



Article Information

Submitted: March 02, 2026
Approved: March 16, 2026
Published: March 18, 2026

How to cite this article: Spadacini D. Clinical MLOps: A Framework for Responsible Deployment and Observability of AI Systems in Cloud-Native Healthcare. *IgMin Res.* March 18, 2026; 4(3): 081-093. IgMin ID: igmin336; DOI: 10.61927/igmin336; Available at: igmin.link/p336

Copyright: © 2026 Spadacini D. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keywords: AI systems; MLOps; Clinical practice; Artificial intelligence

Research Article



Clinical MLOps: A Framework for Responsible Deployment and Observability of AI Systems in Cloud-Native Healthcare

Daniel Spadacini*

DevOps & Support Engineer, Nuvyta, Milan, Italy

***Correspondence:** Daniel Spadacini, DevOps & Support Engineer, Nuvyta, Milan, Italy, Email: daniel@spadacini.eu



Abstract

Machine learning operations (MLOps) practices have reached a notable level of maturity within general-purpose software engineering. Pipelines are standardized, monitoring is automated, and deployment patterns are well rehearsed. Yet when these same practices are transferred into clinical environments, their adequacy becomes less certain. The healthcare domain introduces operational and ethical pressures that conventional MLOps frameworks were not originally designed to address.

Clinical AI systems operate under constraints that extend beyond typical production settings. Stringent data protection regimes — including GDPR and the EU AI Act — shape how data can be processed and retained. Model drift is not merely a statistical inconvenience; it may alter clinical decisions with tangible consequences for patient outcomes. Regulatory compliance demands comprehensive, tamper-evident audit trails. At the same time, decision-critical contexts impose a clear expectation of human oversight. In other words, technical robustness alone is insufficient; governance and accountability become integral system properties.

This paper introduces a Clinical MLOps framework that systematically augments conventional pipelines with healthcare-specific requirements. The framework is structured around four layered components: (1) privacy-preserving deployment patterns, (2) clinical observability mechanisms, (3) compliance-oriented audit trail architecture, and (4) human-in-the-loop governance protocols. Each layer addresses a distinct operational risk while remaining interoperable with established cloud-native tooling.

To evaluate the framework, we construct a demonstrative end-to-end pipeline using the MIMIC-IV dataset — a large, de-identified electronic health record repository from Beth Israel Deaconess Medical Center. Within this pipeline, we implement a patient deterioration prediction model and apply Clinical MLOps controls systematically at every stage, from data ingestion to post-deployment monitoring. This controlled implementation enables us to examine how standard MLOps tooling behaves under clinical constraints.

Our findings indicate that widely adopted MLOps practices, while technically sound, leave critical governance and compliance gaps when applied in healthcare contexts. The proposed Clinical MLOps framework addresses these deficiencies without requiring impractical infrastructure changes. Importantly, it remains compatible with cloud-native architectures, making adoption feasible within contemporary health IT ecosystems. The implications extend to healthcare AI governance more broadly, particularly regarding the operational interpretation of the EU AI Act in clinical settings.

Introduction

The integration of artificial intelligence into clinical practice has accelerated markedly over the past decade. Predictive models for patient deterioration, diagnostic support systems, and resource allocation algorithms are no longer confined to research prototypes; they are increasingly embedded in everyday hospital workflows (Topol, 2019). Yet the operational backbone required to sustain these systems — to keep them safe, reliable, and fair over time — has drawn far less sustained scholarly scrutiny.

MLOps, or machine learning operations, emerged precisely to address the well-known gap between model development

and production deployment. Drawing on DevOps principles, it formalizes versioning strategies, continuous integration and delivery (CI/CD) pipelines, monitoring infrastructures, and retraining cycles for machine learning systems [1,2]. In commercial settings, this toolkit has proven effective. But here's the tension: most canonical MLOps frameworks were designed for enterprise environments where system failure typically carries financial or reputational costs — rarely clinical ones.

Healthcare AI systems operate within a markedly different risk landscape. A patient deterioration model that degrades quietly due to population shift does not merely reduce predictive accuracy; it may delay a life-saving intervention.

A diagnostic system that embeds demographic bias can reinforce inequities already present in care delivery. And consider documentation: an audit trail that omits the model version used at inference time may render a system non-compliant under the European Union’s Artificial Intelligence Act [3]. The Act classifies clinical decision-support tools as high-risk AI applications, subject to rigorous documentation, transparency, and oversight obligations. Compliance, then, is not a bureaucratic afterthought; it is operationally central.

This divergence between general-purpose MLOps and the concrete requirements of clinical environments motivates the present study. We introduce the notion of Clinical MLOps — a structured extension of established MLOps practices tailored specifically to healthcare AI deployment. The proposed four-layer framework addresses the operational constraints unique to clinical contexts in a systematic manner. Importantly, it does not discard existing tooling; rather, it supplements conventional pipelines with domain-specific controls that are otherwise missing. The distinction may seem subtle, but it carries practical weight.

The remainder of the paper proceeds as follows. Section 2 surveys relevant literature on MLOps, healthcare AI governance, and regulatory regimes. Section 3 details the proposed Clinical MLOps framework and its four components. Section 4 outlines the demonstrative pipeline built using the MIMIC-IV dataset and presents the results of applying the framework in practice. Section 4.9 presents a secondary validation on a 30-day readmission prediction task to assess framework generalizability across clinical settings. Section 5 examines implications, limitations, and avenues for further investigation. Section 6 concludes the paper.

A note on manuscript scope: this paper is best classified as a framework proposal with demonstrative empirical instantiation. The four-layer Clinical MLOps framework described in Section 3 is a conceptual and architectural contribution grounded in existing standards and literature. The pipeline presented in Section 4 is a concrete demonstrative implementation: all components listed in Table 1 were operationally instantiated using the technologies specified,

and the results reported reflect actual pipeline executions on MIMIC-IV data. The drift scenario in Section 4.7, however, employs a synthetic perturbation applied to the held-out test set rather than a naturally occurring temporal shift; this distinction is explicitly acknowledged in that section. Readers are encouraged to interpret the empirical findings as a demonstration of framework viability rather than a large-scale clinical validation study. Broader generalizability claims are intentionally bounded accordingly.

Background and related work

MLOps: Principles and limitations

MLOps emerged as a distinct discipline in response to a persistent, almost frustrating reality: deploying and maintaining machine learning models in production is far more complex than training them in isolation [1]. What looks elegant in a notebook can become brittle in a live environment. Over time, the field consolidated a set of engineering practices intended to stabilize that transition — experiment tracking to avoid “mystery models,” versioning to ensure reproducibility, pipeline orchestration to coordinate dependencies, monitoring to detect degradation, and automated retraining to close the feedback loop.

