2024**, *14*, 675. [CrossRef]
85. Amini, M.M.; Jesus, M.; Sheikholeslami, D.F.; Alves, P.; Benam, A.H.; Hariri, F. Artificial intelligence ethics and challenges in healthcare applications: A comprehensive review in the context of the European GDPR mandate. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1023–1035. [CrossRef]
86. European Parliament. EU AI Act: First Regulation on Artificial Intelligence. Available online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 18 October 2025).
87. Meszaros, J.; Minari, J.; Huys, I. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Front. Genet.* **2022**, *13*, 927721. [CrossRef]
88. Personal Data Protection Commission (PDPC). Data Protection Obligations (Personal Data Protection Act, Singapore). Available online: https://www.pdpc.gov.sg/overview-of-pdpa/the-legislation/personal-data-protection-act/data-protection-obligations (accessed on 14 January 2026).
89. Japanese Law Translation (Ministry of Justice, Japan). Act on the Protection of Personal Information. Available online: https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en (accessed on 14 January 2026).
90. Ministry of Government Legislation (Republic of Korea). Personal Information Protection Act (English Text, National Law Information Center). Available online: https://www.law.go.kr/LSW//lsInfoP.do?chrClsCd=010203&lsiSeq=213857&urlMode=engLsInfoR&viewCls=engLsInfoR (accessed on 14 January 2026).
91. Supreme People's Procuratorate (People's Republic of China). Personal Information Protection Law of the People's Republic of China. Available online: https://en.spp.gov.cn/2021-12/29/c_948419.htm (accessed on 14 January 2026).
92. Office for Civil Rights (OCR), U.S. Department of Health and Human Services. HIPAA Security Rule to Strengthen the Cybersecurity of Electronic Protected Health Information. Notice of Proposed Rulemaking; 6 January 2025. Available online: https://www.federalregister.gov/documents/2025/01/06/2024-30983/hipaa-security-rule-to-strengthen-the-cybersecurity-of-electronic-protected-health-information (accessed on 14 January 2026).
93. Office of the Privacy Commissioner of Canada. Privacy and Artificial Intelligence (AI). Available online: https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/ (accessed on 14 January 2026).
94. Office of the Australian Information Commissioner (OAIC). Guidance on Privacy and the Use of Commercially Available AI Products. 21 October 2024. Available online: https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/guidance-on-privacy-and-the-use-of-commercially-available-ai-products (accessed on 14 January 2026).
95. Wang, Y.; Liu, C.; Zhou, K.; Zhu, T.; Han, X. Towards regulatory generative AI in ophthalmology healthcare: A security and privacy perspective. *Br. J. Ophthalmol.* **2024**, *108*, 1349–1353. [CrossRef]
96. Wang, X.; Li, J.; Ding, X.; Zhang, H.; Sun, L. A survey of differential privacy techniques for federated learning. *IEEE Access* **2025**, *13*, 6539–6555. [CrossRef]
97. Shukla, S.; Rajkumar, S.; Sinha, A.; Esha, M.; Elango, K.; Sampath, V. Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity. *Sci. Rep.* **2025**, *15*, 12345. [CrossRef]
98. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
99. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2020**, *3*, 119. [CrossRef]
100. Wu, T.; Deng, Y.; Zhou, Q.; Chen, X.; Zhang, M. ADPHE-FL: Federated learning with adaptive differential privacy and homomorphic encryption. *Peer-to-Peer Netw. Appl.* **2025**, *18*, 210. [CrossRef]
101. Ouyang, J.; Han, R.; Zuo, X.; Cheng, Y.; Liu, C.H. Accuracy-aware differential privacy in federated learning of large transformer models. *J. Inf. Secur. Appl.* **2025**, *81*, 103844. [CrossRef]

102. Mahato, G.K.; Banerjee, A.; Chakraborty, S.K.; Gao, X.-Z. Privacy-preserving verifiable federated learning scheme using blockchain and homomorphic encryption. *Appl. Soft Comput.* **2024**, *156*, 111208. [CrossRef]
