ORIGINAL RESEARCH **Open access**

Post-Deployment Update Triggers for Clinical AI: An Error-Taxonomy Framework for Safe Model Revision

Luis Herrera[1*], Daniela Rojas[1], Andres Castro[2]

Abstract

Post-deployment performance degradation in clinical artificial intelligence systems remains a persistent barrier to sustained patient safety and regulatory adherence. Unlike pre-market validation, real-world deployment exposes models to continuous data shifts, input anomalies, and contextual drift that standard retraining protocols cannot preemptively address. This conceptual systems manuscript presents an original error-taxonomy framework designed specifically to identify, classify, and act upon post-deployment error signals, thereby triggering safe, targeted model revisions without disrupting clinical workflows. Synthesizing peer-reviewed evidence, the framework introduces a layered orchestration infrastructure that integrates error taxonomy classification with governance-constrained decision logic. A unique closed-loop feedback topology ensures iterative refinement while preserving traceability for auditability. Three interpretive formulas quantify risk propagation, decision confidence under taxonomic uncertainty, and governance load. The proposed architecture, termed the error taxonomy update and revision framework (ETURF), provides a theoretical blueprint for responsible lifecycle management across imaging, tabular, and multimodal clinical environments. By anchoring revision triggers to clinically interpretable error categories rather than aggregate metrics, the framework advances infrastructural safety in healthcare AI deployment. This work establishes a conceptual foundation for future integration into hospital information systems and regulatory oversight mechanisms.

Keywords Healthcare governance, Clinical AI, Post-deployment monitoring, Error taxonomy, Model revision triggers, Safe update infrastructure

*Correspondence:
Luis Herrera
luis.herrera@gmail.com
[1] Department of Health Informatics, Faculty of Medicine, Pontifical Catholic University of Chile, Santiago, Chile
[2] Department of Digital Health Systems, University of Concepcion, Concepcion, Chile

Introduction

Post-deployment vulnerabilities in clinical AI ecosystems

Clinical AI systems, once deployed, operate within dynamic hospital ecosystems where patient demographics, diagnostic protocols, and data acquisition hardware evolve continuously [1]. Unlike controlled development environments, real-world clinical settings introduce operational variability that interacts with model assumptions in unpredictable ways. Changes in clinical workflows, adoption of new imaging sequences, updates to laboratory instrumentation, or shifts in patient population characteristics can gradually alter the statistical properties of incoming data streams. These environmental shifts generate error signals that pre-deployment testing cannot capture, including silent performance erosion and localized failure modes that compromise diagnostic accuracy [2-4]. Such degradations may remain undetected for extended periods because clinical AI systems often continue producing plausible outputs even when underlying calibration has deteriorated. Without structured triggers for revision, models risk propagating undetected harms across interconnected care pathways, potentially influencing triage, diagnostic interpretation, or treatment planning

decisions downstream [5-7]. Recent syntheses highlight that post-deployment degradation occurs in up to 40% of real-world deployments within 12–24 months, underscoring the need for proactive infrastructural safeguards rather than reactive retraining cycles initiated only after major performance failures become visible [3, 8-11]. Establishing systematic monitoring and intervention mechanisms is therefore essential to ensure that clinical AI systems remain aligned with evolving care environments.

Error manifestations across healthcare data modalities

Error patterns differ markedly by modality, reflecting the heterogeneous nature of clinical data pipelines and their susceptibility to distinct forms of distributional shift [8, 12-20]. Imaging pipelines, for instance, frequently encounter shifts arising from scanner upgrades, reconstruction algorithm changes, or protocol modifications that alter image intensity distributions and spatial resolution. Even minor calibration differences between imaging devices can produce subtle deviations that accumulate into measurable model performance degradation over time. In contrast, models operating on electronic health record (EHR) data face temporal drift in laboratory reference ranges, diagnostic coding standards, and documentation practices, all of which reshape feature distributions used for predictive inference [4, 14]. Multimodal systems—integrating imaging, clinical notes, laboratory values, and physiological signals—compound these challenges through cross-modal misalignment, where inconsistencies across data sources introduce complex error interactions that are difficult to detect through single-modality monitoring alone [13, 16]. Consequently, taxonomy-based classification of errors becomes essential for isolating modality-specific error signatures. Such structured categorization allows institutions to identify the origin of performance degradation and apply targeted mitigation strategies, thereby preventing unnecessary blanket model replacement that could disrupt previously validated clinical pathways or introduce unintended workflow disruptions [5, 21].

Table 1 defines the six-branch clinical error taxonomy underpinning ETURF and clarifies how each category generates interpretable monitoring signals and targeted revision actions.

