

# Operating Models for Regulated Clinical AI at Scale: From Pilot Systems to Auditable Production-Grade Deployment

Roni Amiel

*Biomedical Informatics, Rutgers University, College Avenue Campus Ringgold Standard Institution, New Jersey, United States*

DOI:10.37819/Jaike.26.0002

## Article history:

Received : 09.01.2026

Revised : 13.07.2026

Accept : 15.07.2026

Published : 21.10.2026

## Corresponding author:

**E-mail:** roniamiel1@gmail.com,  
roniamiel@outlook.com

**Abstract:** Clinical artificial intelligence (AI) has demonstrated strong performance in controlled research and pilot settings, yet most clinical AI systems fail to scale into durable, production-grade deployments. In regulated healthcare environments, technical performance alone is insufficient; success depends on operating models that integrate regulatory constraints, clinical workflows, organizational accountability, and lifecycle governance into system design. While existing regulatory guidance (e.g., FDA and IMDRF Good Machine Learning Practice), quality standards (e.g., ISO 13485/14971), and MLOps paradigms provide necessary principles, they often stop short of defining how those principles are executed as an organizational operating system. This paper proposes an end-to-end operating model for regulated clinical AI, grounded in real-world deployment experience across diagnostic and enterprise healthcare environments. The model extends beyond model-centric or compliance-only approaches by explicitly linking governance structures, technical lifecycle management, and clinical workflow integration to decision rights and accountability. Regulation, auditability, and risk management are treated as foundational design inputs rather than downstream constraints. The framework comprises four interdependent layers: governance and decision rights; technical lifecycle management from development through post-market monitoring; clinical integration with adaptive human-in-the-loop (HITL) controls that evolve from pilot deployment to scaled operations; and continuous performance, safety, and economic feedback loops. Together, these layers define the organizational operating system required to sustain clinical AI safely, auditably, and economically at scale.

**Keywords:** Clinical artificial intelligence, regulated deployment, operating model, AI governance, human-in-the-loop oversight, post-market monitoring, lifecycle management

## 1. INTRODUCTION

Clinical artificial intelligence (AI) has advanced rapidly, driven by improved data availability, computational capacity, and algorithmic innovation. Deep learning systems now match or exceed human benchmarks in selected tasks across imaging, diagnostics, and decision support (Beam & Kohane, 2018; Obermeyer & Emanuel, 2016; Topol, 2019). Yet translation from controlled research into durable

production remains difficult. Prior literature has shown that barriers to clinical impact extend beyond model performance and include implementation logistics, workflow disruption, sociotechnical adoption barriers, local validation requirements, and the need for ongoing post-deployment surveillance (Kelly et al., 2019; Rajkomar et al., 2019; Sendak et al., 2020).

At scale, the dominant failure modes are often operational rather than purely algorithmic.

Models that perform well in retrospective or pilot settings may degrade when exposed to workflow heterogeneity, dataset shift, evolving clinical practice, variable documentation quality, and fragmented governance (Finlayson et al., 2019). Prior work on hidden technical debt, dataset shift, and real-world deployment has shown that successful implementation depends on disciplined lifecycle management, clear accountability, and sustained monitoring after go-live (Raji et al., 2020; Sculley et al., 2015; Sendak et al., 2020; Subbaswamy & Saria, 2020).

Regulators and standards bodies have provided foundational principles for trustworthy AI and lifecycle oversight (National Institute of Standards and Technology, n.d.; Tabassi, 2023; World Health Organization, 2021). FDA and IMDRF Good Machine Learning Practice emphasize data quality, human factors, transparency, and post-market monitoring across the total product lifecycle, while ISO 13485 and ISO 14971 define expectations for quality management and risk control (International Medical Device Regulators Forum, 2025; International Organization for Standardization, 2016, 2019; U.S. Food and Drug Administration, 2021, 2025b). FDA's recent guidance on transparency and predetermined change control plans further reinforces that regulated AI requires explicit mechanisms for explainability, change governance, and lifecycle evidence management (U.S. Food and Drug Administration, 2024, 2025c).

However, these frameworks remain only partially operationalized in practice. The literature provides substantial guidance on model development and validation, but comparatively less on implementation, governance, software deployment, impact assessment, and organizational execution in routine care (Bedoya et al., 2022; Feng et al., 2022; Kelly et al., 2019; Roski et al., 2021). This paper therefore argues that one of the primary barriers to scaling regulated clinical AI is the absence of an operating model that binds governance, lifecycle engineering, workflow integration, and economic accountability into one executable system. The contribution is not a new regulatory standard or model architecture; it is an applied operating model, derived from repeated deployment patterns across regulated environments, that specifies how those domains are operationally connected through decision rights, lifecycle controls, escalation pathways, and monitoring-linked action.

## 2. RELATED WORK AND EXISTING FRAMEWORKS

Prior work spans regulatory guidance, quality and risk standards, MLOps engineering practices, and sociotechnical adoption research. FDA and IMDRF guidance emphasize lifecycle thinking, data and model management, transparency, human factors, and post-market monitoring (International Medical Device Regulators Forum, n.d.; U.S. Food and Drug Administration, n.d.-a, n.d.-b, 2025a). These principles are intentionally nonprescriptive about organizational design, leaving execution to manufacturers and healthcare systems. ISO 13485 and ISO 14971 provide disciplined processes for documentation, verification and validation, and corrective actions; in practice, however, organizations may implement these standards as documentation layers without embedding them into operational decision making. MLOps improves reproducibility and operational deployment of machine learning through versioning, CI/CD, monitoring, and automated testing, but baseline MLOps patterns typically do not encode regulatorily defensible change control, clinical accountability, or the human factors required in clinical settings. Studies of clinical decision support and health IT adoption identify barriers such as workflow disruption, trust calibration, unintended consequences, and alert fatigue — critical observations that are rarely translated into concrete organizational structures such as decision rights, governance cadence, and escalation thresholds that sustain safe operation.