103. Li, K.; Lohachab, A.; Dumontier, M.; Urovi, V. Privacy preservation in blockchain-based healthcare data sharing: A systematic review. *Peer-to-Peer Netw. Appl.* **2025**, *18*, 302. [CrossRef] [PubMed]
104. Conduah, A.K.; Ofoe, S.; Siaw-Marfo, D. Data privacy in healthcare: Global challenges and solutions. *Digit. Health* **2025**, *11*, 20552076241234567. [CrossRef]
105. Alshohoumi, F. Privacy concerns of IoT medical applications: An empirical analysis of the current privacy policies under the GDPR. *Int. J. Electron. Healthc.* **2025**, *15*, 155–175. [CrossRef]
106. Puneeth, R.P.; Parthasarathy, G. Blockchain-based framework for privacy preservation and securing EHR with patient-centric access control. *Acta Inform. Pragensia* **2024**, *13*, 84195–84229. [CrossRef]
107. Chan, H.Y. A proportionality-by-design approach for mobile mental health and well-being applications. *Law Innov. Technol.* **2025**, *17*, 58–83. [CrossRef]
108. Rocher, L.; Hendrickx, J.M.; de Montjoye, Y.-A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **2019**, *10*, 3069. [CrossRef]
109. Schwarz, C.G.; Kremers, W.K.; Therneau, T.M.; Sharp, R.R.; Gunter, J.L.; Vemuri, P.; Arani, A.; Spychalla, A.J.; Kantarci, K.; Knopman, D.S.; et al. Identification of anonymous MRI research participants by face recognition. *N. Engl. J. Med.* **2019**, *381*, 1684–1686. [CrossRef]
110. Pati, S.; Kumar, S.; Varma, A.; Edwards, B.; Lu, C.; Qu, L.; Wang, J.J.; Lakshminarayanan, A.; Wang, S.-H.; Sheller, M.J.; et al. Privacy preservation for federated learning in health care. *Patterns* **2024**, *5*, 100974. [CrossRef]
111. Kaabachi, B.; Despraz, J.; Meurers, T.; Otte, K.; Halilovic, M.; Kulynych, B.; Prasser, F.; Raisaro, J.L. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digit. Med.* **2025**, *8*, 60. [CrossRef]
112. Sella, N.; Guinot, F.; Lagrange, N.; Albou, L.-P.; Desponds, J.; Isambert, H. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *npj Digit. Med.* **2025**, *8*, 49. [CrossRef]
113. Habli, I.; Lawton, T.; Porter, Z. Artificial intelligence in health care: Accountability and safety. *Bull. World Health Organ.* **2020**, *98*, 251–256. [CrossRef] [PubMed]
114. Chan, B. Applying a common enterprise theory of liability to clinical AI systems. *Am. J. Law Med.* **2021**, *47*, 351–385. [CrossRef]
115. Daye, D.; Wiggins, W.F.; Lungren, M.P.; Alkasab, T.; Kottler, N.; Allen, B.; Roth, C.J.; Bizzo, B.C.; Durniak, K.; Brink, J.A.; et al. Implementation of Clinical Artificial Intelligence in Radiology: Who Decides and How? *Radiology* **2022**, *305*, 555–563. [CrossRef] [PubMed]
116. Bedoya, A.D.; Economou-Zavlanos, N.J.; Goldstein, B.A.; Young, A.; Jelovsek, J.E.; O'bRien, C.; Parrish, A.B.; Elengold, S.; Lytle, K.; Balu, S.; et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1631–1636. [CrossRef]
117. Di Palma, G.; Scendoni, R.; Tambone, V.; Alloni, R.; De Micco, F. Integrating enterprise risk management to address AI-related risks in healthcare: Strategies for effective risk mitigation and implementation. *J. Healthc. Risk Manag.* **2025**, *44*, 25–33. [CrossRef]
118. Contaldo, M.T.; Pasceri, G.; Vignati, G.; Bracchi, L.; Triggiani, S.; Carrafiello, G. AI in radiology: Navigating medical responsibility. *Diagnostics* **2024**, *14*, 1506. [CrossRef] [PubMed]
119. Chan, G.K.Y. AI in healthcare: Regulatory guidelines and judge-made negligence principles for AI implementers. *Med. Law Int.* **2025**. [CrossRef]
120. Bottomley, D.; Thaldar, D. Liability for harm caused by AI in healthcare: An overview of the core legal concepts. *Front. Pharmacol.* **2023**, *14*, 1297353. [CrossRef]
121. Naidoo, T. Overview of AI regulation in healthcare: A comparative study of the EU and South Africa. *S. Afr. J. Bioeth. Law* **2024**, *17*, e2294. [CrossRef]
122. Chau, M.; Rahman, M.G.; Debnath, T. From black box to clarity: Strategies for effective AI informed consent in healthcare. *Artif. Intell. Med.* **2025**, *167*, 103169. [CrossRef] [PubMed]