Kreuzberger, et al. [2] offer a structured taxonomy of MLOps principles, distinguishing between technical components and organizational ones. On the technical side, we encounter artifact stores, feature stores, and model registries — the infrastructure that keeps assets traceable and environments consistent. On the organizational side, the focus shifts to team topology, governance routines, and deployment maturity levels. This distinction matters. A well-architected pipeline can still fail if ownership is unclear or oversight is diffuse. Conversely, strong governance without technical rigor produces its own fragility. The two layers are interdependent, whether teams explicitly acknowledge it or not.

Several prominent platforms — including MLflow, Kubeflow, Apache Airflow, and Databricks — operationalize these principles with considerable sophistication. They support experiment management, automate workflows, and orchestrate deployments across cloud-native infrastructures with impressive efficiency. From a systems engineering perspective, the tooling is mature. But maturity does not imply completeness.

As Paley, et al. [4] note in their systematic review of machine learning deployment challenges, these platforms remain largely agnostic to domain-specific constraints. Security controls, regulatory compliance mechanisms, and interpretability requirements are typically treated as adjacent concerns — configurable add-ons rather than embedded design primitives. In many industries, that separation

Table 1: Clinical MLOps pipeline technology stack and layer mapping.

Component	Technology	Clinical MLOps Layer
Data extraction & versioning	DVC + PostgreSQL	Layer 3 (Lineage)
Feature engineering	Apache Spark / PySpark	Layer 1 (Minimization)
Experiment tracking	MLflow	Layer 3 (Audit Trail)
Model training	scikit-learn / XGBoost	Layer 2 (Uncertainty)
CI/CD pipeline	GitHub Actions	Layer 3 (Change Mgmt)
Model serving	FastAPI + Docker	Layer 1 (Isolation)
Monitoring	Grafana + Prometheus	Layer 2 (Observability)
Drift detection	Evidently AI	Layer 2 (Drift)
Secrets management	HashiCorp Vault	Layer 1 (Encryption)
Audit logging	Immutable object store (S3)	Layer 3 (Logging)
Human review interface	Custom dashboard	Layer 4 (HITL)

may be manageable. In high-stakes domains, it becomes more problematic. When compliance and interpretability are externalized, they risk becoming reactive rather than structural. And that subtle shift, while easy to overlook, can have significant downstream implications.

Healthcare AI: Deployment challenges

The deployment of AI systems in clinical environments introduces a layer of complexity that goes well beyond what is typically encountered in commercial settings. In retail or advertising, performance degradation might translate into lost revenue or reduced engagement. In healthcare, the stakes are different — and sharper.

Finlayson, et al. [5] document the phenomenon of dataset shift in clinical AI, showing that models trained on historical electronic health record (EHR) data often degrade when introduced into new hospital systems or even when used in the same institution over time. Why? Clinical practice evolves. Patient populations change. Documentation habits shift. Even subtle modifications in data collection protocols can ripple through a model's behavior. What makes this especially concerning is that such drift may not immediately surface in aggregate performance metrics. Instead, it can manifest quietly — as systematically worse outcomes for particular patient subgroups. The degradation is real, but not always visible at first glance.

Bias presents a related, and equally troubling, dimension. Obermeyer, et al. [6] demonstrated that a widely deployed commercial healthcare algorithm exhibited racial bias, allocating fewer resources to Black patients than to equally sick White patients. The issue was not a trivial calibration error; it reflected structural inequities embedded within the proxy variables used for prediction. The study made something uncomfortably clear: conventional performance metrics such as AUC-ROC do not suffice for governing clinical AI. A model can appear statistically strong while perpetuating inequity. Fairness metrics disaggregated across demographic attributes are therefore not optional add-ons; they are essential evaluative components.

Building on this perspective, Wiens, et al. [7] argued for a more comprehensive evaluation framework for clinical AI — one that extends beyond predictive accuracy to include clinical utility, fidelity of implementation, and sustained performance monitoring. Their argument shifts the focus from model-centric validation to system-level accountability. In practice, this implies infrastructure capable of tracking not just predictions, but consequences; not just metrics, but impact.

Taken together, these contributions point toward a common conclusion. Clinical AI cannot be governed solely

through conventional validation workflows. It requires continuous, structured oversight embedded within the operational lifecycle itself — precisely the gap that a Clinical MLOps framework seeks to address.

Regulatory frameworks: GDPR and the EU AI act

The regulatory landscape for healthcare AI in Europe has undergone significant development over the past decade. The General Data Protection Regulation [8] laid the groundwork by defining strict conditions for processing health data. Explicit consent, data minimization, purpose limitation, and the right to explanation for automated decisions are not abstract legal principles; they carry direct technical implications. For AI systems handling patient data, this means encryption both at rest and in transit, detailed access logging, mechanisms that support explainability, and enforceable data retention controls. Compliance, in this sense, is engineered — not merely declared.

The EU AI Act [3] advances this regulatory trajectory by introducing a risk-based classification scheme for AI systems. Clinical decision-support tools fall under the category of high-risk applications. This designation entails conformity assessments, extensive technical documentation, post-market monitoring, and clearly defined human oversight mechanisms. Notably, Article 9 mandates the establishment and maintenance of risk management systems across the entire lifecycle of high-risk AI systems. The requirement is continuous rather than episodic. It aligns closely with what a Clinical MLOps framework seeks to institutionalize: structured oversight embedded within operational workflows.

Taken together, these regulatory instruments expand the compliance perimeter in ways that conventional MLOps architectures were not originally designed to accommodate. A deployment pipeline may be technically refined — orchestrated, automated, cloud-native — yet still fall short. Without immutable audit logging, systematic model card generation, or demographic fairness monitoring, it cannot reasonably be considered aligned with the EU AI Act's expectations. Technical sophistication alone does not equate to regulatory adequacy. In healthcare AI, governance must be built into the pipeline itself, not layered on after deployment.

Research gap

The literature reviewed thus far points to a noticeable disconnect. On one side, MLOps practices have matured considerably in general-purpose engineering contexts. On the other hand, the risks and operational fragilities of healthcare AI deployment have been extensively documented. Yet a systematically articulated framework that integrates these two strands remains absent. The bridge, in other words, is still under construction.

Existing contributions to responsible AI in healthcare — including Char, et al. [9] and Topol [10] — provide valuable ethical guidance. They articulate principles such as transparency, accountability, and patient-centered design with clarity and urgency. However, these proposals largely remain at the normative level. They outline what ought to be achieved, but stop short of specifying how those principles should be embedded into day-to-day engineering workflows. The operational layer is implied rather than formalized.

This gap is not merely academic. Without concrete implementation pathways, ethical aspirations risk becoming aspirational checklists rather than enforceable practices. What does accountability look like in a CI/CD pipeline? How is transparency encoded in logging infrastructure? These questions require technical answers, not only conceptual ones.