This paper complements these domains by specifying the operating system that makes the above principles executable at scale: who decides what evidence is required, how change is controlled, how humans remain in the loop as scale increases, and how performance signals drive action. While much of the regulatory discussion references U.S. guidance, the operating model is intentionally jurisdiction-agnostic. In the European context, the framework aligns with EU MDR 2017/745, particularly its requirements for post-market surveillance, clinical evaluation continuity, and corrective and preventive actions (Articles 83–86), and with the EU AI Act's treatment of high-risk AI systems, including requirements for human oversight, traceability, and risk management (European Union, 2017).

2.1. Evidence Base and Methodological Approach

The operating model presented here is derived using a structured qualitative cross-case synthesis informed by thematic analysis methodology (Braun & Clarke, 2006; Eisenhardt, 1989) and implementation-science approaches to sociotechnical systems. The study adopts a qualitative analytical approach focused on identifying recurring operational patterns rather than estimating causal effects.

**Case definition and inclusion criteria:** A deployment was defined as a clinical AI system that was actively integrated into clinical workflows (not retrospective only), operated within a regulated or governed healthcare environment, and had progressed beyond initial pilot into operational use with monitoring or oversight mechanisms in place. Deployments remaining in pilot or proof-of-concept stage were excluded.

**Sample selection:** A total of 12 deployments spanning 2016 to 2024 were included based on three criteria: exposure to multiple lifecycle stages (deployment and monitoring), representation across diagnostic, clinical decision support, and predictive AI systems, and variation in regulatory context (FDA SaMD, EU MDR, and institutional governance). Case identification drew on the author’s direct operational access across participant-observer and advisory roles; no systematic search of external databases was conducted, and the sample is accordingly purposive rather than representative.

**Data sources:** Sources included implementation records, lifecycle and monitoring artifacts, audit and regulatory-readiness documentation, and incident and escalation logs. No patient-level data were accessed or analyzed.

**Analytical method:** A six-phase thematic synthesis was conducted following Braun and Clarke (2006): familiarization with deployment artifacts; initial coding of operational observations; identification of recurring themes; cross-case validation of themes; mapping of dependencies across governance, lifecycle, integration, and feedback dimensions; and synthesis into the four-layer operating model. Patterns were retained only when observed in three or more deployments and associated with consistent operational consequences. The recurrence threshold was set prior to analysis and applied without exception. The four-layer model emerged inductively from the pattern set rather than being imposed a priori. All analytical decisions were documented contemporaneously in operational notes and cross-referenced against the published implementation literature to assess convergent validity.

**Bias and limitations:** The author acted as a participant-observer across multiple deployments, introducing potential selection bias — the sample may skew toward environments where operating-model discipline was already partially present — and interpretation bias, as the author cannot fully separate analytical from operational judgments. These were mitigated through the cross-case recurrence threshold, alignment with existing literature, and explicit non-causal framing. These mitigations reduce but do not eliminate the limitation; readers should treat the framework as hypothesis-generating and the reported associations as motivating evidence for the independent empirical validation described in Section 7.

**Ethics statement:** This work does not constitute human-subjects research. No identifiable patient data were used. The analysis is limited to system-level and organizational processes.

Table 1. Deployment contexts informing the operating model

| Deployment Type                   | Regulatory Context                      | Typical Operational Scale | Recurring Operational Challenge                 |
|-----------------------------------|-----------------------------------------|---------------------------|-------------------------------------------------|
| Diagnostic AI platforms           | FDA SaMD / CLIA-regulated environments  | 20–50 clinical sites      | Governance fragmentation during scale expansion |
| Clinical decision support systems | Hospital clinical governance frameworks | 3–10 hospitals            | Workflow integration variability                |

Table 1. Continued

| Deployment Type                         | Regulatory Context                | Typical Operational Scale        | Recurring Operational Challenge                    |
|-----------------------------------------|-----------------------------------|----------------------------------|----------------------------------------------------|
| Predictive analytics within EHR systems | HIPAA and institutional oversight | Single enterprise health systems | Inconsistent monitoring and escalation             |
| AI-enabled monitoring platforms         | EU MDR / ISO 13485 environments   | Multi-country deployments        | Lifecycle validation and change-control complexity |

These recurring deployment patterns informed both the abstraction of the four-layer operating model and the generalized case anchors presented in Section 5 (Table 1).

## 2.2. Execution Layer and Novelty

This work is best understood as an execution-layer operating model that formalizes how regulatory, technical, and clinical domains are operationally connected through decision rights, lifecycle controls, and escalation mechanisms. The contribution extends beyond synthesis by defining a structured model of operational dependencies that can be empirically tested and implemented across regulated environments.

## 3. CONSTRAINTS OF REGULATED CLINICAL AI SYSTEMS

Regulated clinical AI operates under intertwined regulatory, clinical, organizational, and economic constraints that intensify with scale. Regulatory constraints include validation, traceability, transparency, controlled change management, and post-market surveillance, extending beyond initial clearance into continuous monitoring and auditable evidence of real-world behavior (Centers for Medicare & Medicaid Services, n.d.; U.S. Department of Health and Human Services, Office for Civil Rights, n.d.). Clinical constraints include workflow heterogeneity, time pressure, liability, uncertainty communication, and the need for clinically interpretable outputs; systems adding cognitive burden or low-value interruptions are often ignored or worked around regardless of nominal technical performance (Ancker et al., 2017; Van Dort et al., 2021). Organizational constraints include siloed teams, competing priorities, local variation across sites, and ambiguity in ownership — conditions under which deployment quality depends on explicit structures for approval, monitoring,

and intervention across technical, clinical, and operational stakeholders (Bedoya et al., 2022; Roski et al., 2021). Economic constraints arise because remediation, revalidation, downtime, and manual review can erode expected value, making structured operating models essential not only for safety and compliance but also for reducing avoidable operational friction.

