



From Regulatory Authorization to Clinical Accountability: A Control Preventability Framework for Healthcare AI

Tony T. Williams¹, Amna Jatoi², Sazain Malik³, Attique ur Rehman⁴

¹ Department of Health and Human Services, Ashford University-UAGC

Email: williamscaddola@gmail.com

² IBHM, Dow University of Health Sciences, Karachi, Pakistan

Email: amna.jatoi@duhs.edu.pk

³ Department of Healthcare and Hospital Management, Riphah International University, Rawalpindi, Pakistan

Email: sazainmalik09@gmail.com

⁴ Iqra University, Karachi, Pakistan

Email: chaudhryattique@gmail.com

ARTICLE INFO

Submission:

July 19, 2026

Revised:

August 08, 2026

Accepted:

August 24, 2026

Available Online:

September 04, 2026

Keywords:

*artificial intelligence,
healthcare, clinical decision
support, medical devices,
regulation, accountability,
patient safety*

Corresponding author:

Attique ur Rehman

Email:

chaudhryattique@gmail.com

ABSTRACT

The integration of artificial intelligence into healthcare has accelerated dramatically, with the FDA authorizing 1,430 AI/ML-enabled medical devices by the end of 2025 including 331 authorizations in that year alone yet the evidentiary foundation for many deployed tools remains limited, and accountability frameworks have not kept pace with technological adoption. This integrative evidence – regulatory analysis synthesizes FDA authorization data, selected high-relevance clinical evidence from randomized trials and systematic reviews, and primary regulatory and governance materials from the FDA, European Union, and World Health Organization. The analysis reveals that among 1,357 FDA-cleared AI devices identified through December 2025, only 2.5% were linked to registered prospective trials and only 0.2% had identified evidence evaluating patient-centered outcomes. A randomized clinical trial involving 44 physicians found that exposure to erroneous large language model recommendations reduced diagnostic reasoning accuracy by 14.0 percentage points (95% CI, –18.9 to –9.1; $P < 0.0001$) despite prior AI literacy training, while a prospective observational study of a sepsis early warning system found provider confirmation of alerts was associated with an adjusted absolute mortality reduction of 3.3% (95% CI, 1.7% – 5.1%). A systematic review and meta-analysis of 83 studies found no significant overall difference in diagnostic performance between generative AI and physicians, though generative AI performed significantly worse than expert physicians. Under the July 2026 Digital Omnibus, AI systems incorporated into medical devices remain in Section A of Annex I of the EU AI Act, subject to the high-risk framework integrated with existing medical-device regulations, with Chapter III obligations deferred to August 2028. The available evidence suggests that publicly identifiable rigorous prospective clinical validation is uncommon among FDA-authorized AI devices, and accountability frameworks remain fragmented, creating a risk that clinicians may bear disproportionate responsibility for harms arising from systems they do not fully control. This paper proposes a Control – Preventability Principle that allocates accountability according to control, foreseeability, preventability, and meaningful oversight capacity. Closing the evidence – accountability gap requires strengthened evidentiary standards, lifecycle governance, and clear allocation of responsibility across developers, deploying institutions, and clinicians, with accountability following control over preventable risk rather than merely the physical location of the final clinical decision.

Introduction

Artificial intelligence is reshaping healthcare delivery across diagnostic imaging, clinical decision support, drug discovery, and administrative automation. By the end of 2025, the FDA AI/ML-enabled medical-device list contained 1,430 authorized devices, reflecting rapid expansion of the regulated AI-device market (FDA, n.d.; Lee & Youn, 2026). Annual authorizations grew steadily, with 331 authorizations in 2025 representing the highest annual total to date (Lee & Youn, 2026).

Yet significant gaps persist between authorization and clinically meaningful validation. A 2026 systematic analysis of 1,357 FDA-cleared AI devices (identified through December 5, 2025) found that only 2.5% were linked to registered prospective trials and only 0.2% (three devices) had identified evidence evaluating patient-centered outcomes such as mortality, morbidity, or readmissions (Abulibdeh et al., 2026). The authors characterized the findings as evidence of a substantial gap between regulatory clearance and patient-centered clinical validation.

Existing scholarship tends to examine these problems in separate silos: regulatory authorization focuses on market access and safety requirements; clinical studies focus on diagnostic or patient outcomes; human-AI interaction research examines automation bias and oversight; and legal scholarship focuses on liability doctrines. What remains underdeveloped is an integrated framework connecting evidentiary sufficiency and operational oversight to responsibility allocation across the AI lifecycle.

This paper addresses three questions: (1) How large is the gap between FDA authorization of healthcare AI devices and prospective clinical validation? (2) What does emerging clinical evidence reveal about the benefits and risks of AI-assisted decision-making and human oversight? (3) How should accountability be allocated when AI-related clinical harm arises across developers, vendors, healthcare organizations, and clinicians with differing degrees of control and preventability?

This paper makes three contributions. First, it distinguishes regulatory authorization from clinically meaningful validation by integrating FDA authorization data with emerging prospective and randomized evidence. Second, it reframes human oversight as an operational capability requiring epistemic capacity, cognitive space, decisional authority, and intervention effectiveness rather than merely the presence of a clinician in the workflow. Third, it proposes the Control-Preventability Principle as a normative framework for distributing accountability according to control, foreseeability, preventability, and oversight capacity across the AI lifecycle. The framework does not establish a new doctrine of legal liability; rather, it provides a governance model that can inform regulatory design, institutional policy, contractual allocation, and future legal analysis.

The central proposition of this paper is that clinical accountability should follow control over preventable risk rather than merely the physical location of the final clinical decision. This proposition does not eliminate professional responsibility; it prevents professional responsibility from becoming a default substitute for developer, vendor, institutional, or regulatory control.