The present paper responds directly to this need. It proposes a concrete and implementable framework that is grounded simultaneously in regulatory mandates and established engineering practices. Rather than introducing new abstract principles, it translates existing obligations and governance expectations into structured operational controls — making them actionable within the lifecycle of healthcare AI systems.

The clinical MLOps framework

The Clinical MLOps framework is organized into four interdependent layers, each addressing a distinct set of requirements that arise at the intersection of MLOps practice and healthcare AI governance. The layers are designed to be composable: they can be implemented progressively, with each layer building upon the infrastructure established by the previous one. Figure 1 illustrates the overall architecture.

Layer 1 — Privacy-preserving deployment patterns

The first layer establishes a baseline that is, in healthcare, non-negotiable: patient data must be handled in strict accordance with GDPR and applicable national data protection law. This is not simply a compliance checkbox. It is an architectural constraint that shapes how models are designed, deployed, and maintained. If privacy protections are bolted on after deployment, the system is already misaligned.

This layer is structured around four interrelated design patterns.

- **Data minimization at inference:** Production models should operate on the smallest possible feature set necessary to generate a prediction. That principle sounds straightforward, yet it demands discipline in practice. Feature selection must be explicitly documented and auditable; every retained variable requires justification. Moreover, feature extraction

pipelines should avoid persisting intermediate representations of patient data beyond the inference window. Temporary transformations are acceptable; lingering artifacts are not. Minimization, here, is both a technical safeguard and a governance signal.

- **Encryption and secrets management:** All patient data in transit must be encrypted using TLS 1.3 or an equivalent protocol. Data at rest must rely on AES-256 or a comparably robust standard. However, encryption strength alone is insufficient if key management is ad hoc. Encryption keys must be administered through a dedicated secrets management service — such as AWS Secrets Manager or HashiCorp Vault — with rotation policies enforced programmatically rather than manually. Automation reduces the risk of silent lapses; in security engineering, predictability is protection.
- **Inference environment isolation:** Model serving containers should operate within network-isolated environments equipped with strict egress controls to prevent unauthorized data exfiltration. Isolation reduces the attack surface and limits the blast radius in the event of compromise. In parallel, container images must undergo vulnerability scanning as part of the CI/CD workflow. Security review is therefore not a one-time gate; it becomes a recurring operational checkpoint embedded directly into deployment routines.
- **Consent and purpose tracking:** When inference relies on patient data collected under specific consent conditions, the deployment pipeline must verify that the intended use aligns with the documented scope of consent before execution. This requirement effectively transforms consent from a static record into a runtime constraint. The system must “know” why it is permitted to act — and refrain when the purpose diverges.

Together, these design patterns translate legal obligations into enforceable technical controls. Privacy, in this layer, is not treated as a peripheral concern. It is operationalized as a core property of the deployment architecture itself.

Layer 2 — Clinical observability mechanisms

Standard MLOps monitoring typically concentrates on technical indicators — latency, throughput, error rates — alongside statistical measures such as prediction distributions and feature drift. These signals are necessary; without them, operational stability quickly erodes. Yet in clinical environments, they are not sufficient. Observability must extend beyond system health to encompass signals that are clinically meaningful.

Clinical observability, therefore, broadens the monitoring surface in several critical ways.

- **Clinical outcome linkage:** Where outcome data becomes available within an acceptable latency window, model predictions should be systematically linked to observed clinical outcomes. This enables ongoing calibration assessment rather than one-time validation. Achieving this requires direct integration between the inference pipeline and the clinical data repository. The technical implication is straightforward but non-trivial: prediction logs must be persistently joinable with outcome records. Calibration, in this context, is not a retrospective academic exercise; it becomes a live operational metric.
- **Demographic fairness monitoring:** Model performance indicators — including precision, recall, and F1 score — must be computed and monitored separately across protected attributes such as age group, sex, and ethnicity, where available. Aggregated performance can obscure subgroup disparities. Threshold-based alerts should be configured to activate when performance gaps exceed predefined tolerances. The goal is not merely to detect bias retrospectively, but to surface disparities early enough to prompt structured review.
- **Concept drift detection:** Statistical drift detection methods — including ADWIN, Page-Hinkley, and the Population Stability Index — should be applied to both input feature distributions and prediction output distributions. In clinical settings, however, drift detection must be coupled with governance safeguards. Alerts should trigger human review rather than automated retraining. Automated adaptation to distributional shift may be efficient in commercial systems; in healthcare, the unintended consequences of unsupervised retraining can be significant. Human oversight, therefore, functions as a stabilizing control.
- **Uncertainty quantification:** Production models should expose uncertainty estimates alongside point predictions. These may take the form of calibration metrics or conformal prediction intervals. Presenting uncertainty is not an admission of weakness; it is a communication of epistemic limits. Clinicians should be explicitly informed when the model operates in low-confidence regimes, enabling informed judgment rather than blind reliance.

Taken together, these extensions redefine observability in clinical AI. Monitoring is no longer confined to system performance metrics; it becomes a structured mechanism for safeguarding clinical validity, equity, and interpretability over time.

Layer 3 — Compliance-oriented audit trail architecture

The EU AI Act stipulates that high-risk AI systems must

maintain logs capable of reconstructing system behavior over an appropriate retention period. In clinical environments, this obligation becomes concrete and technically demanding. Logging is no longer a debugging convenience; it is a regulatory artifact.

In practice, this requirement translates into several structured audit trail controls.

- **Immutable inference logging:** Every model inference must generate a durable log entry containing, at minimum, the timestamp, the exact model version (including a hash of the training dataset), pseudonymized input feature values, the output prediction with an associated confidence score, and the identity of the requesting system or authenticated user. Immutability is central. Logs must be tamper-evident, ensuring that retrospective review cannot be compromised by post hoc modification. Reconstruction of a single prediction event should be possible without ambiguity.
- **Model lineage tracking:** Each deployed model must possess a fully traceable lineage. This includes the specific training dataset version, the preprocessing pipeline version, hyperparameter configurations, the training environment, and the final deployed artifact. The lineage record should be stored within an immutable model registry. Traceability here functions as both a quality assurance mechanism and a compliance safeguard. If a clinical outcome is questioned, the underlying computational pathway must be reproducible.
- **Change management integration:** Model deployment events should be formally integrated into the organization's change management framework. Releases are not informal updates; they are governed modifications subject to documented review and approval procedures aligned with clinical governance standards. Treating deployment as a controlled change event reinforces accountability and ensures multidisciplinary oversight.
- **Data retention and right to erasure:** Audit logs that contain patient-linked elements must comply with GDPR retention constraints and support data erasure requests. This necessitates architectural foresight. Patient identifiers should be logically separated from inference metadata at the point of design, enabling selective deletion without compromising system-level audit integrity. The separation is not merely technical housekeeping; it is a structural prerequisite for lawful operation.

Collectively, these requirements formalize auditability

as a first-order system property. In high-risk clinical AI applications, traceability is not an optional enhancement — it is an operational obligation embedded in the lifecycle of the system itself.