## 4. OPERATING MODEL FOR REGULATED CLINICAL AI AT SCALE

This section sets out the paper's conceptual contribution. The four-layer operating model presented here is proposed as a structured framework derived from deployment experience; it is offered as a conceptual model rather than as a statistically validated empirical result (Jaakkola, 2020). The model defines how organizations operationalize regulatory principles, quality standards, and MLOps practices as a unified execution system, specifying how governance, engineering, workflow design, and monitoring functions are linked through decision authority, lifecycle controls, operational cadence, and intervention pathways.

### 4.1. Design Principles

Five principles underpin the operating model. First, regulation is treated as a design input: traceability, evidence lineage, and post-market monitoring are built into architecture and workflows from the outset, not appended at compliance review. Second, decision rights precede technology: accountability for clinical outcomes, model behavior, and regulatory posture is made explicit before deployment. Third, lifecycle continuity: development through post-market monitoring is managed as a single lifecycle with controlled transitions rather than discrete handoffs. Fourth, economic sustainability: scaling requires reducing manual oversight load and preventing rework cycles. Fifth, global regulatory

alignment is treated as an execution concern rather than a structural constraint — by defining governance authority, lifecycle controls, and post-market feedback loops independently of any single regulator's documentation format, the model enables organizations to satisfy divergent regulatory regimes through localized artifacts while preserving a unified operating system, allowing scale across geographies without re-engineering operating capability for each jurisdiction.

#### 4.2. Comparative Positioning

Good Machine Learning Practice defines lifecycle principles; this operating model defines the execution mechanisms — governance forums, approvals, escalation thresholds, and operating cadence — through which those principles are enacted. ISO 13485 and ISO 14971 define quality and risk processes; this model embeds them into day-to-day decision making and clinical operations rather than

treating them as documentation overlays. MLOps defines technical deployment discipline; this model adds regulatorily defensible change control, clinical accountability, and HITL evolution that clinical settings require.

#### 4.3. Four-Layer Architecture

The operating model comprises four interdependent layers. Layer 1 establishes governance and decision rights, defining who approves deployment and authorizes change. Layer 2 covers technical lifecycle management from development through post-market monitoring. Layer 3 addresses clinical integration with adaptive human-in-the-loop controls that evolve as scale increases. Layer 4 provides continuous performance, risk, and economic feedback loops that link monitoring signals to governance authority and controlled intervention. Figure 1 illustrates how these layers interact as an integrated system.

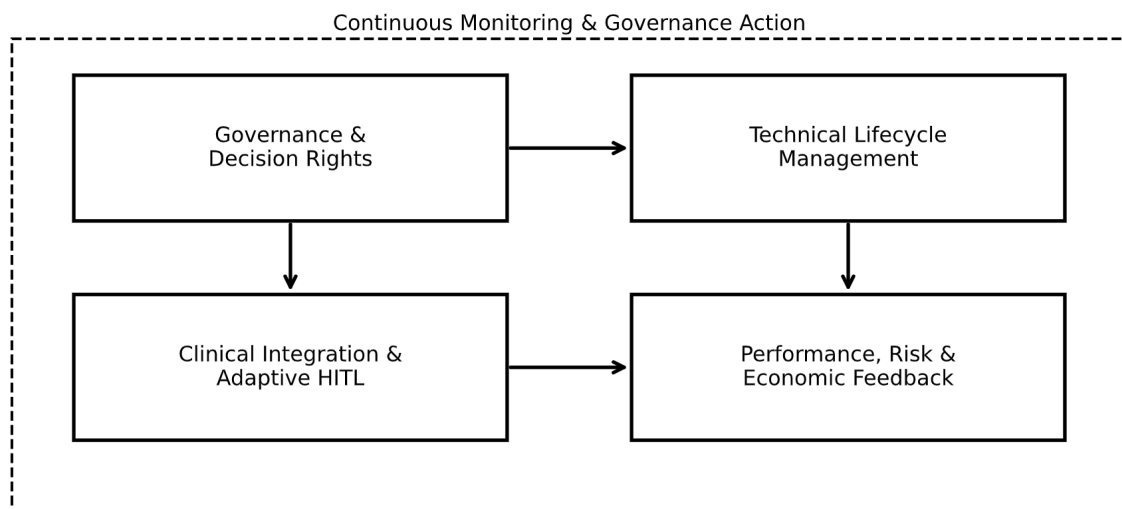


**Figure 1.** Operating model architecture for regulated clinical AI, integrating governance, lifecycle management, clinical workflow integration, and continuous performance and risk feedback loops.

Source: Author-generated figure derived from deployment patterns and regulatory requirements; created specifically for this manuscript

#### 4.4. Governance and Decision Rights

Governance is operational authority, not policy. It defines who owns clinical outcomes, who approves releases, who can pause or roll back a system, how incidents and drift are escalated, and how audit findings trigger corrective and preventive action (CAPA). Key governance artifacts include a

governance charter, role definitions, a RACI map for lifecycle stages, release and change-approval thresholds, and documented escalation timelines. Effective governance separates model-development accountability from clinical-approval accountability, reducing conflicts of interest and improving defensibility. Governance cadence matters: standing

forums (weekly operational review, monthly risk and quality review, and quarterly product review) create structured venues for escalation and cross-functional accountability, reducing reliance on ad hoc decision making. Governance also establishes predefined thresholds that trigger review, rollback, or constrained use, limiting ambiguity during incidents or audit pressure.

#### 4.5. Technical Lifecycle Management

This layer extends MLOps to meet regulated requirements through controlled development and validation environments, versioned datasets and artifacts, reproducible training, structured verification and validation, and change control aligned to regulatory strategy. Four core components form its foundation. Evidence lineage captures dataset provenance, labeling protocols, observed data shifts, and representativeness analyses. Model lineage records architecture, hyperparameters, training runs, and calibration decisions. Release packages assemble the model card, intended-use statement, contraindications, validation summaries, human-factors notes, monitoring plan, and rollback plan. Post-market change control defines triggers for retraining, revalidation tiers, and the documentation updates required at each tier.