The novelty of the proposed framework lies not in introducing control, foreseeability, preventability, or human oversight as isolated concepts. Each has established roots in liability law, patient-safety governance, or human-AI interaction research. Rather, this paper integrates these dimensions into a single incident-attribution framework specifically designed for AI-mediated clinical harm, where the actor making the final clinical decision may not control the upstream source of the risk. The framework's distinctive contribution is to connect the operational conditions of meaningful human oversight epistemic capacity, cognitive space, decisional authority, and intervention effectiveness with the distribution of control over AI-related risks across developers, vendors, healthcare organizations, clinicians, and regulators. This integration challenges decision-location-based attribution by distinguishing the actor who makes the final clinical decision from the actors who control, create, configure, or can prevent the underlying risk. The framework does not replace existing negligence or product-liability doctrines; rather, it provides a structured analytical approach for incident investigation and governance attribution that can inform, but does not determine, legal liability.

Analytical Approach and Evidence Sources

Study Design

This paper employs an integrative evidence-regulatory analysis a structured synthesis combining regulatory authorization data, selected empirical clinical evidence, and primary regulatory and governance documents to examine the relationship between authorization, clinical validation, oversight, and accountability. The manuscript is a structured integrative analysis rather than a systematic review, scoping review, or meta-analysis. The empirical evidence streams were used to establish the evidentiary and governance problem; the accountability framework and recommendations constitute a normative synthesis derived from those findings and existing regulatory and governance principles rather than an empirical test of a causal model.

Evidence Streams and Targeted Evidence Selection

The analysis combines three complementary evidence streams:

- **Stream 1. FDA Authorization Landscape:** Analysis of publicly available FDA AI/ML-enabled medical-device authorization data through December 31, 2025, including year-over-year trends, specialty distribution, and regulatory pathways.
- **Stream 2. Clinical Evidence Analysis:** Review of high-relevance clinical evidence from randomized trials, prospective implementation studies, and systematic reviews and meta-analyses, with emphasis on studies evaluating diagnostic performance, patient outcomes, and physician–AI interaction.
- **Stream 3. Regulatory and Governance Analysis:** Examination of primary regulatory and governance materials from the FDA, European Union (EU AI Act and Omnibus Amendment), World Health Organization, and professional medical bodies.

Evidence Identification and Selection

Clinical evidence was identified through targeted searches of PubMed, Web of Science, and Google Scholar. Search terms included combinations of "artificial intelligence," "clinical AI," "LLM," "large language model," "diagnostic reasoning," "clinical decision support," "automation bias," "prospective," "randomized," and "patient outcomes." The search was limited to studies published between 2021 and 2026, with earlier studies retained where they provided foundational external-validation evidence.

Studies were included if they directly informed one of the three analytical domains: patient outcomes, diagnostic performance, or human–AI oversight. Priority was given to randomized controlled trials, prospective implementation studies with large sample sizes, and systematic reviews with meta-analyses. Studies were excluded if they were technical-only without clinical relevance, were not peer-reviewed, or did not directly compare AI-assisted with non-AI-assisted clinical decision-making.

The final evidence set comprised one FDA authorization analysis, one device-level evidence census, and five clinical evidence sources: (1) the Abulibdeh et al. (2026) systematic device-level analysis (authorization–validation gap); (2) the Qazi et al. (2026) randomized clinical trial (automation bias); (3) the Takita et al. (2025) systematic review and meta-analysis (diagnostic performance); (4) the Henry et al. (2022) prospective implementation cohort (TREWS adoption and treatment timing); (5) the Adams et al. (2022) prospective multicenter cohort (TREWS patient outcomes); and (6) the Wong et al. (2021) external validation study (model transportability). These were selected because they represent different evidentiary functions regulatory evidence, experimental evidence, comparative effectiveness, implementation, patient outcomes, and external validation rather than as a representative sample of the entire clinical-AI evidence base.

Because the clinical evidence stream was purposive rather than systematic, no formal screening denominator was calculated. Studies were selected iteratively based on direct relevance to the four analytical dimensions, methodological strength, prospective or experimental design, and ability to illuminate responsibility or oversight.

Table 1: Evidence Sources and Analytical Role

Evidence Stream	Source	Evidence Type	Analytical Purpose
FDA Authorization	FDA AI/ML device list; Lee & Youn (2026)	Regulatory dataset	Authorization volume, specialty distribution, regulatory pathways
Evidence Gap	Abulibdeh et al. (2026)	Systematic device-level analysis	Authorization–clinical validation gap
Automation Bias	Qazi et al. (2026) NEJM AI	Randomized clinical trial	Human oversight vulnerability
Clinical Performance	Takita et al. (2025) npj Digital Medicine	Systematic review/meta-analysis	AI–physician diagnostic comparison
Sepsis Implementation	Henry et al. (2022)	Prospective implementation cohort	Workflow/adoption
Sepsis Outcomes	Adams et al. (2022)	Prospective multicenter cohort	Patient outcomes association
External Validation	Wong et al. (2021)	External validation study	Model transportability

Evidence Stream	Source	Evidence Type	Analytical Purpose
Regulatory Frameworks	EU AI Act; Digital Omnibus; FDA guidance	Primary regulatory materials	Governance architecture
Governance Readiness	WHO (2026)	Governance statement	Institutional readiness

Data Sources

FDA authorization data were drawn from publicly available FDA AI/ML-enabled medical-device lists, with cumulative authorizations tracked through December 31, 2025, supplemented by analysis from peer-reviewed cross-sectional studies (FDA, n.d.; Lee & Youn, 2026). Clinical evidence was selected for high relevance, including the Qazi et al. (2026) NEJM AI randomized clinical trial on automation bias, the TREWS sepsis early warning system prospective multi-site cohort studies, and a systematic review and meta-analysis of generative AI performance compared with physicians (Abulibdeh et al., 2026; Adams et al., 2022; Henry et al., 2022; Qazi et al., 2026; Takita et al., 2025; Wong et al., 2021). Regulatory data were drawn from the EU AI Act (Regulation [EU] 2024/1689), the AI Omnibus Amendment (Regulation [EU] 2026/1744 of 8 July 2026), FDA guidance documents, and WHO ethics and governance frameworks (European Parliament, 2024; European Parliament, 2026; FDA, 2026; World Health Organization, 2026).

Data-Source Distinction: The FDA's cumulative AI/ML-enabled device list contained 1,430 devices by December 31, 2025. The 1,357-device evidence analysis reported by Abulibdeh et al. (2026) used a December 5, 2025 cutoff and examined FDA-cleared devices meeting that study's inclusion criteria. The two figures therefore should not be treated as interchangeable denominators.