Layer 4 — Human-in-the-loop governance protocols

Article 14 of the AI Act establishes that high-risk AI systems must be designed in a way that meaningfully enables human intervention. Oversight is not symbolic; it must be actionable. Humans should not merely observe system outputs — they must be able to understand, question, and, when necessary, interrupt them. In clinical environments, this requirement becomes operationally precise.

Clinical MLOps translates these oversight obligations into concrete governance controls.

- **Override mechanisms:** Every AI-generated recommendation must be overridable by a qualified clinician without procedural friction. The interface should not penalize disagreement or require justification beyond standard clinical documentation norms. Override events must be systematically logged, and override frequency must be monitored as an indicator of model trust, usability, and potential misalignment with clinical workflow. A consistently high override rate may signal calibration drift or contextual blind spots; a zero-override pattern may raise different questions. Both extremes warrant examination.
- **Escalation thresholds:** Automated systems must define explicit escalation thresholds — predefined conditions under which AI-assisted recommendations are suspended, and cases are routed to human review. These thresholds may be triggered by uncertainty estimates, drift alerts, or anomalous prediction patterns. Crucially, escalation criteria must be clinically validated and formally documented. Suspension is not a failure state; it is a safeguard embedded within system design.
- **Periodic human review:** Irrespective of automated monitoring signals, model performance metrics, and governance documentation must undergo review by qualified personnel at predetermined intervals, such as quarterly cycles. This review should encompass calibration status, fairness assessments, incident logs, and documentation completeness. Outcomes must be recorded and linked directly to the model's lifecycle registry. Scheduled review reinforces the principle that oversight is continuous rather than reactive.
- **Incident response protocol:** A formally documented incident response protocol must specify actions to

be taken in the event of confirmed model failure or harmful behavior. This includes containment measures, communication pathways, and — where required by clinical ethics or regulation — patient notification procedures. Clear role delineation is essential. When responsibility is diffuse, response becomes delayed; when response is delayed, harm may escalate.

Together, these mechanisms operationalize human oversight as a structured component of system governance. The objective is not to diminish automation, but to ensure that automated assistance remains bounded by accountable, informed human control.

Demonstrative pipeline: Patient deterioration prediction with MIMIC-IV

Dataset description

MIMIC-IV (Medical Information Mart for Intensive Care, version 2.2) is a large, publicly accessible database containing de-identified health data from patients admitted to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, between 2008 and 2019 [11]. The scale alone is substantial: over 40,000 ICU stays and approximately 190,000 hospital admissions. But scale is only part of the story.

The dataset combines structured elements — vital signs, laboratory measurements, medication administrations, procedure codes — with unstructured artifacts such as clinical notes and radiology reports. This duality makes it particularly suitable for realistic pipeline construction. It resembles the messy, heterogeneous data landscapes encountered in operational hospital systems. Clean in governance, complex in structure. That combination is rare.

Access to MIMIC-IV is mediated through PhysioNet under a formal Data Use Agreement (DUA). Credentialing is required, alongside completion of a recognized human research data use training program. These access conditions are not incidental; they reinforce a culture of responsible data stewardship even in research settings. Importantly, because the dataset is fully de-identified in accordance with HIPAA Safe Harbor guidelines, its use in the present pipeline does not necessitate additional Institutional Review Board (IRB) approval. The regulatory footing is therefore clear.

The selection of MIMIC-IV as the validation substrate for the Clinical MLOps framework is deliberate. First, it is widely adopted within the clinical AI research community, enabling comparability with prior work. Second, its structural complexity approximates real-world electronic health record systems, stress-testing pipeline design choices. Third, and perhaps most pragmatically, its public availability ensures that the proposed pipeline remains reproducible. Reproducibility is not a rhetorical commitment here; it is technically achievable.

In short, MIMIC-IV provides a realistic yet governed environment in which Clinical MLOps controls can be implemented, examined, and scrutinized without ambiguity about data provenance or regulatory posture.

Prediction task

The demonstrative pipeline implements a binary classification task: prediction of in-hospital mortality within 48 hours for adult ICU patients, based on the first 24 hours of available clinical data. This task is well-established in the clinical AI literature [12] and provides a realistic context for demonstrating Clinical MLOps controls, as it involves: sensitive patient data, time-critical predictions, potential for demographic bias, and clear clinical consequences of model failure. Cohort construction followed pre-specified inclusion and exclusion criteria applied to MIMIC-IV v2.2. Inclusion criteria required: (1) adult patients aged 18 years or older; (2) first ICU admission during the study period 2008–2019; and (3) ICU length of stay of at least 24 hours, to ensure sufficient data availability for the feature extraction window. Exclusion criteria removed: (1) admissions with incomplete vital-sign records across more than 50% of the 17-variable feature set within the first 24-hour window; (2) patients with missing outcome labels; and (3) repeat ICU admissions within the same hospitalisation, retaining only the index admission. After applying these criteria, the final analysis cohort comprised 23,841 ICU admissions. Outcome prevalence (in-hospital mortality within 48 hours) was 11.3% ($n = 2,694$), reflecting the moderate class imbalance characteristic of ICU deterioration tasks. The temporal train/test split yielded 16,751 admissions in the training set (2008–2016) and 7,090 in the held-out test set (2017–2019). No separate validation fold was used for hyperparameter tuning; a five-fold cross-validation scheme was applied within the training set for this purpose.

Pipeline architecture

The pipeline is implemented on a cloud-native infrastructure using containerized microservices. The technology stack reflects current industry practice and is compatible with the Clinical MLOps framework:

Feature engineering and data minimization

In alignment with the data minimization principle articulated in Layer 1, we extract a deliberately parsimonious feature set of 17 variables from MIMIC-IV. The selected variables include: age, sex, primary ICD-10 diagnosis category, Glasgow Coma Scale (GCS) score, mean arterial pressure, heart rate, respiratory rate, temperature, SpO₂, serum creatinine, serum lactate, albumin, bilirubin, platelet count, INR, vasopressor use flag, and mechanical ventilation flag.

The selection is not arbitrary. It draws from established

clinical severity scoring frameworks — notably SOFA and APACHE II — and aligns with prior validation work demonstrating predictive relevance for patient deterioration [12]. In other words, the feature space is intentionally constrained yet clinically grounded. More variables might increase apparent model flexibility, but constraints here functions as discipline. It reduces privacy exposure and supports interpretability without sacrificing clinical signal.

Feature extraction is implemented through a time-windowed aggregation pipeline built in PySpark. Temporal aggregation is necessary to reconcile heterogeneous sampling frequencies across vital signs and laboratory measurements. The aggregation logic itself is version-controlled alongside the training codebase. At training time, the hash of the aggregation script is recorded within the model metadata, thereby satisfying the lineage requirements specified in Layer 3. This ensures that feature engineering steps are not merely documented but cryptographically traceable.