#### 4.6. Clinical Integration and HITL Evolution

Human-in-the-loop oversight is a lifecycle variable rather than a fixed setting. During pilots, oversight is intentionally high-touch to validate appropriateness and build trust. At early scale, oversight shifts to structured sampling and exception-driven review with defined escalation criteria. At mature scale, HITL becomes targeted: monitoring-driven interventions

with periodic audits and review. Embedding HITL in the primary workflow is critical, as parallel review systems create alert fatigue and workarounds. Integration must also address human factors: clarity of recommendations, uncertainty communication, and user feedback loops.

**Extension to generative and non-deterministic systems:** Generative and large language model (LLM)-based systems introduce additional risk dimensions not typically present in conventional discriminative models, including non-deterministic outputs, probabilistic reasoning chains, and the potential for hallucinated or otherwise unverifiable responses. Within the proposed operating model, these risks are mitigated primarily through governance and lifecycle controls rather than model-level constraints: intended-use definitions are narrowed, human-in-the-loop oversight is elevated from sampling-based review to mandatory gating at defined decision points, and output monitoring focuses on semantic validity, override patterns, and downstream clinical actions. Change control thresholds for generative systems are correspondingly stricter. This reinforces the central thesis: as model behavior becomes less deterministic, organizational operating discipline becomes the primary safety mechanism.

#### 4.7. Continuous Feedback Loops

Monitoring must be linked to authority. Performance, safety, fairness, and economic signals are mapped to thresholds and response playbooks. Economic signals — cost of oversight, rework burden, and downtime — must be visible to governance to prevent unsustainable scaling. Feedback loop maturity is the difference between late discovery of drift and structured, preemptive intervention.

**Table 2.** Failure modes by missing operating layer

| Missing Layer        | Common Failure Mode    | Typical Consequence                | Downstream Impact                   |
|----------------------|------------------------|------------------------------------|-------------------------------------|
| Governance           | Unclear accountability | Indecision during audits/incidents | Audit exposure; trust loss          |
| Lifecycle Management | Uncontrolled changes   | Drift; revalidation                | Cost escalation; downtime           |
| Clinical Integration | Workflow misalignment  | Low adoption; workarounds          | Value erosion; safety risk          |
| Feedback Loops       | Delayed intervention   | Late remediation                   | Reputational damage; avoidable harm |

#### 4.8. Failure Mode Matrix

Failures map predictably to missing operating layers. Without governance, organizations experience indecision and audit exposure. Without lifecycle control, they accumulate uncontrolled changes and unplanned revalidation. Without integration, they see low adoption and workarounds. Without feedback loops, intervention is delayed and risk compounds. Table 2 summarizes these relationships. Sequencing also matters: governance first, lifecycle management second, integration and HITL design third, and feedback-loop maturation last. Deployments that invert this sequence are consistently associated with higher remediation burden, slower scale transitions, and greater operational risk (Raji et al., 2020; Sculley et al., 2015).

### 5. GENERALIZED CASE ANCHORS: OPERATING MODELS IN PRACTICE

The material in this section constitutes illustrative, case-based evidence. The cases document observed patterns from real-world deployment programs to show how the framework operates in practice; they are not presented as controlled trials or statistically

generalizable findings. The case anchors are qualitative abstractions derived from repeated deployment patterns observed across multiple regulated clinical AI implementations, normalized to preserve confidentiality while enabling reproducibility of insight. Source material includes diagnostic AI systems, enterprise clinical decision support platforms, predictive analytics embedded within electronic health record infrastructures, and post-market monitoring programs supporting regulated AI products. Each case isolates structural dependencies within the operating model rather than describing a single system implementation.

#### 5.1. Case A: Pilot to First Regulated Deployment

Clinical AI pilots often demonstrate strong technical performance and local clinical acceptance under conditions of informal oversight and close collaboration between clinicians and technical teams. In this case, an AI system was initially deployed within a narrowly scoped pilot addressing a single clinical use case. Validation artifacts were created iteratively, and decisions regarding model updates and usage were made through consensus rather

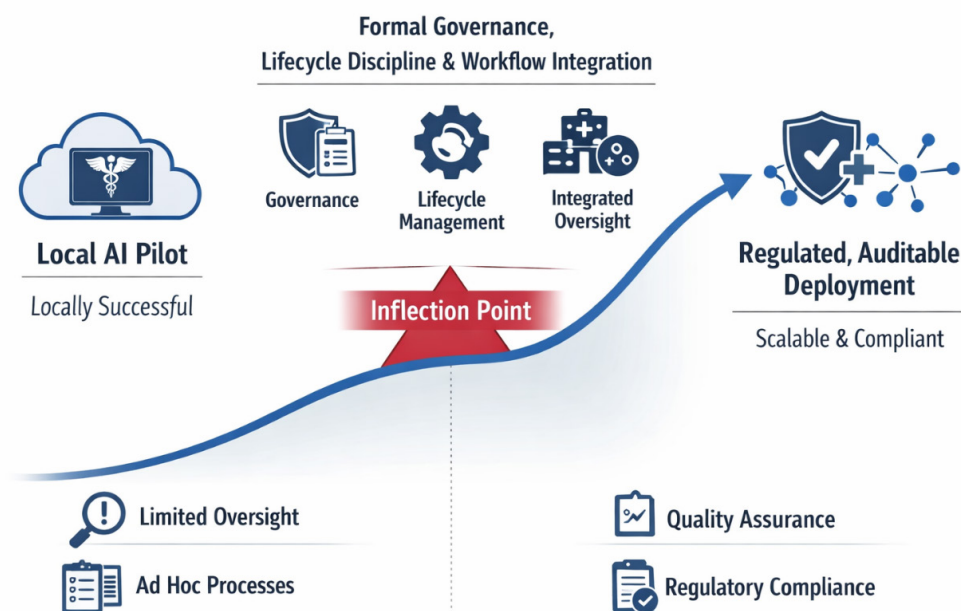


**Figure 2.** Transition from pilot deployment to regulated clinical AI production enabled by operating model maturity rather than algorithmic change. Source: Author-generated figure derived from deployment patterns and regulatory transition requirements; created specifically for this manuscript

than formal approval structures. These conditions supported rapid iteration but masked structural weaknesses.