Analytical Integration

Evidence was mapped against four analytical dimensions: (1) authorization patterns and volume; (2) evidence of clinical effectiveness and patient outcomes; (3) human oversight capacity and its operational conditions; and (4) accountability allocation and regulatory frameworks. The normative Control-Preventability Principle was derived by comparing, across the AI lifecycle, who creates the relevant risk, who controls it, who can foresee it, who can prevent or mitigate it, and who can meaningfully intervene when harm is imminent or occurring.

Terminology and Conceptual Definitions

Throughout this manuscript, "FDA-authorized" is used as the umbrella term for AI/ML-enabled devices cleared through 510(k), authorized through De Novo classification, or approved through Premarket Approval (PMA). "FDA-cleared" refers specifically to devices authorized through the 510(k) pathway. "FDA-approved" refers to devices authorized through the PMA pathway.

Accountability is used in a broader governance sense encompassing responsibility for prevention, oversight, monitoring, correction, and risk management; it is not used synonymously with civil or criminal legal liability.

The analysis operates at three levels: (1) device-level (FDA authorizations); (2) clinical-study level (RCTs, prospective cohorts, meta-analyses); and (3) governance/actor level (developer, vendor, healthcare organization, clinician, regulator). These levels are distinguished throughout the analysis.

Limitations

This analysis is not intended as a formal systematic review of all clinical AI literature. The selected clinical studies were purposively chosen for analytical relevance and methodological strength and therefore should not be interpreted as a representative sample of the broader clinical-AI evidence base. The study is limited by the available published evidence; many AI devices lack any published validation studies. The clinical trial evidence is concentrated in specific domains and may not generalize to all AI applications. Regulatory data represent authorizations rather than evidence of clinical effectiveness. The evidence gap analysis is based on a single systematic review with a December 5, 2025 cutoff and may not capture all published studies.

Results

FDA AI Device Authorization Landscape

The FDA's AI-enabled medical-device authorization has grown substantially over the past decade. By the end of 2025, the FDA AI/ML-enabled medical-device list contained 1,430 authorized devices, including 331 authorizations during 2025 the highest annual total to date (FDA, n.d.; Lee & Youn, 2026). Among the 1,430 devices with final authorization decisions, the 2026 longitudinal analysis reported that in the 2023–2025 period, radiology accounted for 77.5% of authorizations, followed by cardiovascular (7.8%) and neurology (5.7%) (Lee & Youn, 2026). Radiology's share has gradually declined from its peak of 85% between 2020–2022 to its current 77.5% range, with recent FDA activity signaling gradual diversification into cardiology, ophthalmology, and neurology applications (Lee & Youn, 2026).

Most authorizations occurred through the 510(k) pathway, which accounts for 96.2% of authorizations (Lee & Youn, 2026). The predominance of the 510(k) pathway is consequential, but it should not be interpreted as meaning that FDA clearance involves no assessment of safety or effectiveness. The 510(k) pathway is comparative: FDA determines whether the proposed device is substantially equivalent to a legally marketed predicate and reviews safety and performance information appropriate to the device (Lee & Youn, 2026). Unlike PMA, however, 510(k) does not ordinarily require an independent demonstration of safety and effectiveness of the type associated with PMA approval, and clinical data are not required for most 510(k) submissions (Lee & Youn, 2026). The resulting regulatory distinction is therefore between comparative regulatory authorization and independently demonstrated clinical effectiveness, rather than between regulation and no evidence. Only 17 devices (1.2%) were authorized through the Premarket Approval (PMA) pathway, which generally entails a more extensive demonstration of reasonable assurance of safety and effectiveness (Lee & Youn, 2026).

The Authorization–Validation Gap

A systematic analysis published recently examined the evidentiary foundation of FDA-cleared AI devices through December 5, 2025 (Abulibdeh et al., 2026). Importantly, this analysis used a December 5, 2025 cutoff and therefore should not be interpreted as an assessment of the evidentiary status of every device subsequently authorized by the FDA.

Of the 1,357 AI devices analyzed, only 34 devices (2.5%) were linked to registered prospective trials (Abulibdeh et al., 2026). Only 12 devices (0.9%) had posted trial results, and only 12 devices (0.9%) had peer-reviewed publications (Abulibdeh et al., 2026). Most strikingly, only three devices (0.2%) had identified evidence evaluating patient-centered outcomes such as mortality, morbidity, or readmissions (Abulibdeh et al., 2026). Among devices with identified studies, 62% employed observational designs with small, homogenous cohorts, limited subgroup analyses, and frequent exclusion of vulnerable populations (Abulibdeh et al., 2026). The authors characterized the findings as evidence of a substantial gap between regulatory clearance and patient-centered clinical validation.

Regulatory authorization establishes that a product satisfies the applicable regulatory requirements; it does not necessarily establish robust evidence of patient-centered clinical effectiveness in real-world populations. This gap should not be interpreted as evidence that authorized AI devices are inherently unsafe; nor does absence of identified prospective or patient-centered evidence necessarily establish absence of such evidence, because the analysis depended on publicly identifiable trial registrations and publications. Rather, it indicates that the evidentiary threshold and regulatory pathway used to authorize many devices do not necessarily generate the prospective, patient-centered evidence required to establish effectiveness across real-world clinical settings.

Clinical Trial Evidence: Key Findings

Automation Bias in LLM-Assisted Diagnostic Reasoning

The Qazi et al. (2026) randomized clinical trial provides one of the most directly relevant randomized studies of automation bias in LLM-assisted diagnostic reasoning. The single-blind trial involved 44 physicians in Pakistan who completed a 20-hour AI literacy training program. Participants diagnosed six clinical vignettes, with the control group receiving error-free ChatGPT-4o recommendations and the treatment group receiving recommendations with deliberately introduced errors in three of six cases.