Missing values are addressed through median imputation. Crucially, imputation statistics are computed exclusively on the training set to prevent data leakage into validation or test partitions. While median imputation is methodologically simple, its transparency supports reproducibility and auditability. More complex imputation schemes could be considered, yet simplicity here enhances traceability — an important trade-off in regulated environments. Missing data rates varied considerably across features. Laboratory values showed the highest rates of missingness: serum lactate (38.2%), albumin (31.7%), and bilirubin (29.4%) were the most affected, reflecting the selective-ordering patterns typical of ICU clinical practice. Vital signs and binary flags were largely complete: heart rate (0.4%), respiratory rate (1.1%), mean arterial pressure (1.8%), SpO₂ (2.3%), mechanical ventilation flag (0.2%), and vasopressor use flag (0.3%). GCS score had a missingness rate of 7.6%, attributable predominantly to sedated patients. These rates were computed on the training set only and applied uniformly at test time via the frozen median values. The outcome class imbalance (11.3% positive rate) was addressed using class-weight balancing in the XGBoost objective function (scale_pos_weight set to the ratio of negative to positive instances in the training fold), rather than resampling, to preserve the original temporal structure of the data.

The resulting dataset is compact, clinically interpretable, and fully versioned from raw extraction through aggregation. This disciplined construction establishes a controlled foundation for subsequent model training and governance controls.

Model training and uncertainty quantification

A gradient-boosted decision tree model, implemented via

XGBoost, is trained on hospital admissions recorded between 2008 and 2016 — roughly 70% of the available dataset. Evaluation is conducted on a temporally distinct holdout set comprising admissions from 2017 to 2019. This temporal partitioning is intentional. Random splits may inflate apparent performance; temporal splits better approximate real deployment conditions, where models confront future data shaped by evolving clinical practice and shifting patient populations. In effect, the split introduces a controlled test of distributional drift across time. Model selection followed a two-stage procedure. In the first stage, three candidate algorithms were evaluated on the training set using five-fold cross-validation stratified by outcome: logistic regression (L2 regularisation, $C = 1.0$), a random forest ($n_estimators = 300$, $max_depth = 10$), and XGBoost. XGBoost was selected based on superior mean cross-validated AUC-ROC (0.861, 95% CI 0.853–0.869), with logistic regression and random forest yielding 0.832 and 0.849, respectively. In the second stage, XGBoost hyperparameters were tuned via random search (100 iterations) over the following ranges: $n_estimators$ [100, 500], max_depth [3, 8], $learning_rate$ [0.01, 0.3], $subsample$ [0.6, 1.0], $colsample_bytree$ [0.6, 1.0], and min_child_weight [1, 10]. The final selected configuration was: $n_estimators = 400$, $max_depth = 5$, $learning_rate = 0.05$, $subsample = 0.8$, $colsample_bytree = 0.8$, $min_child_weight = 3$. All tuning was conducted exclusively within the training set; the held-out temporal test set was accessed only once for final evaluation.

To comply with the uncertainty quantification requirements defined in Layer 2, we apply post-hoc isotonic regression calibration [13] to transform raw model scores into calibrated probability estimates. Calibration is not cosmetic. In clinical decision-making, the reliability of predicted probabilities directly influences threshold selection and risk communication. Calibration quality is assessed using the Expected Calibration Error (ECE) and visualized through reliability diagrams, which expose systematic overconfidence or underconfidence across probability bins.

Beyond calibration, we incorporate conformal prediction methods [14] to generate prediction sets with a guaranteed 90% coverage level. This approach produces explicit uncertainty bounds for each case rather than a single point estimate. For clinicians, this distinction matters. A prediction accompanied by a quantifiable uncertainty interval communicates epistemic limits in a structured manner. It signals when the model is operating within familiar territory — and when it is not.

Together, temporal validation, probabilistic calibration, and conformal prediction embed statistical rigor within the training and evaluation workflow. The model is not merely optimized for discrimination; it is structured to communicate risk with calibrated confidence under realistic deployment conditions.

Fairness evaluation

Consistent with Layer 2 demographic fairness monitoring requirements, we compute model performance metrics disaggregated by sex (male/female) and age group (18–44, 45–64, 65–79, 80+). We evaluate three fairness metrics: equalized odds [15], demographic parity difference, and predictive parity. Results are presented in Table 2.

The results indicate broadly consistent performance across demographic subgroups, with the most notable gap observed between the youngest (18–44) and oldest-middle (65–79) age groups on AUC-ROC (0.821 vs. 0.853). This gap, while modest, would trigger a Layer 2 fairness alert under a monitoring threshold of ± 0.03 AUC points, prompting review of potential systematic differences in data quality or clinical management patterns for younger ICU patients. The three fairness metrics were computed across both stratification attributes and are reported here with 95% confidence intervals derived via 1,000-iteration bootstrap resampling. For the sex stratification, the equalized odds difference (defined as the maximum absolute difference in true positive rate or false positive rate across groups) was 0.029 (95% CI 0.012–0.046), indicating near-parity in error rate distributions. The demographic parity difference (difference in positive prediction rates at the operating threshold of 0.40) was 0.031 (95% CI 0.014–0.048). The predictive parity difference (difference in positive predictive values) was 0.027 (95% CI 0.008–0.046). For the age stratification, the equalized odds difference was 0.040 (95% CI 0.021–0.059), driven primarily by the 18–44 group, which showed a lower true positive rate (0.748 vs. overall 0.781). Demographic parity difference across age groups was 0.044 (95% CI 0.025–0.063), and predictive parity difference was 0.080 (95% CI 0.056–0.104). The predictive parity gap across age groups exceeds the other metrics and is attributable in part to the lower baseline prevalence of the outcome in younger patients, which affects PPV independently of model calibration. This pattern is consistent with prior ICU literature and does not necessarily indicate bias in the causal sense; however, it warrants contextual review before deployment. All three metrics remain within ranges that would not trigger automated retraining under the framework's governance rules, though the age-related predictive parity gap would generate a monitoring alert requiring human assessment.

Table 2: Model performance metrics disaggregated by demographic subgroup (held-out temporal test set, 2017–2019).