123. Srinivasu, P.N.; Sandhya, N.; Jhaveri, R.H.; Raut, R. From blackbox to explainable AI in healthcare: Existing tools and case studies. *Mob. Inf. Syst.* **2022**, *1*, 8167821. [CrossRef]
124. Price, W.N., II; Gerke, S.; Cohen, I.G. Potential liability for physicians using artificial intelligence. *JAMA* **2019**, *322*, 1765–1766. [CrossRef]
125. Hillis, J.M.; Visser, J.J.; Cliff, E.R.S.; van der Geest–Aspers, K.; Bizzo, B.C.; Dreyer, K.J.; Adams-Prassl, J.; Andriole, K.P. The lucent yet opaque challenge of regulating artificial intelligence in radiology. *npj Digit. Med.* **2024**, *7*, 69. [CrossRef]
126. Carvalho, E.; Mascarenhas, M.; Pinheiro, F.; Correia, R.; Balseiro, S.; Barbosa, G.; Guerra, A.; Oliveira, D.; Moura, R.; dos Santos, A.M.; et al. Predetermined change control plans: Guiding principles for advancing safe, effective, and high-quality AI-ML technologies. *JMIR AI* **2025**, *4*, e76854. [CrossRef]
127. Weerakoon, A.T.; Girdis, T.; Peters, O. Artificial intelligence in Australian dental and general healthcare: A scoping review. *Aust. Dent. J.* **2025**, *online ahead of print*. [CrossRef] [PubMed]

128. Warraich, H.J.; Tazbaz, T.; Califf, R.M. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA* **2025**, *333*, 241–247. [CrossRef]
129. Stetson, P.D.; Choy, J.; Summerville, N.; Baldwin-Medsker, A.; Mak, J.; Chatterjee, A.; Kim, K.; Kumar, C.; Samedy, P.; Halperin, J.; et al. Responsible artificial intelligence governance in oncology. *npj Digit. Med.* **2025**, *8*, 407. [CrossRef]
130. Rozenblit, L.; Price, A.; Solomonides, A.; Joseph, A.L.; Koski, E.; Srivastava, G.; Labkoff, S.; Bray, D.; Lopez-Gonzalez, M.; Singh, R.; et al. Toward responsible AI governance: Balancing multi-stakeholder perspectives on AI in healthcare. *Int. J. Med. Inform.* **2025**, *203*, 106015. [CrossRef]
131. Waeiss, Q.; Cho, M.K. An ecosystem approach to governing commercial actors in healthcare AI. *Policy Stud.* **2025**, 1–14. [CrossRef]
132. Salwei, M.E.; Davis, S.E.; Reale, C.; Novak, L.L.; Walsh, C.G.; Beebe, R.; Nelson, S.; Sundrani, S.; Rose, S.; Wright, A.; et al. Human-Centered Design of an Artificial Intelligence Monitoring System: The Vanderbilt Algorithmovigilance Monitoring and Operations System. *JAMIA Open* **2025**, *8*, ooaf136. [CrossRef]
133. Balendran, A.; Benchoufi, M.; Evgeniou, T.; Ravaud, P. Algorithmovigilance, lessons from pharmacovigilance. *npj Digit. Med.* **2024**, *7*, 270. [CrossRef]
134. Sridharan, K.; Sivaramakrishnan, G. Leveraging artificial intelligence to detect ethical concerns in medical research: A case study. *J. Med. Ethics* **2025**, *51*, 126–134. [CrossRef]
135. Friesen, P.; Douglas-Jones, R.; Marks, M.; Pierce, R.; Fletcher, K.; Mishra, A.; Lorimer, J.; Véliz, C.; Hallowell, N.; Graham, M.; et al. Governing AI-driven health research: Are IRBs up to the task? *Ethics Hum. Res.* **2021**, *43*, 35–42. [CrossRef] [PubMed]
136. Anderson, E.E.; Johnson, A.; Lynch, H.F. Inclusive, engaged, and accountable institutional review boards. *Account. Res.* **2024**, *31*, 1287–1295. [CrossRef] [PubMed]
137. Labkoff, S.; Oladimeji, B.; Kannry, J.; Solomonides, A.; Leftwich, R.; Koski, E.; Joseph, A.L.; Lopez-Gonzalez, M.; Fleisher, L.A.; Nolen, K.; et al. Toward a responsible future: Recommendations for AI-enabled clinical decision support. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2730–2739. [CrossRef]
138. Bignami, E.G.; Russo, M.; Semeraro, F.; Bellini, V. The European Union AI Act in an era of global uncertainty. *JMIR AI* **2025**, *4*, e75527. [CrossRef] [PubMed]
139. NHS Digital. Artificial Intelligence (AI) Buyer's Guide Assessment Template. Available online: https://digital.nhs.uk/services/ai-knowledge-repository/develop-ai/a-buyers-guide-to-ai-in-health-and-care/assessment-template (accessed on 14 January 2026).