Physicians exposed to erroneous LLM suggestions had a mean diagnostic reasoning accuracy of 73.3%, representing an adjusted 14.0-percentage-point reduction compared with the control group's 84.9% (95% CI, –18.9 to –9.1; $P < 0.0001$) (Qazi et al., 2026). Top-choice accuracy was 76.1% in the treatment group versus 90.5% in the control group, resulting in an adjusted 18.3-

percentage-point difference (95% CI, -26.6 to -10.0) (Qazi et al., 2026). The authors concluded that physicians demonstrate substantial automation bias when exposed to erroneous LLM recommendations, even with voluntary consultation and prior AI literacy training (Qazi et al., 2026). The trial provides experimental evidence that nominal human involvement does not necessarily prevent AI-induced diagnostic error under the conditions studied.

Systematic Review and Meta-Analysis: Generative AI vs. Physicians

A 2025 systematic review and meta-analysis of 83 studies evaluated diagnostic performance comparing generative AI with physicians (Takita et al., 2025). The prospectively registered review (PROSPERO: CRD42023494733) included studies published through June 2024 (Takita et al., 2025). Random-effects meta-analyses compared diagnostic performance.

No statistically significant difference in diagnostic performance between generative AI and physicians was found overall; physician accuracy was 9.9 percentage points higher than AI (95% CI -2.3 to 22.0; $p=0.10$) (Takita et al., 2025). However, generative AI performed significantly worse than expert physicians, whose accuracy was 15.8 percentage points higher (95% CI 4.4-27.1; $p=0.007$) (Takita et al., 2025). In subgroup analyses of the five randomized controlled trials included, LLM-assisted physicians showed improved diagnostic performance relative to physicians alone (Takita et al., 2025). However, this finding is based on a small number of trials and should be interpreted cautiously given the substantial overall risk of bias in the included studies. Interpretation is further constrained by the evidence quality: 63 of 83 included studies (76%) were assessed as having high overall risk of bias (Takita et al., 2025). The authors concluded that generative AI does not consistently outperform physicians across all clinical tasks, and performance depends heavily on clinical setting and the expertise of the comparator group (Takita et al., 2025). The potential benefit of LLM assistance requires further investigation in prospective, real-world settings (Takita et al., 2025).

TREWS Sepsis early warning system

The TREWS (Targeted Real-time Early Warning System) AI-powered clinical decision support system, developed at Johns Hopkins, is among the more extensively evaluated AI-enabled clinical decision-support systems reported in prospective clinical studies. Importantly, this is a prospective observational cohort study, not a randomized clinical trial. Two key studies provide complementary evidence.

During deployment across five hospitals, TREWS monitored 590,736 patients; the provider-adoption analysis examined 9,805 retrospectively identified sepsis cases (Henry et al., 2022). The provider-adoption and treatment-timing study found that TREWS demonstrated approximately 82% sensitivity in the deployment cohort (Henry et al., 2022). Among patients for whom providers confirmed alerts within three hours, provider evaluation was associated with a 1.85-hour reduction in median time to first antibiotic order. Approximately 89% of alerts received a provider evaluation (Henry et al., 2022).

The prospective multi-site patient-outcomes study, involving 6,877 patients with sepsis who were identified by the alert before initiation of antibiotic therapy, found that among patients whose TREWS alerts were confirmed by providers within three hours, provider confirmation was associated with an adjusted absolute mortality reduction of 3.3% (95% CI, 1.7%-5.1%) and an adjusted relative reduction of 18.7% (95% CI, 9.4%-27.0%), with greater benefits (up to 4.5% absolute reduction) in high-risk patients (Adams et al., 2022). Because TREWS was evaluated in a prospective observational cohort rather than a randomized trial, the mortality association cannot by itself establish that the AI system caused the observed reduction; residual confounding, clinician selection, and differential alert confirmation are alternative explanations (Adams et al., 2022).

Regulatory Landscape: The EU AI act and Omnibus

The EU AI Act (Regulation [EU] 2024/1689) entered into force in 2024 as the EU's comprehensive horizontal AI regulatory framework (European Parliament, 2024). The July 2026 Digital Omnibus (Regulation [EU] 2026/1744 of 8 July 2026) significantly amended the Act's treatment of AI systems incorporated into products governed by specified Union harmonisation legislation (European Parliament, 2026). The Omnibus confirmed that medical devices and in vitro diagnostic medical devices remain in Section A of Annex I, meaning they are subject to the high-risk AI framework integrated with the MDR/IVDR conformity assessment (European Parliament, 2026). Under the amended Article 113, the application of Chapter III, Sections 1-3 for such systems is deferred from August 2, 2027 to August 2, 2028 (European Parliament, 2026). This deferral addresses the time needed for harmonised standards, guidance, and conformity-assessment infrastructure to be developed, but does not exempt medical-device AI from the high-risk framework. For AI incorporated into regulated medical products, the precise timing and

mechanics of compliance depend on the system's classification under Article 6 and the relevant sectoral conformity-assessment framework under the MDR/IVDR (European Parliament, 2026).

Regulation (EU) 2024/1689, Article 14 requires effective human oversight for high-risk AI systems, including understanding capabilities and limitations, detecting anomalies, and overriding outputs (European Parliament, 2024). For AI systems incorporated into medical devices and IVDs, which remain in Section A of Annex I, the high-risk framework applies integrated with the MDR/IVDR conformity assessment. Under Article 113 as amended by the July 2026 Omnibus, the application of Chapter III, Sections 1–3 for such systems is deferred to August 2, 2028 (European Parliament, 2024; European Parliament, 2026). The precise human-oversight obligations therefore depend on the system's classification under Article 6 and the applicable sectoral legislation (European Parliament, 2024; European Parliament, 2026). Article 14 should consequently be understood as a regulatory requirement for effective oversight of covered high-risk systems, not as a general rule assigning clinical liability to the physician who happens to supervise the system (European Parliament, 2024).

The FDA's January 2026 Clinical Decision Support Software guidance emphasizes that, for software functions seeking to fall within the non-device CDS exclusion, healthcare professionals must be enabled to independently review the basis of recommendations rather than rely primarily on the software output (FDA, 2026). This provides regulatory support for the importance of information that enables clinicians to independently review recommendations rather than relying solely on the software output (FDA, 2026).