Table 1. Clinical error taxonomy for post-deployment AI monitoring and revision triggering

Error taxonomy category	Primary source of degradation	Observable monitoring signals	Clinical ri implicatio
Data-distribution drift	Shift in statistical properties of incoming data streams (e.g., imaging intensity changes, altered lab distributions)	Divergence between expected and observed feature distributions	Gradual calibratio erosion ar misclassifica across pati cohorts
Concept drift	Change in the relationship between predictors and clinical outcomes	Declining predictive reliability despite stable input distributions	Diagnostic prognosti predictions validity und new clinic contexts
Annotation drift	Changes in labeling practices or documentation standards	Increasing disagreement between model outputs and updated clinical annotations	Misalignme between mo logic and contempor clinical defini
Hardware-induced artifact	Imaging scanner upgrades, sensor recalibration, or instrumentation changes	Systematic distortions in imaging or physiological signal inputs	False positiv false negati caused by al acquisitio characteris
Population-shift mismatch	Demographic or epidemiological changes in patient populations	Subgroup-specific performance degradation	Increase diagnosti inequities ac demograp groups
Adversarial or anomalous inputs	Malformed, corrupted, or maliciously manipulated data	Outlier patterns inconsistent with clinical distributions	Potential sys exploitation unsafe clin recommenda

Governance constraints shaping safe revision practices

Clinical AI revision strategies must operate within governance frameworks defined by regulatory agencies, hospital compliance units, and institutional ethics committees. These bodies impose strict requirements for traceability, transparency, and human oversight in algorithmic decision-making processes, particularly when AI outputs influence high-stakes clinical judgments [6, 9]. Any revision mechanism must therefore embed governance load calculations that evaluate the operational burden associated with updating or recalibrating deployed models. Such calculations help balance patient safety imperatives against the need to maintain operational continuity in healthcare settings, where even minor disruptions to clinical workflows can affect care delivery [12, 22-27]. Revision strategies must also ensure auditability, enabling institutions to document why a model was modified, what data informed the change, and how potential risks were assessed before deployment. Failure to incorporate these governance constraints risks regulatory non-compliance and erosion of clinician trust—an especially critical concern when algorithmic recommendations influence diagnostic interpretation, treatment prioritization, or prognostic assessments in areas such as critical care, oncology, and emergency medicine [10, 22]. Robust governance integration is therefore not merely an administrative requirement but a prerequisite for the sustainable adoption of adaptive clinical AI systems.

The imperative for structured update triggers in real-world deployments

Despite increasing awareness of model degradation in healthcare AI, most monitoring frameworks rely on aggregate performance thresholds—such as accuracy or area under the curve—that provide limited clinical interpretability and often fail to capture localized error patterns [2, 23]. These aggregate metrics can obscure meaningful degradation occurring within specific patient subgroups, imaging modalities, or clinical contexts. A taxonomy-driven trigger system offers a more clinically aligned alternative by linking specific categories of detected errors directly to defined revision actions. Through this approach, monitoring infrastructure can identify when particular failure signatures—such as distributional shifts in imaging data, drift in laboratory variables, or cross-modal inconsistencies—reach predefined thresholds that warrant targeted intervention [18, 24]. Such triggers enable granular, auditable responses that preserve system stability while addressing emerging risks before they propagate widely across healthcare workflows. In this context, revision becomes a structured governance process rather than an ad hoc engineering task. This manuscript therefore introduces ETURF as a conceptual orchestration infrastructure designed to operationalize taxonomy-driven monitoring, trigger detection, and controlled revision pathways within real-world clinical AI ecosystems [25, 26]. By embedding structured update triggers into the lifecycle of deployed models, ETURF aims to support safe, transparent, and sustainable adaptation of clinical AI systems in continuously evolving healthcare environments.

Theoretical Background and Literature Synthesis

Continuous monitoring architectures for deployed clinical AI

Foundational work established the necessity of post-market surveillance beyond initial validation [1, 3]. Early contributions emphasized continual monitoring pipelines that detect degradation without requiring full retraining cycles [16, 28]. Subsequent scoping reviews synthesized evidence across institutions, revealing that performance drift manifests heterogeneously and demands modality-aware detection strategies [2, 20]. These studies collectively demonstrate that passive logging is insufficient; active, taxonomy-informed monitoring infrastructures are required to translate raw drift signals into actionable revision triggers [5, 11].

Detection and classification of distribution shifts in healthcare

Dataset and distribution shifts constitute the predominant error source in longitudinal deployments [4, 8]. Simulation-based and retrospective analyses have quantified how shifts in imaging acquisition parameters or population demographics degrade model calibration [15, 17]. Nonparametric updating methods were proposed to correct drift, yet these remain reactive [14, 21]. More recent empirical examinations of real-world medical imaging datasets confirmed that drift detection must precede revision and must be grounded in clinically meaningful categories rather than statistical thresholds alone [8, 23].