Subgroup	AUC-ROC	Sensitivity	Specificity	PPV
Overall	0.847	0.781	0.833	0.612
Male	0.851	0.793	0.828	0.624
Female	0.839	0.764	0.841	0.597
Age 18–44	0.821	0.748	0.861	0.541
Age 45–64	0.844	0.779	0.834	0.608
Age 65–79	0.853	0.788	0.829	0.621
Age 80+	0.838	0.761	0.842	0.589

Drift simulation and monitoring

To evaluate the Layer 2 drift detection mechanisms under controlled conditions, we simulate a plausible concept drift scenario. Specifically, we introduce an artificial distributional shift in the test set by modifying the mechanical ventilation flag distribution to mirror documented changes in ICU ventilation practices observed during the COVID-19 pandemic [16]. The objective is not to recreate the pandemic in full complexity, but to approximate a realistic structural shift in care patterns — the kind that models trained on historical data might struggle to absorb. The perturbation was implemented as follows. In the unmodified test set (2017–2019), the proportion of ICU admissions with the mechanical ventilation flag set to 1 was 38.4%. In alignment with the Grasselli, et al. [16] report, which documented invasive mechanical ventilation rates of approximately 88% among critically ill COVID-19 patients admitted to Lombardy ICUs, we upsampled the mechanical ventilation flag to 72% prevalence in a modified test subset ($n = 2,000$ randomly selected admissions from the held-out set), representing a relative increase of approximately 87% from the baseline rate. This magnitude was chosen to reflect a clinically plausible worst-case shift rather than an incremental one: ICU ventilation protocols underwent rapid and substantial revision during pandemic surges, and a monitoring system must be able to detect even extreme structural changes reliably. The perturbation was applied exclusively to the mechanical ventilation flag; all other feature distributions were held constant. This controlled scope is important: it ensures that the observed performance degradation can be attributed to the ventilation feature shift, but it also constitutes a limitation. In real distributional shifts, correlated covariates would change simultaneously. Modifying only one feature may underestimate the sensitivity of the full feature space to realistic distributional perturbations, and may also fail to capture interactions between ventilation status and other variables (such as vasopressor use, SpO₂, or respiratory rate) that would co-shift in an actual pandemic surge. The scenario, therefore, demonstrates the drift detection mechanism's sensitivity to targeted feature perturbations; it should not be interpreted as evidence that the framework would perform equivalently under more complex, multi-covariate distributional shifts. A more rigorous evaluation would require temporal validation against naturally occurring shifts in real clinical data. This limitation is acknowledged in Section 5.3 and represents a priority for future work.

We then apply ADWIN drift detection [17] to the distribution of prediction scores and compute the Population Stability Index (PSI) on the input feature distributions. This dual-layer approach allows us to monitor both upstream input shifts and downstream output instability. Drift, after all, can manifest in different parts of the pipeline. Observing only one layer would be insufficient.

Under continuous monitoring with a 15-minute evaluation window, drift detection triggers identify the introduced shift within 48 hours of simulated operation. The performance impact is measurable: AUC-ROC declines from 0.847 to 0.793 under the shifted distribution — a 6.4% reduction. While numerically modest, this decrease surpasses the predefined clinical alert threshold of 5% AUC degradation. The system therefore activates the escalation protocol, suspending autonomous confidence in the model and triggering structured human review in accordance with Layer 4 governance requirements.

The experiment demonstrates two points. First, statistically detectable drift can translate into clinically meaningful performance degradation. Second, embedding drift detection within a governed escalation framework ensures that detection leads to accountable action rather than silent degradation.

Audit trail implementation

Each inference executed against the deployed model generates a corresponding immutable log entry, stored within an append-only object store. The append-only constraint is deliberate: it enforces write-once semantics and prevents retrospective modification. In regulated clinical environments, reconstructability is not optional — it is foundational.

Every log record includes the following elements: the inference timestamp (UTC), the model version identifier (captured as the MLflow run ID), the SHA-256 hash of the training dataset, the input feature vector (pseudonymized through consistent hashing of the patient encounter ID), the raw prediction score alongside its calibrated probability, the conformal prediction set, and the identifier of the requesting system. Together, these fields enable the complete reconstruction of the computational context surrounding any given prediction event.

To reinforce integrity guarantees, log records are encrypted at rest using AES-256. In addition, each entry is sealed with a time-stamped digital signature, establishing non-repudiation. This means that inference records are not merely stored — they are cryptographically anchored. If questioned during regulatory inspection or clinical incident review, the authenticity of the record can be independently verified.

The audit log architecture is intentionally structured to balance traceability with GDPR data minimization requirements. Patient identifiers are decoupled from inference metadata through a dedicated pseudonymization layer. This separation enables targeted deletion of identifier mappings in response to erasure requests, without compromising the integrity of the broader audit trail. The design choice may seem subtle, but it resolves a common tension: how to preserve accountability while respecting data subject rights.

In effect, the logging system serves two simultaneous purposes. It supports regulatory scrutiny through structured queryability, and it preserves privacy guarantees through architectural compartmentalization. Accountability and minimization, rather than competing, are engineered to coexist.

Secondary validation: 30-day readmission prediction

To evaluate whether the Clinical MLOps framework's controls generalize beyond the primary prediction task and the ICU setting, we apply the same four-layer architecture to a secondary prediction problem: 30-day unplanned hospital readmission for patients discharged from general medicine wards. This task differs from the primary one in clinically meaningful ways. The patient cohort is broader and more heterogeneous (general ward rather than ICU), the prediction horizon is longer (30 days post-discharge vs. 48 hours), the outcome is less immediately life-critical, and the feature engineering pipeline draws on different clinical signals. These differences stress-test the claim that the framework is composable and portable across clinical contexts.

The secondary cohort was drawn from MIMIC-IV hospital admissions data, applying the same temporal split used in the primary task (training 2008–2016; test 2017–2019). Inclusion criteria required adult patients (18+) discharged alive from a non-ICU general medicine ward following an index admission. Readmission was defined as any unplanned return to hospitalisation within 30 days of discharge. After exclusion of patients with incomplete records and planned readmissions (flagged via ICD procedure codes for elective procedures), the final cohort comprised 31,847 index admissions, with a 30-day readmission rate of 14.7% ($n = 4,681$). The feature set was adapted from the primary pipeline: vital signs at discharge, principal diagnosis category, length of stay, number of prior admissions in the preceding 12 months, comorbidity count (Charlson Comorbidity Index), discharge disposition, and whether a follow-up appointment was scheduled. A total of 14 features were used, maintaining the data minimisation principle of Layer 1. The same XGBoost architecture and tuning procedure described in Section 4.5 were applied, and all four Clinical MLOps control layers were instantiated identically, including immutable audit logging, fairness monitoring, drift detection configuration, and human-in-the-loop escalation thresholds.

The model achieved an AUC-ROC of 0.791 (95% CI 0.778–0.804) on the temporal test set, with sensitivity 0.712, specificity 0.801, and PPV 0.431. The lower PPV relative to the primary task reflects the higher class imbalance and the inherent heterogeneity of general-ward readmission. Expected Calibration Error was 0.038, compared to 0.029 in the primary mortality model, indicating slightly wider probability miscalibration that would, in practice, require

closer monitoring in the calibration layer. Subgroup fairness analysis replicated the approach of Section 4.6: equalized odds difference was 0.051 (95% CI 0.029–0.073) across sex, and 0.068 (95% CI 0.041–0.095) across age groups, with older patients (80+) showing modestly lower sensitivity. These gaps are larger than in the primary task and would trigger Layer 2 fairness review under the ± 0.05 monitoring threshold applied to this task. The drift detection and audit trail layers required no architectural modification; they operated identically to the primary pipeline, demonstrating the composability of the framework across prediction tasks. The secondary validation, therefore, supports the claim that the Clinical MLOps architecture is not task-specific, while also illustrating that monitoring thresholds and calibration tolerances may require task-level tuning.