As the organization attempted to transition the system into a regulated production environment, operating gaps became apparent. Documentation practices varied across development cycles, evidence lineage was incomplete, and accountability for clinical outcomes versus technical performance was not clearly delineated. Release readiness was assessed primarily through engineering criteria, while regulatory and clinical sign-off processes remained loosely defined. These gaps surfaced during regulatory review, resulting in delayed approval, additional evidence requests, and rework.

Stabilization required changes to the operating model rather than to the underlying algorithm. The organization introduced formal governance structures defining decision rights and escalation pathways, standardized lifecycle documentation, and embedded human-in-the-loop oversight directly into clinical workflows. Approval thresholds and audit-ready artifacts were established before resubmission. Following these changes, the system achieved

regulated deployment with greater predictability, reduced remediation cycles, and increased organizational confidence. This case suggests that failure in the transition from pilot to regulated deployment is frequently driven by operating model deficiencies rather than core algorithmic shortcomings (Figure 2).

## 5.2. Case B: Multi-Site Scaling

Following initial regulated deployment, the AI system was expanded across multiple clinical sites with heterogeneous workflows, staffing models, and local governance norms. Although the core model and technical infrastructure remained consistent, operational execution varied significantly: local teams adapted workflows independently, oversight practices diverged, and escalation thresholds were interpreted inconsistently across sites. As scale increased, these variations compounded. Aggregate model-performance metrics remained stable, obscuring underlying operational fragility, and expansion slowed as leadership struggled to reconcile local autonomy with enterprise-level risk and compliance expectations.

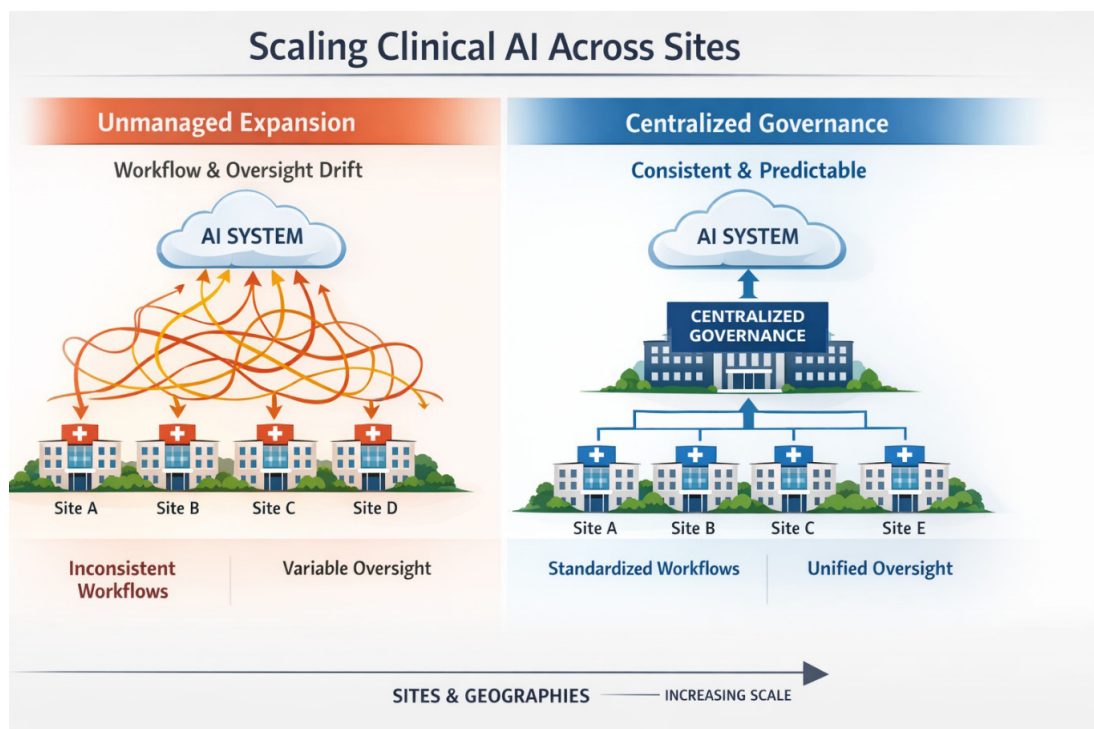


**Figure 3.** Multi-site scaling dynamics illustrating how centralized operating models mitigate workflow and oversight drift as clinical AI scales across sites and geographies. Source: Author-generated figure informed by common multi-site clinical deployment patterns; created specifically for this manuscript

Resolution required re-centering the scale-up effort around the operating model rather than the model alone. Central governance defined minimum operating standards, clarified decision rights, and established uniform escalation pathways, while permitting controlled local variation where clinically appropriate. Human-in-the-loop oversight evolved from universal review to structured sampling aligned with site maturity and risk profile, and economic metrics were introduced to surface hidden costs associated with fragmented execution. The key takeaway is that scaling amplifies organizational variance; without a unifying operating model, local adaptations accumulate into systemic inefficiency and risk (Figure 3).

### 5.3. Case C: Post-Market Drift and Audit Stress Test

After sustained production use, the clinical AI system entered a period of apparent stability. Over time, subtle performance changes emerged due to shifts in data characteristics, evolving clinical practices, and upstream system modifications (Koch et al., 2024). Concurrently, the organization underwent

an external regulatory audit, placing heightened scrutiny on post-market monitoring and change-management practices.

Monitoring infrastructure detected performance drift and variation across subpopulations, but organizational response differed based on operating model maturity. In environments lacking clear decision authority, technical findings prompted extended debate rather than decisive action. Where escalation thresholds and governance pathways were predefined, monitoring signals triggered structured review, documented justification, and controlled intervention. A mature operating model enabled the organization to address drift without halting deployment or initiating emergency remediation: retraining and workflow adjustments were executed through controlled lifecycle processes, and audit inquiries were addressed using existing evidence rather than reactive reconstruction. Predefined escalation thresholds shortened the path from drift detection to corrective action, allowing performance changes to be addressed through structured lifecycle processes before regulatory or clinical escalation became necessary (Figure 4).