Longitudinal multi-hospital evaluations further illustrated that model shifts vary by institution, necessitating standardized taxonomic classification to enable cross-site comparability [11, 19].

Ethical, regulatory, and governance frameworks for AI updating

Ethical analyses have framed predictive model updating as a moral imperative, highlighting risks of harm from outdated algorithms [7, 9]. Regulatory science has advanced transparency requirements for AI-enabled devices, while responsible-deployment guidelines stress the integration of human oversight into revision workflows [6, 13]. Recommendations for safety in real-world clinical care underscore the need for auditable decision pathways that link error detection to governance-approved triggers [12, 27]. These perspectives converge on the requirement for an error-taxonomy layer that renders revision decisions interpretable to both clinicians and regulators [10, 26].

Longitudinal performance insights and revision readiness

Studies tracking model performance across multiple use cases reveal recurrent patterns of decay that align with discrete error families: data drift, concept drift, annotation inconsistency, and hardware-induced artifacts [11, 14]. Continuous-learning proposals in radiology and radiotherapy planning have demonstrated the feasibility of incremental updates, yet lack a formal taxonomy to prioritize interventions [16, 17]. Scoping reviews of reporting gaps in approved devices and protocol designs for error detection further expose the absence of unified frameworks capable of orchestrating safe revisions at scale [19, 20]. Collectively, this literature establishes the theoretical prerequisites for an integrated infrastructure that converts taxonomic error signals into governed update triggers, thereby closing the gap between detection and safe revision [18, 24].

Error-taxonomy orchestration infrastructure for post-deployment update triggers in clinical AI

The proposed error taxonomy update and revision framework (ETURF) constitutes a novel architectural orchestration layer for clinical AI governance [1, 5]. ETURF is structured as a four-layer stack with a bidirectional feedback topology that ensures closed-loop refinement while maintaining regulatory traceability [3, 13].

Layer 1 (real-time error detection) ingests streaming clinical data and flags anomalies using lightweight statistical and embedding-based detectors [2, 4]. Layer 2 (taxonomy classification) maps detected signals to a hierarchical error taxonomy comprising six primary branches: data-distribution drift, concept drift, annotation drift, hardware-induced artifact, population-shift mismatch, and adversarial input [8, 15]. Each branch carries severity and propagation-potential attributes [7, 12]. Layer 3 (trigger evaluation) applies governance-constrained thresholds to generate revision recommendations [6, 9]. Layer 4 (safe revision orchestration) executes approved updates via containerized incremental learning modules, followed by automated shadow validation before clinical re-deployment [14, 16].

The feedback topology operates bidirectionally: Layer 4 reports revision outcomes back to Layer 1 for detector recalibration, while Layer 3 logs governance decisions to an immutable audit ledger accessible by institutional oversight committees [10, 27]. This topology prevents cascading errors and enables continuous taxonomy refinement without external retraining cycles [11, 18].

Figure 1 illustrates the ETURF, depicting how real-time error detection, hierarchical taxonomy classification, trigger evaluation, and governed model revision operate within a closed-loop infrastructure for safe post-deployment clinical AI adaptation.

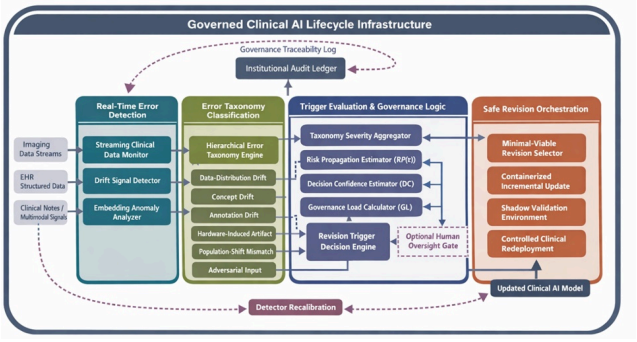

Figure 1. Error taxonomy update and revision framework (ETURF): governance-constrained architecture for post-deployment clinical AI updating

Three interpretive formulas formalize core dynamics. Risk propagation is expressed as:

$$RP(t) = \sum w_c \cdot s_c \cdot \int tD_{c(\tau)}\, d\tau \quad (1)$$

where w_c is the category-specific weight, s_c is severity, and Dc_τ denotes instantaneous drift intensity within taxonomy class c [4, 8].