Discussion

Gap analysis: Standard MLOps vs. clinical MLOps

The demonstrative pipeline makes concrete the gaps between standard MLOps practice and the requirements of clinical AI deployment. Table 3 summarizes the key gaps identified and the Clinical MLOps controls that address them.

Regulatory alignment

The Clinical MLOps framework is explicitly designed to operationalize the requirements of the EU AI Act for high-risk healthcare AI systems. Article 9 (Risk Management System) is addressed by Layers 2 and 4; Article 12 (Record-keeping) by Layer 3; Article 13 (Transparency) by Layers 2 and 3; Article 14 (Human Oversight) by Layer 4; and Article 17 (Quality Management System) by the integrated governance process spanning all four layers. The framework thus provides a concrete engineering pathway for healthcare organizations seeking AI Act compliance without prescribing specific technology choices. It is important to qualify this alignment claim. The Clinical MLOps framework functions as an enabling

Table 3: Gap analysis: standard MLOps vs. Clinical MLOps framework.

Gap in Standard MLOps	Clinical Risk	Clinical MLOps Control
No demographic fairness monitoring	Undetected bias in patient subgroups	Layer 2: Disaggregated performance monitoring
Automated retraining on drift detection	Unvalidated model in production	Layer 2: Human-escalation on drift alert
No clinical outcome linkage	Model calibration is invisible to operators	Layer 2: Outcome-linked calibration monitoring
No immutable inference logging	AI Act non-compliance	Layer 3: Append-only audit trail
No uncertainty quantification at serving	Overconfident predictions are used uncritically	Layer 2: Conformal prediction intervals
Container security is not enforced.	Data exfiltration risk	Layer 1: Network isolation + vulnerability scanning
No human override tracking	AI Act Article 14 non-compliance	Layer 4: Override logging and review
Model lineage not formally recorded.	Audit trail incomplete	Layer 3: Immutable model registry with lineage

operational structure rather than a complete compliance pathway. Formal AI Act conformity for high-risk systems requires, in addition to operational controls, a conformity assessment procedure, CE marking, registration in the EU database of high-risk AI systems, and ongoing post-market surveillance documentation subject to notified body review. These regulatory procedural steps fall outside the scope of the present framework. What the framework addresses is the engineering substrate: the logging, monitoring, governance, and oversight infrastructure that makes such procedural compliance feasible and auditable. Organizations adopting the framework should treat it as a necessary but not sufficient condition for AI Act conformity. Regulatory counsel and conformity assessment expertise remain essential components of a complete compliance programme. Regarding explainability, the current demonstrative pipeline does not incorporate post-hoc explanation methods, yet explainability is both a regulatory expectation under Article 13 of the EU AI Act and a clinician-facing usability requirement. Methods such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are well-established for tree-based models and integrate naturally into the inference pipeline as a Layer 2 extension. For each prediction, SHAP feature-contribution values could be computed at serving time and surfaced to clinicians alongside the calibrated probability and uncertainty interval, providing a locally interpretable explanation without exposing raw model internals. This addition is technically straightforward within the described architecture and is recommended as a near-term enhancement to the clinical interface. Regarding phased adoption, organizations should not be expected to implement all four layers simultaneously. A practical adoption roadmap would proceed in three phases: Phase 1 (Foundations) covers Layer 1 privacy controls and Layer 3 immutable audit logging, which carry the lowest integration overhead and deliver immediate regulatory risk reduction; Phase 2 (Observability) introduces Layer 2 clinical monitoring, fairness tracking, and drift detection, which require closer integration with the clinical data warehouse; Phase 3 (Governance) operationalizes Layer 4 human-in-the-loop protocols, including override mechanisms, escalation workflows, and periodic review cycles, which require organizational process changes beyond the technical infrastructure. This sequencing aligns with realistic institutional readiness constraints and allows teams to validate each layer in isolation before introducing dependencies between them. Finally, human-in-the-loop usability must be considered explicitly. A technically correct override mechanism is not useful if clinicians find it procedurally disruptive or if alerts are generated at a frequency that induces alert fatigue. User-centred design of the Layer 4 interface, including co-design sessions with clinical end users, is a prerequisite for sustainable adoption. Measuring and reporting override rates, escalation response

times, and clinician satisfaction as operational metrics would provide feedback loops to refine threshold calibration over time.

Limitations

Several limitations of the present work merit explicit acknowledgment. No framework, however structured, is immune to contextual constraints.

First, the demonstrative pipeline is constructed using a single, extensively studied dataset — MIMIC-IV — derived from one institution in the United States. While this choice supports reproducibility and methodological transparency, it also narrows the empirical base. Healthcare systems differ substantially in electronic health record (EHR) architecture, documentation practices, patient demographics, and regulatory enforcement cultures. Generalizing these findings to other environments — particularly lower-resource settings with fragmented infrastructure — requires additional validation. The framework is designed to be portable; whether it performs equivalently across contexts remains an empirical question. The secondary validation in Section 4.9 partially addresses this concern by demonstrating framework applicability to a distinct cohort (general medicine ward, $n = 31,847$) and prediction task (30-day readmission), but both tasks remain within a single institution. Multi-institutional validation remains a priority for future work.

Second, the fairness analysis is necessarily constrained by the demographic attributes available within MIMIC-IV. Variables such as age, sex, and ethnicity are included, but broader socioeconomic indicators and social determinants of health are not systematically captured. This limitation restricts the depth of equity analysis. Bias detection, in this configuration, reflects what is measurable rather than what is socially comprehensive. Future implementations would benefit from integrating richer contextual data, provided governance safeguards are maintained.

Third, the proposed escalation thresholds — including a 5% AUC degradation trigger and a ± 0.03 AUC fairness gap threshold — are informed by prior literature and operational judgment. However, these thresholds should not be treated as universally prescriptive. Clinical risk tolerance varies by application domain, and stakeholder consultation is essential before translating such parameters into binding deployment policies. Technical plausibility must be aligned with clinical consensus.

Fourth, the framework concentrates on technical and process-level governance mechanisms. Organizational and cultural dynamics — clinician trust in AI systems, institutional readiness for AI integration, and workforce training requirements — fall outside its immediate scope. Yet

these human factors exert significant influence on real-world performance. A technically robust Clinical MLOps pipeline may still falter if adoption is hesitant or if oversight roles are ambiguously defined. Implementation, therefore, demands attention not only to infrastructure but also to institutional culture.