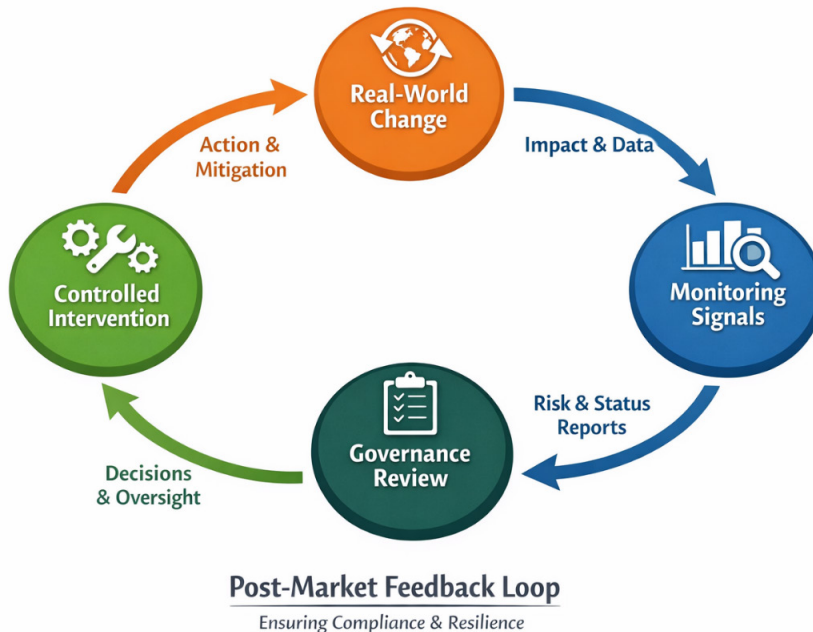


**Figure 4.** Post-market feedback loop linking monitoring signals to governance authority and controlled intervention. Source: Author-generated figures aligned to post-market surveillance and quality-system operating practices; created specifically for this manuscript

## 6. DISCUSSION: TRADEOFFS, MATURITY, AND ECONOMIC VIABILITY

Organizational maturity, not model performance alone, is often the practical limiter of scale. Durable deployment requires more than retrospective accuracy or one-time validation; it requires explicit structures for approval, local oversight, workflow integration, model surveillance, and intervention after go-live (Bedoya et al., 2022; Feng et al., 2022; Kelly et al., 2019; Roski et al., 2021).

The observational material synthesized in this manuscript suggests that structured operating models are associated with shorter remediation cycles, faster

scale transitions, and more disciplined response to drift and workflow issues. These associations are implementation-based and observational: they reflect patterns recorded during operational deployments rather than effects estimated from controlled studies, and no causal inference is warranted. Alternative explanations—including secular trends, regression to the mean, and site-level practice change—cannot be excluded. The figures in Table 3 should not be interpreted as statistically validated effect estimates; they are author-observed operational ranges presented to illustrate plausible magnitude and direction, not universal performance claims.

**Table 3.** Observed operational impact of structured operating models

| Operational Metric                   | Prior to Structured Operating Model | After Operating Model Implementation | Observed Range      |
|--------------------------------------|-------------------------------------|--------------------------------------|---------------------|
| Regulatory remediation cycles        | 6–12 months                         | 2–5 months                           | ~40–60% reduction   |
| Pilot-to-production transition       | 12–24 months                        | 8–14 months                          | ~25–40% improvement |
| Drift detection to response interval | Weeks–months                        | Days–weeks                           | 3–5× faster         |
| Clinical escalation resolution time  | Ad hoc and inconsistent             | Structured escalation pathways       | ~30–50% faster      |

Two limitations of this evidence base warrant explicit statement. First, selection bias: the deployments informing this synthesis were accessible precisely because of the author's operational and advisory roles, which means they are not a random or representative sample of regulated clinical AI deployments. Second, interpretation bias: as a participant-observer, the author cannot fully separate the analytical judgments made in this paper from the operational judgments made during the deployments themselves. In each case, the comparison group is a matched historical cohort at the same sites during the period immediately preceding AI-system activation, not a concurrent randomised control; temporal confounding, unobserved site-level changes, and the absence of blinding are therefore acknowledged limitations. Readers should treat the framework as hypothesis-generating and the reported associations as motivating evidence for the independent empirical validation described in Section 7.

Economic viability is likewise an operating-model question. Controlled change management, targeted human oversight, and earlier detection of workflow

or performance degradation can reduce avoidable rework, downtime, and manual-review burden. Operating capability is not administrative overhead; it is part of the safety, compliance, and scalability infrastructure required for production-grade AI. Operating models also make explicit the tradeoffs between speed, safety, and scale, enabling leadership to tune HITL intensity, validation thresholds, and monitoring cadence to institutional risk tolerance (Tabassi, 2023; U.S. Food and Drug Administration, 2025b).

## 7. EMPIRICAL VALIDATION PATHWAY

This section defines the agenda for future empirical validation. The propositions implied by the framework, and the patterns illustrated by the cases, are not yet empirically confirmed; the prospective, comparative studies outlined here are proposed as the means to test them. The current study is positioned as a hypothesis-generating and measurement-defining contribution, enabling future empirical validation by independent groups.

**Dataset design:** The unit of analysis is the deployment-site-month. Metrics span five domains: governance (escalation frequency, decision time); lifecycle (revalidation cycles, change frequency); clinical (adoption rate, override rate); monitoring (drift detection time, response time); and economic (cost per case, review burden).

**Statistical methods:** Appropriate methods include pre-post analysis, difference-in-differences, regression modeling, and survival analysis for time-to-event outcomes.

**Testable hypotheses:** Four hypotheses follow directly from the model: governance maturity reduces audit-resolution time (H1); lifecycle control reduces revalidation frequency (H2); HITL evolution reduces override rates (H3); and closed feedback loops reduce time to intervention (H4).