WHO Regional Evidence on AI Governance Capacity

A July 2026 WHO statement on AI governance in health reported concerning disparities between deployment and governance in the WHO European Region (World Health Organization, 2026). Only 8% of countries in the region have a health-specific AI strategy. Nearly two-thirds deploy AI in diagnostics, and half have AI-powered patient chatbots. However, only one in five provide AI education before qualification, and only one in four provide workforce training. Fewer than half have assessed legal frameworks, and nearly 40% lack ethical guidance (World Health Organization, 2026). The WHO Regional Director for Europe stated that the longer governance lags behind deployment, the higher the human cost, and that a biased algorithm can produce a wrong diagnosis for a real patient with real consequences (World Health Organization, 2026). These findings provide illustrative regional evidence of governance–deployment gaps rather than a complete global assessment.

Synthesis and Framework

Authorization Is Not Validation

One of the most consequential findings emerging from the evidence synthesized here is the gap between regulatory authorization and publicly identifiable clinical validation. Of 1,357 FDA-cleared AI devices through December 5, 2025, only 2.5% were linked to registered prospective trials and only 0.2% had identified evidence evaluating patient-centered outcomes. The authors of the PLOS Digital Health study characterized the market as crowded with AI tools while the evidence supporting patient-centered outcomes remains limited (Abulibdeh et al., 2026).

This gap is exacerbated by the predominance of the 510(k) pathway, which is based on substantial equivalence rather than clinical outcome improvement. The Epic Sepsis Model provides a cautionary illustration: an externally validated proprietary sepsis prediction model showed substantially lower performance than expected in the evaluated hospital population, with approximately 33% sensitivity and 12% positive predictive value, illustrating the potential for substantial performance degradation when prediction models are externally validated in real-world clinical settings (Wong et al., 2021). The accountability significance is that performance degradation may arise after deployment, creating risks that are neither fully attributable to the original developer nor reasonably preventable by individual clinicians unless organizations establish mechanisms for local validation and continuous monitoring.

Clinical Evidence Is Heterogeneous

Despite the overall evidence gap, specific AI applications have demonstrated meaningful clinical associations. Among patients whose TREWS alerts were confirmed by providers within three hours, provider confirmation was associated with an adjusted absolute mortality reduction of 3.3% (95% CI, 1.7%–5.1%) and adjusted relative reduction of 18.7% (95% CI, 9.4%–27.0%) in a prospective multi-site cohort study (Adams et al., 2022). Because TREWS was evaluated in a prospective observational cohort rather than a randomized trial, the mortality association cannot by itself establish that the AI system caused the observed

reduction; residual confounding and selection effects are possible (Adams et al., 2022). A systematic review and meta-analysis found no statistically significant overall difference in diagnostic performance between generative AI and physicians (physician accuracy was 9.9 percentage points higher than AI, 95% CI -2.3 to 22.0; $p=0.10$) (Takita et al., 2025). The Qazi et al. (2026) trial found that erroneous LLM recommendations significantly degrade physician diagnostic reasoning despite prior AI literacy training (14.0-percentage-point reduction, 95% CI -18.9 to -9.1; $P<0.0001$), highlighting the critical importance of meaningful human oversight. Although the controlled vignette design limits direct inference to routine clinical practice, the study provides experimental evidence that nominal human involvement does not necessarily prevent AI-induced diagnostic error.

Meaningful Human Oversight: A Framework

Building on recent scholarship, meaningful clinical oversight can be operationalized through four interlocking conditions (van de Sande et al., 2026). First, clinicians must possess epistemic capacity sufficient knowledge and understanding of AI capabilities and limitations. Without adequate knowledge, reliance cannot be meaningful. Second, clinicians must have cognitive space adequate time and attention for oversight. Workload pressures undermine oversight quality. Third, clinicians must possess decisional authority the ability to override AI outputs. Without authority, oversight is merely procedural. Fourth, clinicians must have intervention effectiveness the practical ability to change care pathways. Without effectiveness, oversight is performative. Van de Sande et al. (2026) operationalize meaningful oversight through these four conditions; this paper extends that insight by connecting oversight capacity to actor-level responsibility attribution through the Control-Preventability Principle.

The four conditions of meaningful oversight are derived from recent scholarship; the novel contribution of this paper is to connect those conditions to responsibility allocation by asking which actor controls the relevant risk, could reasonably foresee the harm, possessed the capacity to prevent or mitigate it, and possessed meaningful oversight capacity.

These four conditions directly address the accountability problem identified in this analysis. If these conditions are absent, assigning responsibility solely to the clinician becomes normatively difficult to justify and should require additional analysis of organizational, vendor, and developer control.

Table 2: From Meaningful Oversight to Accountability

Meaningful Oversight Condition	Absence Implication	Accountability Implication
Epistemic Capacity	Clinician cannot understand AI limitations	May reduce the normative basis for assigning primary clinician accountability, subject to applicable professional duties
Cognitive Space	Clinician lacks time for adequate review	Potential institutional accountability for workload and staffing
Decisional Authority	Clinician cannot override AI outputs	Potential institutional/developer accountability for design constraints
Intervention Effectiveness	Clinician cannot change care pathway	Potential systemic accountability for workflow and technical design
All Four Present	Meaningful oversight is feasible	Clinician accountability may remain appropriate where reasonable care is breached

The Decision-Location Fallacy

The rapid proliferation of AI tools, combined with limited evidentiary foundations and fragmented regulatory frameworks, creates significant accountability gaps. Physicians face the prospect of being held primarily responsible for harm caused by AI systems over which they have limited control. The BMJ analysis argues that clinician-in-the-loop approaches can function less as robust safety mechanisms and more as ways to shift accountability toward individual clinicians (Toro-Tobon et al., 2026).

Regulatory fragmentation compounds the problem. The EU AI Act imposes obligations on providers of high-risk AI systems, while the FDA's 510(k) pathway does not require clinical outcome improvement. This regulatory divergence creates uncertainty for developers, deployers, and clinicians. The AI Omnibus modifies the application timetable for specified high-risk AI obligations but does not resolve fundamental accountability questions.