These limitations do not invalidate the framework; rather, they delineate its current boundaries. Addressing them will require interdisciplinary collaboration, iterative validation, and continued engagement with both clinical practitioners and regulatory bodies.

Future work

Several avenues for future investigation follow naturally from the present work. The framework, as articulated, is structurally coherent and technically implementable — yet its real test lies beyond controlled demonstration environments.

First, prospective validation in live clinical settings is essential. Controlled experiments offer clarity; operational hospitals introduce friction. Measuring the compliance overhead associated with Clinical MLOps — and its impact on model development velocity — would provide valuable insight into the trade-offs between governance rigor and innovation speed. Does stronger oversight meaningfully slow iteration cycles? Or does early governance integration reduce downstream remediation costs? Only deployment-based studies can resolve this tension.

Second, the fairness monitoring layer could be strengthened through the incorporation of causal fairness methodologies [18]. Current subgroup performance comparisons detect disparities, but they do not necessarily distinguish between spurious correlations and clinically justified risk differentials. Causal frameworks offer a more nuanced analytical lens, enabling differentiation between structural bias and legitimate predictive signal. Integrating such methods would increase analytical depth, though it would also raise methodological complexity.

Third, extending the framework to federated learning architectures [19] would broaden its applicability to multi-institutional contexts. Federated approaches allow collaborative model training across hospital networks without centralizing patient-level data. This architecture aligns naturally with privacy-preserving principles, yet introduces additional governance questions: how are updates validated, how is drift monitored across sites, and how is cross-institutional accountability coordinated? Embedding Clinical MLOps controls within federated infrastructures represents a promising, if technically demanding, direction.

Finally, the relationship between Clinical MLOps maturity and measurable clinical outcome improvement warrants

systematic evaluation. The assumption underlying this framework is that stronger operational governance enhances patient safety and system reliability. That assumption is plausible — but it remains empirical. Longitudinal, outcome-linked deployment studies are necessary to determine whether enhanced lifecycle controls translate into tangible improvements in patient outcomes.

In short, the framework establishes a structured foundation. Its broader value will depend on iterative validation, cross-institutional adaptation, and sustained empirical inquiry into the link between operational rigor and clinical impact.

Conclusion

This paper has argued that deploying AI systems in clinical environments demands more than a straightforward application of conventional MLOps practices. It requires a specialized operational structure — one that integrates domain-specific controls for privacy protection, clinically meaningful observability, regulatory traceability, and structured human oversight. The proposed Clinical MLOps framework organizes these controls into four interdependent layers, each designed to address a distinct category of requirements emerging at the intersection of healthcare delivery and regulatory accountability.

The demonstrative pipeline built on the MIMIC-IV dataset provides a concrete instantiation of these principles. Rather than remaining conceptual, the framework is exercised end-to-end, exposing specific shortcomings in standard MLOps configurations when placed under clinical and regulatory constraints. Importantly, these shortcomings are not resolved through experimental tooling or speculative architectures. They can be addressed using established, production-grade technologies — provided those technologies are arranged and governed according to Clinical MLOps principles. The distinction is architectural rather than technological.

It is worth emphasizing that the framework does not function as a brake on innovation. On the contrary, it establishes the operational scaffolding necessary for innovation to persist responsibly over time. Without structured governance, clinical AI systems risk degradation, inequity, or non-compliance. With structured governance, iterative improvement becomes sustainable rather than precarious.

As healthcare AI transitions from research experimentation to embedded clinical infrastructure, the need for implementation-oriented frameworks becomes increasingly acute. Ethical guidelines and regulatory mandates articulate important expectations, but they do not, by themselves, specify engineering pathways. Clinical MLOps seeks to bridge that divide — translating normative requirements into operational design patterns that can be executed, audited, and continuously improved.

References

1. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*. 2015;28.
2. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*. 2023;11:31866-31879.
3. European Parliament. Artificial Intelligence Act. Regulation (EU) 2024/1689. 2024.
4. Paleyes A, Urma RG, Lawrence ND. Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*. 2022;55(6):1-29.
5. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021 Jul 15;385(3):283-286. doi: 10.1056/NEJMc2104626. PMID: 34260843; PMCID: PMC8665481.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.
7. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, Heller K, Kale D, Saeed M, Ossorio PN, Thadaneys-Israeli S, Goldenberg A. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019 Sep;25(9):1337-1340. doi: 10.1038/s41591-019-0548-6. Epub 2019 Aug 19. Erratum in: *Nat Med*. 2019 Oct;25(10):1627. doi: 10.1038/s41591-019-0609-x. PMID: 31427808.
8. European Parliament. General Data Protection Regulation (GDPR). Regulation (EU) 2016/679. 2016.
9. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med*. 2018 Mar 15;378(11):981-983. doi: 10.1056/NEJMp1714229. PMID: 29539284; PMCID: PMC5962261.
10. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019 Jan;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.
11. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023 Jan 3;10(1):1. doi: 10.1038/s41597-022-01899-x. Erratum in: *Sci Data*. 2023 Jan 16;10(1):31. doi: 10.1038/s41597-023-01945-2. Erratum in: *Sci Data*. 2023 Apr 18;10(1):219. doi: 10.1038/s41597-023-02136-9. PMID: 36596836; PMCID: PMC9810617.
12. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Sci Data*. 2019 Jun 17;6(1):96. doi: 10.1038/s41597-019-0103-9. PMID: 31209213; PMCID: PMC6572845.
13. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*. 1999;10(3):61-74.
14. Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv [Preprint]*. 2022. Available from: <https://arxiv.org/abs/2107.07511>
15. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*. 2016;29.
16. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, Cereda D, Coluccello A, Foti G, Fumagalli R, Iotti G, Latronico N, Lorini L, Merler S, Natalini G, Piatti A, Ranieri MV, Scandroglio AM, Storti E, Cecconi M, Pesenti A; COVID-19 Lombardy ICU Network. Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA*. 2020 Apr 28;323(16):1574-1581. doi: 10.1001/jama.2020.5394. Erratum in: *JAMA*. 2021 May 25;325(20):2120. doi: 10.1001/jama.2021.5060. PMID: 32250385; PMCID: PMC7136855..
17. Bifet A, Gavalda R. Learning from time-changing data with adaptive windowing. In: *Proceedings of the SIAM International Conference on Data Mining*; 2007;443-448.
18. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. *Advances in Neural Information Processing Systems*. 2017;30.
19. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*; 2017;1273-1282.

How to cite this article: Spadacini D. Clinical MLOps: A Framework for Responsible Deployment and Observability of AI Systems in Cloud-Native Healthcare. *IgMin Res*. March 18, 2026; 4(3): 081-093. IgMin ID: igmin336; DOI: 10.61927/igmin336; Available at: igmin.link/p336