## 8. CONCLUSION

Scaling regulated clinical AI is best understood primarily as an operating-capability problem rather than solely as a model-performance problem. By integrating governance, lifecycle controls, clinical workflow design, adaptive HITL, and feedback loops into one executable system, organizations can convert pilot success into auditable, resilient, and economically sustainable deployments. The four-layer operating model presented here provides a practical and testable structure for that integration, grounded in documented deployment experience and aligned with existing regulatory, quality, and engineering frameworks. The empirical validation agenda in Section 7 defines the path from framework to evidence.

## 9. APPENDIX: PRACTICAL ARTIFACTS, METRICS, AND PLAYBOOKS

The governance charters, release-approval checklists, incident-response playbooks, and role definitions described here are provided as supplementary materials to support direct adoption and adaptation by practitioners, intended to be instantiated within existing quality management systems and tailored to local regulatory requirements while preserving the operating model's core execution logic.

### 9.1. Operating Artifacts

Four core artifacts operationalize the governance and lifecycle layers of the model. The governance charter defines scope, membership, decision authority, meeting cadence, quorum rules, and conflict-of-interest provisions. The release-approval checklist verifies evidence-package completeness, clinical sign-off, QA/RA review, the deployment plan, and the rollback plan before any system goes live. Change-control tiers classify updates as minor (documentation only, no revalidation), moderate (limited revalidation, internal approval), or major (full revalidation and updated regulatory submission). The incident-response playbook defines classification criteria, immediate mitigation actions, internal and external communication plans, CAPA initiation protocols, and the post-incident review process.

### 9.2. Metrics That Matter

Conventional accuracy metrics are necessary but not sufficient. The operating model requires parallel tracking across four categories. Safety metrics include false-negative and false-positive rates in clinically meaningful strata, near-miss signals, and override patterns. Operations metrics include review workload per 1,000 cases, time to escalation, time to mitigation, and rollback frequency. Compliance metrics include evidence-package completeness rates, audit-trail coverage, and CAPA cycle time. Economic metrics include cost per deployment episode, cost of oversight per 1,000 cases, remediation cost per incident, and time to measurable clinical value from go-live. Each category is tied to governance thresholds and defined response playbooks.

### 9.3. Example Escalation Thresholds

Escalation thresholds convert monitoring signals into governance decisions. Organizations should calibrate specific values — delta magnitudes, time windows, and consecutive-period criteria — to their clinical context, regulatory environment, and system risk profile. Four signal classes illustrate the mapping. A drift signal (statistically significant shift in key feature distributions sustained over a predefined number of monitoring days) triggers constrained use pending governance review. A performance signal (degradation beyond a predefined delta from validated baseline in one or more critical clinical

strata) triggers mandatory governance review and documented justification before continued operation. A workflow signal (clinician override rates exceeding a predefined threshold for two consecutive reporting periods) triggers a workflow-integration review and HITL recalibration. A safety signal (any sentinel event attributable to AI output) triggers immediate governance review, potential deployment pause, and CAPA initiation.

#### 9.4. Role Archetypes

Five role archetypes define the human operating infrastructure of the model, each carrying distinct accountability that cannot be absorbed by an adjacent function without creating governance gaps. The Clinical AI Product Owner defines and maintains intended-use scope, owns clinical outcome targets, and serves as the primary liaison between clinical and technical teams. The AI Quality Lead owns audit-readiness, evidence-package discipline, and the CAPA lifecycle. The ML Platform Lead owns model training pipelines, monitoring infrastructure, and reproducibility controls. Clinical Champions

own workflow-fit assessment at each site and serve as first-line escalation for workflow-level concerns. The Governance Chair maintains the integrity of decision-rights allocation and holds final escalation authority and rollback-trigger decision rights.

#### 9.5. Transferability Beyond Healthcare

The operating model generalizes to other regulated AI domains — finance, energy, transportation — where change control, auditability, and accountable decision rights are required. Domain adaptation primarily involves different risk taxonomies and regulatory artifacts; the core structure of governance, lifecycle management, integration, and feedback loops remains consistent. The primary adaptation point in non-healthcare domains is typically the substitution of clinical-workflow integration requirements with domain-specific human-oversight mechanisms appropriate to the consequential decision context.

### 10. GLOSSARY OF TERMS

| Term                                           | Definition                                                                                                                                                                                   |
|------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Clinical Artificial Intelligence (Clinical AI) | AI systems intended to support or inform clinical decision making, diagnosis, treatment, or healthcare workflows within environments carrying clinical liability.                            |
| Regulated Clinical AI                          | Clinical AI systems subject to formal regulatory oversight, quality management systems, post-market surveillance, and auditability requirements across their full operational lifecycle.     |
| Operating Model                                | The organizational system that defines how clinical AI is governed, deployed, monitored, changed, and scaled, including decision rights, lifecycle processes, and accountability mechanisms. |
| Governance and Decision Rights                 | The formal allocation of authority and accountability for approving deployment, authorizing changes, escalating risk, and responding to safety or compliance signals.                        |
| Technical Lifecycle Management                 | Controlled processes governing model development, validation, deployment, monitoring, retraining, and change management, ensuring traceability and regulatory defensibility.                 |
| Human-in-the-Loop (HITL)                       | A design pattern in which humans participate in oversight, validation, or decision making at defined lifecycle stages; treated here as an adaptive control that evolves with scale.          |
| Workflow Integration                           | Embedding AI outputs and controls directly into existing clinical workflows and systems of record to minimize cognitive burden and avoid parallel review processes.                          |
| Pilot Deployment                               | A limited-scope implementation conducted under heightened oversight and constrained exposure, typically not representative of scaled, regulated operation.                                   |

## Continued

| Term                        | Definition                                                                                                                                |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Production-Grade Deployment | An auditable, resilient deployment capable of sustained operation at scale under defined governance, lifecycle controls, and monitoring.  |
| Post-Market Monitoring      | Ongoing surveillance of performance, safety, and operational impact following deployment, linked to predefined governance responses.      |
| Performance Drift           | Degradation or change in model behavior over time due to shifts in data distributions, clinical practice, or upstream systems.            |
| Auditability                | The ability to reconstruct, justify, and defend system behavior and changes through documented evidence and logs.                         |
| Economic Sustainability     | The ability to operate and scale clinical AI without disproportionate growth in manual oversight, remediation, or revalidation costs.     |
| Operating-Model Maturity    | The extent to which governance, lifecycle discipline, and workflow integration are institutionalized and enforced within an organization. |

### Acknowledgement

The author thanks the clinical, engineering, governance, and operational teams associated with the deployment program described in this paper. Specific individual and institutional acknowledgements will be completed on the title page at final submission to preserve double-blind review.