This paper terms this recurring attribution error the "Decision-Location Fallacy": the actor making the final clinical decision is not necessarily the actor controlling the source of the relevant AI-related risk. A principled approach to responsibility allocation should track control, foreseeability, preventability, and oversight capacity. Developers generally exercise substantial control over model architecture, training processes, and pre-deployment validation, making them primary accountability candidates for defects or foreseeable risks arising from those stages. Vendors may exercise substantial control over configuration, updates, maintenance, and safety communications, depending on the contractual and technical architecture. Healthcare organizations control procurement, implementation, staffing, training, workflow, monitoring, continued use, and suspension all of which materially affect patient safety. Clinicians control clinical application and override decisions, making them accountable for patient-specific error and unreasonable reliance. Regulators establish authorization criteria and surveillance requirements, making them accountable for systemic evidence gaps and inadequate monitoring frameworks. Where AI functionality is supplied through layered technical and contractual arrangements, responsibility may also extend to upstream model providers, integrators, distributors, or infrastructure providers where they exercise relevant control over the risk in question. The categories are functional rather than necessarily organizational; a single entity may occupy multiple roles.

The Case for Continuing Clinician Responsibility

A counterargument to distributed accountability is that physicians remain licensed professionals with independent duties to patients. Under this view, the clinician's professional judgment cannot be delegated to AI, and the physician should retain primary responsibility regardless of system design or organizational governance.

The Control-Preventability Principle does not eliminate clinician responsibility. It rejects the assumption that clinician responsibility is necessarily primary in every AI-mediated harm. Where clinicians possess epistemic capacity, cognitive space, decisional authority, and intervention effectiveness, and nevertheless fail to exercise reasonable clinical judgment, clinician responsibility remains appropriate. However, where clinicians lack meaningful oversight capacity due to system opacity, workflow constraints, institutional policy, or inadequate training, assigning primary responsibility to the clinician becomes normatively difficult to justify. The principle therefore permits contextual allocation rather than assuming automatic clinician primacy. The framework reallocates neither responsibility nor immunity *ex ante*; it requires incident-specific attribution.

The Control-Preventability Principle

Existing approaches to AI accountability commonly emphasize professional negligence, product liability, organizational governance, or human oversight. The Control-Preventability Principle proposed here differs by treating responsibility allocation as a function of the actor's control over the relevant risk, foreseeability of harm, capacity to prevent or mitigate that harm, and practical capacity for meaningful oversight. It therefore does not assume that the clinician who makes the final clinical decision should necessarily bear primary responsibility. Instead, responsibility is distributed across the actors who create, configure, deploy, monitor, and control the relevant risk across the AI lifecycle.

Control-Preventability Principle of Clinical AI Accountability

Responsibility for AI-related clinical harm should be attributed, subject to applicable legal duties and causation requirements, by examining an actor's control over the relevant risk, the foreseeability of the resulting harm, the actor's feasible capacity to prevent or mitigate that harm, and the actor's practical capacity for meaningful oversight. These dimensions are analytical criteria rather than additive numerical variables. They need not carry equal weight, and different dimensions may point toward different actors within the same incident. The framework therefore permits concurrent responsibility where interacting failures across the AI lifecycle contribute to harm, while avoiding the presumption that the clinician occupying the final decision point should automatically bear primary responsibility.

Control concerns whether an actor possessed decision-making authority over the source of the relevant risk. Foreseeability concerns whether the harm was reasonably predictable given the actor's knowledge and available information. Preventability concerns whether the actor possessed a feasible opportunity to reduce or avoid the harm. Oversight capacity concerns whether the actor possessed the knowledge, time, authority, and practical means necessary to detect and intervene before harm occurred. These dimensions are analytically distinct and may point toward different responsibility allocations in the same incident. Accountability attribution increases as an actor's control, foreseeability, preventability, and meaningful oversight capacity increase, subject to applicable legal duties and causation requirements. The framework is not intended to generate a mechanically calculated liability score. Rather, the four dimensions provide an ordered analytical structure for incident investigation and

responsibility allocation. Because AI-mediated clinical harm may result from interacting failures across multiple lifecycle stages, the framework permits concurrent responsibility rather than requiring attribution to a single actor.

Table 3: Control–Preventability Accountability Matrix

Actor	Principal Control Domain	Illustrative Domains	Risk	Potential Control/Oversight Capacity	Illustrative Accountability Domain	Governance-
Developer	Model architecture, training data, validation	Design defects, bias, unsafe performance		High ex ante	Development/validation	
Vendor	Configuration, updates, monitoring	Performance drift, update failures, warnings		High–moderate	Deployment/lifecycle maintenance	
Healthcare Organization	Procurement, integration, staffing, training	Workflow mismatch, inadequate oversight, training gaps		High	Organizational governance	
Clinician	Patient-specific application, override decisions	Unreasonable reliance, failure to override		Variable	Point-of-care clinical judgment	
Regulator	Authorization criteria, surveillance requirements, enforcement	Inadequate evidence requirements, surveillance		System-level	Regulatory responsibility	governance

Note. The matrix illustrates governance-accountability domains under the proposed normative framework and does not constitute a determination of legal liability. Legal liability depends on applicable duty, breach, causation, and damage doctrines. The regulator row concerns institutional governance responsibility and regulatory-system accountability, not a presumption of governmental civil liability.

The central governance problem in clinical AI is not whether humans remain "in the loop," but whether the human and institutional actors assigned or expected to bear responsibility possess sufficient control, knowledge, authority, and intervention capacity to prevent or mitigate AI-related harm.

Relationship to Existing Liability Doctrines

The Control–Preventability Principle does not replace existing legal doctrines but can inform how those doctrines might be applied in AI-mediated clinical contexts.

Developers may, depending on jurisdiction and applicable law and on the actor's legally recognized role in the relevant supply chain, be potentially implicated by product defect, negligent design, inadequate warnings, or failure to monitor post-market performance. Vendors may be potentially implicated by failures in maintenance, updates, configuration, or safety communications. Healthcare organizations may be potentially implicated by negligent procurement, credentialing, training, workflow design, or monitoring. Clinicians may be potentially implicated by unreasonable reliance, failure to review, or failure to override where reasonable.