140. Khan, S.D.; Hoodbhoy, Z.; Raja, M.H.R.; Kim, J.Y.; Hogg, H.D.J.; Manji, A.A.A.; Gulamali, F.; Hasan, A.; Shaikh, A.; Tajuddin, S.; et al. Frameworks for procurement, integration, monitoring, and evaluation of artificial intelligence tools in clinical settings: A systematic review. *PLoS Digit. Health* **2024**, *3*, e0000514. [CrossRef]
141. Bidenko, N.V.; Stuchynska, N.V.; Palamarchuk, Y.V.; Matviienko, M.M. Integrating artificial intelligence in healthcare practice: Challenges and future prospects. *Wiad. Lek.* **2025**, *78*, 1199–1205. [CrossRef]
142. Kim, J.Y.; Hasan, A.; Kellogg, K.C.; Ratliff, W.; Murray, S.G.; Suresh, H.; Valladares, A.; Shaw, K.; Tobey, D.; Vidal, D.E.; et al. Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. *PLoS Digit. Health* **2024**, *3*, e0000390. [CrossRef]
143. Arnaout, A.; Gill, P.; Virani, A.; Flatt, A.; Prodan-Balla, N.; Byres, D.; Stowe, M.; Saremi, A.; Coss, M.; Tatto, M.; et al. Shaping the future of healthcare in British Columbia: Establishing provincial clinical governance for responsible deployment of artificial intelligence tools. *Healthc. Manag. Forum* **2024**, *37*, 320–328. [CrossRef]
144. Torkilsheyggi, A. Flexibility first, then standardize: A strategy for growing inter-departmental systems. *Stud. Health Technol. Inform.* **2015**, *216*, 477–481.
145. Lukkien, D.R.M.; Nap, H.H.; Peine, A.; Minkman, M.M.N.; Moors, E.H.M.; Boon, W.P.C. Responsible scaling of artificial intelligence in healthcare: Standardization meets customization. *Ethics Inf. Technol.* **2025**, *27*, 34. [CrossRef]
146. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **2022**, *377*, e070904. [CrossRef] [PubMed]
147. Sakly, H.; Guetari, R.; Kraiem, N. Deployment and Continuous Integration of AI in Healthcare. In *Scalable Artificial Intelligence for Healthcare: Advancing AI Solutions for Global Health Challenges*; CRC Press: Boca Raton, FL, USA, 2025; pp. 95–112.
148. Brady, A.P.; Allen, B.; Chong, J.; Kotter, E.; Kottler, N.; Mongan, J.; Oakden-Rayner, L.; dos Santos, D.P.; Tang, A.; Wald, C.; et al. Developing, purchasing, implementing and monitoring AI tools in radiology: Practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. *J. Med. Imaging Radiat. Oncol.* **2024**, *68*, 7–26. [CrossRef]
149. Rajendra, J.B.; Thuraisingam, A.S. The role of explainability and human intervention in AI decisions: Jurisdictional and regulatory aspects. *Inf. Commun. Technol. Law* **2025**, *34*, 1–32. [CrossRef]
150. Wang, Y.; Li, N.; Chen, L.; Wu, M.; Meng, S.; Dai, Z.; Zhang, Y.; Clarke, M. Guidelines, consensus statements, and standards for the use of artificial intelligence in medicine: Systematic review. *J. Med. Internet Res.* **2023**, *25*, e46089. [CrossRef] [PubMed]
151. Hou, J.; Cheng, X.; Liao, J.; Zhang, Z.; Wang, W. Ethical concerns of AI in healthcare: A systematic review of qualitative studies. *Nurs. Ethics* **2025**, *online ahead of print*. [CrossRef]

152. Čartolovni, A.; Tomičić, A.; Lazić Mosler, E. Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. *Int. J. Med. Inform.* **2022**, *161*, 104738. [CrossRef]
153. Cestonaro, C.; Delicati, A.; Marcante, B.; Caenazzo, L.; Tozzo, P. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: A systematic review. *Front. Med.* **2023**, *10*, 1305756. [CrossRef] [PubMed]