### Conflict of Interest

The author holds an equity interest in the organization in which the deployment described in this paper took place. This interest is disclosed in the interest of full transparency. The case material has been anonymized and reported to reflect operational experience accurately, and the conceptual contributions of this paper are independent of this financial interest.

### Funding

This work was supported in part by a federally funded initiative. The specific funding agency and award number will be confirmed on the title page at final submission to preserve double-blind review.

### REFERENCES

- Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*, 17, 36. <https://doi.org/10.1186/s12911-017-0430-8>
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>
- Bedoya, A. D., Economou-Zavlanos, N. J., Goldstein, B. A., Young, A., Jelovsek, J. E., O'Brien, C., ... & Pencina, M. J. (2022). A framework for the oversight and local deployment of safe and high-quality prediction models. *Journal of the American Medical Informatics Association*, 29(9), 1631–1636. <https://doi.org/10.1093/jamia/ocac078>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Centers for Medicare & Medicaid Services. (n.d.). *Conditions of participation for hospitals*. <https://www.cms.gov/medicare/health-safety-standards/conditions-coverage-participation/hospitals>
- Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, 14(4), 532–550. <https://doi.org/10.2307/258557>
- European Union. (2017). *Regulation (EU) 2017/745 of the European Parliament and of the Council on*

- medical devices. <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32017R0745>
- Feng, J., Phillips, R. V., Malenica, I., Bishara, A., Hubbard, A. E., Celi, L. A., & Pirracchio, R. (2022). Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *npj Digital Medicine*, 5, 66. <https://doi.org/10.1038/s41746-022-00611-y>
- Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, 363(6433), 1287–1289. <https://doi.org/10.1126/science.aaw4399>
- International Medical Device Regulators Forum. (n.d.). *Artificial intelligence/machine learning-enabled medical device working group resources*. <https://www.imdrf.org/working-groups/artificial-intelligencemachine-learning-enabled>
- International Medical Device Regulators Forum. (2025, January 29). *Good machine learning practice for medical device development: Guiding principles (IMDRF/AIML WG/N88 FINAL:2025)*. [https://www.imdrf.org/sites/default/files/2025-01/IMDRF\\_AIML%20WG\\_GMLP\\_N88%20Final\\_0.pdf](https://www.imdrf.org/sites/default/files/2025-01/IMDRF_AIML%20WG_GMLP_N88%20Final_0.pdf)
- International Organization for Standardization. (2016). *ISO 13485:2016 medical devices—Quality management systems—Requirements for regulatory purposes*. ISO.
- International Organization for Standardization. (2019). *ISO 14971:2019 medical devices—Application of risk management to medical devices*. ISO.
- Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18–26. <https://doi.org/10.1007/s13162-020-00161-0>
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 195. <https://doi.org/10.1186/s12916-019-1426-2>
- Koch, L. M., Baumgartner, C. F., & Berens, P. (2024). Distribution shift detection for the postmarket surveillance of medical AI algorithms: A retrospective simulation study. *npj Digital Medicine*, 7, 120. <https://doi.org/10.1038/s41746-024-01085-w>
- National Institute of Standards and Technology. (n.d.). *Artificial intelligence risk management framework*. <https://www.nist.gov/itl/ai-risk-management-framework>
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- Rajkumar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
- Roski, J., Maier, E. J., Vigilante, K., Kane, E. A., & Matheny, M. E. (2021). Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association*, 28(7), 1582–1590. <https://doi.org/10.1093/jamia/ocab069>
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28* (pp. 2503–2511). <https://papers.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>
- Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., ... & O'Brien, C. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics*, 8(7), e15182. <https://doi.org/10.2196/15182>
- Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2), 345–352. <https://doi.org/10.1093/biostatistics/kxz041>
- Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST*

- AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- U.S. Department of Health and Human Services, Office for Civil Rights. (n.d.). *HIPAA privacy rule*. <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>
- U.S. Food and Drug Administration. (n.d.-a). *Digital Health Center of Excellence*. <https://www.fda.gov/medical-devices/digital-health-center-excellence>
- U.S. Food and Drug Administration. (n.d.-b). *Software as a medical device (SaMD)*. <https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd>
- U.S. Food and Drug Administration. (2021, October). *Good machine learning practice for medical device development: Guiding principles*. <https://www.fda.gov/media/153486/download>
- U.S. Food and Drug Administration. (2024, June 13). *Transparency for machine learning-enabled medical devices: Guiding principles*. <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles>
- U.S. Food and Drug Administration. (2025a, March 25). *Artificial intelligence in software as a medical device*. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device>
- U.S. Food and Drug Administration. (2025b, December 19). *Good machine learning practice for medical device development: Guiding principles*. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>
- U.S. Food and Drug Administration. (2025c, August). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff*. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence>
- Van Dort, B. A., Zheng, W. Y., Sundar, V., & Baysari, M. T. (2021). Optimizing clinical decision support alerts in electronic medical records: A systematic review. *Journal of the American Medical Informatics Association*, 28(1), 177–183. <https://doi.org/10.1093/jamia/ocaa279>
- World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. <https://www.who.int/publications/i/item/9789240029200>



Publisher's note: Eurasia Academic Publishing Group (EAPG) and AAKS remain neutral about jurisdictional claims in published maps and institutional affiliations.

**Open Access.** This article is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) licence, which permits copy and redistribute the material in any medium or format for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the licence terms. Under the following terms you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorsed you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nd/4.0/>.