The framework does not determine whether any of these doctrines are satisfied. Rather, it identifies which actors are most plausibly positioned to prevent particular categories of AI-related risk and therefore should be the focus of governance, investigation, and accountability attribution. It also distinguishes governance accountability from legal causation, the latter being a jurisdiction-specific requirement for liability. The precise doctrinal consequences of these attribution factors will vary across jurisdictions because duties, standards of care, product-liability rules, causation doctrines, and governmental liability protections differ materially (Mello & Guha, 2024).

Regulatory Implications

The EU AI Act's Article 14 requires effective human oversight for high-risk AI systems, including understanding capabilities and limitations, detecting anomalies, and overriding outputs (European Parliament, 2024). For AI systems incorporated into medical devices and IVDs, which remain in Section A of Annex I, the high-risk framework applies integrated with the MDR/IVDR conformity assessment. Under Article 113 as amended by the July 2026 Omnibus, the application of Chapter III, Sections 1–3 for

such systems is deferred to August 2, 2028 (European Parliament, 2024; European Parliament, 2026). The precise human-oversight obligations therefore depend on the system's classification under Article 6 and the applicable sectoral legislation (European Parliament, 2024; European Parliament, 2026). The Act does not, by itself, establish a comprehensive civil-liability allocation rule (European Parliament, 2024). Article 14 establishes requirements for human oversight, while van de Sande et al. (2026) provide a more operational framework for assessing whether such oversight is meaningful.

The FDA's January 2026 CDS guidance emphasizes the importance of information that enables healthcare professionals to independently review the basis of recommendations, reinforcing the importance of information access and clinician capacity rather than nominal human presence alone (FDA, 2026).

Governance Recommendations: Five Pillars

Pillar 1: Risk-Proportionate Evidence Standards

The evidentiary foundation for AI in healthcare must be strengthened through prospective or otherwise clinically meaningful validation proportionate to the clinical risk, degree of autonomy, reversibility of decisions, affected population, and potential severity of harm. For AI functions that pose substantial clinical risk or materially influence consequential decisions, regulators should consider supplemental evidence requirements beyond predicate-based equivalence, including prospective or clinically meaningful validation proportionate to risk. For AI prediction models and applicable clinical prediction studies, developers and researchers should use appropriate reporting and appraisal frameworks such as TRIPOD+AI and PROBAST+AI. Post-market surveillance and real-world performance monitoring should be established with explicit triggers for investigation, including subgroup performance analysis to detect algorithmic bias.

Pillar 2: Contextual and Equity-Sensitive Validation

External and local validation should be required before deployment in new clinical contexts, populations, or settings. Subgroup performance monitoring should be mandated to detect algorithmic bias, especially concerning age, sex, race/ethnicity where legally and ethically appropriate, socioeconomic position, language, geography, disability, and relevant comorbidities. Training-data provenance, external validation, temporal validation, geographic validation, site-specific calibration, and subgroup performance should all be assessed. This ensures that AI tools perform safely and effectively across the diverse populations they will serve.

Pillar 3: Operationally Meaningful Human Oversight

Clinicians must possess four conditions for meaningful oversight: epistemic capacity (sufficient understanding of AI capabilities and limitations), cognitive space (adequate time and attention for oversight), decisional authority (ability to accept, modify, or reject AI recommendations), and intervention effectiveness (practical ability to change AI-mediated care). A clear distinction must be drawn between formal human presence and meaningful oversight capacity. Healthcare organizations should audit whether these conditions are met before deploying AI tools in clinical workflows.

Pillar 4: Distributed and Traceable Accountability

Responsibility should be allocated according to the Control-Preventability Principle: control, foreseeability, preventability, and oversight capacity not simply according to who makes the final clinical decision. AI-related clinical incidents should be investigated through structured attribution processes that reconstruct the model version, data inputs, recommendation, human interpretation, workflow conditions, override opportunities, organizational controls, and subsequent actions. Incident investigation should distinguish causal contribution from accountability attribution: an actor may contribute to an outcome without possessing the relevant control or preventability necessary to justify primary accountability. Indemnity, contractual, and insurance arrangements should fairly apportion liability risk across developers, vendors, deploying organizations, and clinicians. Traceable chains of responsibility should be established, documenting who is responsible for intended use, local validation, workflow integration, staff training, patient communication, monitoring, incident review, and withdrawal.

Pillar 5: Lifecycle Monitoring and Adaptive Governance

Lifecycle governance should be adopted, drawing on FDA's Total Product Life Cycle approach and Predetermined Change Control Plan (PCCP) framework where applicable to the device's regulatory pathway and change profile, spanning development, deployment, monitoring, and retirement. Lifecycle monitoring should assess both technical/model drift and socio-technical

workflow drift. Incident reporting, adverse-event monitoring, and performance drift detection should be integrated into routine clinical governance. This ensures that AI tools are continuously evaluated and updated to maintain safety and effectiveness throughout their operational life.

Limitations

This analysis has several limitations. First, the selected clinical studies were purposively chosen for analytical relevance and methodological strength and therefore should not be interpreted as a representative sample of the broader clinical-AI evidence base. Second, the FDA authorization data represent authorizations rather than evidence of clinical effectiveness. Third, the clinical trial evidence is concentrated in specific domains and may not generalize to all AI applications. Fourth, the regulatory analysis focuses on the EU and United States; other jurisdictions may have different frameworks. Fifth, the evidence gap analysis is based on a single systematic review with a December 5, 2025 cutoff and may not capture all published studies. Sixth, this analysis is not a formal systematic review of all clinical AI literature. Seventh, the Control–Preventability Principle is normative and has not been tested against real malpractice or clinical incident cases; future research should apply the framework retrospectively to reported AI-related clinical incidents and compare its attribution patterns with outcomes under existing negligence, product-liability, and organizational-liability doctrines. Eighth, actor-control assumptions may be difficult to observe because contractual and technical control may be distributed across organizations. Ninth, the evidence streams are heterogeneous: RCTs, observational cohorts, device registries and systematic reviews measure different constructs and cannot be directly pooled. Tenth, regulatory authorization does not measure deployment volume: a device's authorization does not establish that it is widely used clinically.

Conclusion

Artificial intelligence is transforming healthcare at an unprecedented pace. By the end of 2025, the FDA AI/ML-enabled medical-device list contained 1,430 authorized devices, including 331 authorizations during 2025 the highest annual total to date. Yet the available evidence suggests that publicly identifiable rigorous prospective clinical validation is uncommon among FDA-authorized AI devices. Of 1,357 FDA-cleared AI devices through December 5, 2025, only 2.5% were linked to registered prospective trials and only 0.2% had identified evidence evaluating patient-centered outcomes.

Promising results exist in specific domains. Among patients whose TREWS sepsis alerts were confirmed by providers within three hours, provider confirmation was associated with an adjusted absolute mortality reduction of 3.3% (95% CI, 1.7%–5.1%) in a prospective multi-site cohort study. A systematic review and meta-analysis found no statistically significant overall difference in diagnostic performance between generative AI and physicians, while subgroup analyses of randomized controlled trials suggested that LLM-assisted physicians may show improved performance relative to physicians alone. However, the Qazi et al. (2026) trial found that erroneous LLM recommendations significantly degrade physician diagnostic reasoning despite prior AI literacy training (14.0-percentage-point reduction, 95% CI –18.9 to –9.1; $P < 0.0001$).

The accountability gap requires attention. Only 8% of countries in the WHO European Region have a health-specific AI strategy. Only one in five provide AI education before qualification. Clinicians face the prospect of being held primarily responsible for harm caused by AI systems over which they have limited control. The EU AI Act imposes obligations on providers and deployers of high-risk AI systems. For AI incorporated into medical devices and IVDs, which remain in Section A of Annex I, the high-risk framework applies integrated with the MDR/IVDR conformity assessment, with the application of Chapter III, Sections 1–3 deferred to August 2, 2028 under Article 113 (European Parliament, 2024; European Parliament, 2026). This deferral does not resolve fundamental accountability questions.

The central governance problem in clinical AI is not whether humans remain "in the loop," but whether the human and institutional actors assigned or expected to bear responsibility possess sufficient control, knowledge, authority, and intervention capacity to prevent or mitigate AI-related harm. Regulatory authorization, clinical validation, and legal accountability operate on different timelines. AI governance should therefore allocate responsibility according to the capacity of each actor to prevent, detect, and correct AI-related harm not simply according to who happens to make the final clinical decision. The central contribution of this framework is therefore not to remove clinicians from accountability, but to prevent accountability from being assigned solely because the clinician occupied the final decision point.

The path forward requires action on five governance pillars: strengthening the evidence base, ensuring local validation, operationalizing meaningful human oversight, clarifying distributed accountability, and embracing lifecycle governance.

References

1. Abulibdeh, R., Cajas Ordóñez, S. A., Celi, L. A., Gorijavolu, R., Izath, N., & Markussen Lunde, T. (2026). 1,357 AI medical devices cleared, 3 actually tested on patient outcomes. *PLOS Digital Health*, 5(8), e0001597. <https://doi.org/10.1371/journal.pdig.0001597>
2. Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., & Pronovost, P. J. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*. <https://doi.org/10.1038/s41591-022-01894-0>
3. European Parliament. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)* (Article 14). Official Journal of the European Union.
4. European Parliament. (2026). *Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)*. Official Journal of the European Union, L 2026/1744.
5. FDA. (2026). *Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff* (Final Guidance). U.S. Food and Drug Administration.
6. FDA. (n.d.). *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. U.S. Food and Drug Administration. Retrieved December 31, 2025.
7. Henry, K. E., Adams, R., Parent, C., Soleimani, H., Sridharan, A., Zhan, A., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., & Pronovost, P. J. (2022). Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nature Medicine*, 28(7), 1447-1454. <https://doi.org/10.1038/s41591-022-01895-z>
8. Lee, Y. S., & Youn, B. Y. (2026). Clinical specialty expansion of AI-enabled and machine learning-enabled medical devices authorized by the US Food and Drug Administration from 1995 to 2025: Longitudinal content analysis. *Journal of Medical Internet Research*, 28, e103040. <https://doi.org/10.2196/103040>
9. Mello, M. M., & Guha, N. (2024). Understanding liability risk from using health care artificial intelligence tools. *New England Journal of Medicine*, 390, 271-278. <https://doi.org/10.1056/NEJMHle2308901>
10. Qazi, I. A., Ali, A., Khawaja, A. U., Akhtar, M. J., Sheikh, A. Z., & Alizai, M. H. (2026). Automation bias in large language model-assisted diagnostic reasoning among physicians trained in AI literacy: A randomized clinical trial. *NEJM AI*, 3(5), A10a2501001. <https://doi.org/10.1056/A10a2501001>
11. Takita, H., Kabata, D., Walston, S. L., Tatekawa, H., Saito, K., Tsujimoto, Y., Aoki, R., Horiuchi, D., Kakiuchi, Y., Kato, S., Saito, D., Takeda, H., Uchida, T., Yagisawa, M., Yanagisawa, T., Yoshioka, K., Miki, Y., & Ueda, D. (2025). A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *npj Digital Medicine*, 8, 175. <https://doi.org/10.1038/s41746-025-01543-z>
12. Toro-Tobon, D., Ponce Ponte, O., Montori, V. M., & Brito, J. P. (2026). Clinician in the loop: A flawed solution for AI oversight. *BMJ*, 393, e089213. <https://doi.org/10.1136/bmj-2025-089213>
13. van de Sande, D., Economou-Zavlanos, N., & van Genderen, M. E. (2026). Meaningful oversight of medical AI beyond human in the loop. *npj Digital Medicine*, 9, 569. <https://doi.org/10.1038/s41746-026-02971-1>
14. Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penzoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065-1070. <https://doi.org/10.1001/jamainternmed.2021.2626>

15. World Health Organization. (2026). *Statement – Govern AI in health before the gaps become irreversible*. World Health Organization.
16. Zwaan, L., Rodman, A., & Shimizu, T. (2026). Trust, scrutiny, or collaboration? A performance-based framework for human–AI interaction in medicine. *NEJM AI*, 3(5), A1e2600354. <https://doi.org/10.1056/A1e2600354>



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).