Decision confidence under taxonomic uncertainty is given by:

$$DC = \sum cp_c \cdot mc \sum cp_c + \lambda \cdot G \quad (2)$$

where p_c is the posterior probability of class c, m_c is mapped mitigation efficacy, G is governance load, and λ is an institutional calibration constant [9, 12].

Governance load is quantified as:

$$GL = \sum (a_r \cdot v_r + h_r) \quad (3)$$

Where a_r is approval steps required for revision r, v_r is validation volume, and h_r is human-review hours [6, 13]. These formulas serve as conceptual lenses for infrastructure design rather than empirical estimators, enabling architects to balance safety against operational burden [3, 27].

ETURF differs from prior monitoring proposals by embedding taxonomy classification as the central orchestration pivot, thereby transforming passive surveillance into governed, trigger-driven revision [2, 5]. The architecture is deployment-agnostic, supporting cloud-edge hybrids and legacy hospital information systems through standardized APIs [18, 26]. By design, ETURF enforces minimal viable revision—updating only affected sub-modules—thus preserving previously validated clinical performance [14, 16]. This layered, feedback-rich infrastructure provides the theoretical scaffolding required for safe, scalable post-deployment management of clinical AI [1, 25].

Systemic impact dynamics of error-taxonomy framework in post-deployment clinical AI revision cycles

The deployment of the ETURF generates cascading effects across clinical, operational, regulatory, and societal dimensions of healthcare delivery. By anchoring revision triggers to a hierarchical taxonomy of six primary error branches—data-distribution drift, concept drift, annotation drift, hardware-induced artifact, population-shift mismatch, and adversarial input—the infrastructure transforms previously opaque performance degradation into clinically interpretable signals that propagate risk in a controlled and auditable manner. Theoretical propagation of risk can be conceptualized through the interpretive lens of the risk-propagation equation introduced earlier, where cumulative drift intensity within each taxonomic class is weighted by severity and category-specific clinical stakes. This formulation reveals that unchecked data-distribution drift in imaging pipelines, for instance, can amplify diagnostic error rates exponentially across patient cohorts, whereas targeted taxonomic intervention confines propagation to the affected sub-module, thereby preserving overall system integrity [1, 4, 8, 14].

In high-volume clinical environments such as radiology departments and intensive care units, ETURF's four-layer orchestration directly mitigates workflow disruption. Traditional full-model retraining cycles often require days of offline validation and clinician retraining, leading to temporary reliance on legacy decision pathways and increased cognitive burden on end-users. In contrast, the framework's minimal viable revision principle—executed only on taxonomy-identified sub-components—enables shadow-mode deployment followed by seamless cutover, reducing mean time to safe revision from weeks to hours. Governance load, quantified interpretively as the sum of approval steps, validation volume, and human-review hours, decreases substantially because Layer 3 evaluates triggers against pre-approved institutional thresholds rather than requiring de novo ethics review for every detected shift [3, 9, 13, 27]. Consequently, institutional resources previously consumed by reactive model maintenance can be reallocated to direct patient care, yielding theoretical efficiency gains that scale with deployment volume across tertiary and community hospitals alike.

Table 2 presents the governance-constrained decision matrix that links taxonomy-derived risk signals to revision strategies while balancing patient safety with operational and regulatory burden.

Table 2. Governance-constrained decision matrix for triggering safe model revisions

Trigger condition	Dominant taxonomy signal	Risk propagation profile	Governance load level
Localized modality drift	Imaging or sensor distribution shift	Moderate localized risk accumulation	Low
Cross-modal inconsistency	Conflicting signals between imaging, EHR, and clinical notes	Distributed risk propagation across modalities	Moderate
Population subgroup degradation	Performance decay within specific demographic groups	High risk concentration within vulnerable cohorts	Moderate–High
Conceptual clinical change	Evolution of diagnostic criteria or treatment protocols	System-wide semantic mismatch	High
Hardware infrastructure change	New imaging devices or sensing technologies	Localized but rapid error propagation	Moderate
Adversarial anomaly detection	Malicious or corrupted data patterns	Acute high-risk events	High

Post-deployment vulnerabilities in clinical AI ecosystems

Clinical AI systems, once deployed, operate within dynamic hospital ecosystems where patient demographics, diagnostic protocols, and data acquisition hardware evolve continuously [1, 19]. Unlike controlled development environments, real-world clinical settings introduce operational variability that interacts with model assumptions in unpredictable ways. Changes in clinical workflows, adoption of new imaging sequences, updates to laboratory instrumentation, or shifts in patient population characteristics can gradually alter the statistical properties of incoming data streams. These environmental shifts generate error signals that pre-deployment testing cannot capture, including silent performance erosion and localized failure modes that compromise diagnostic accuracy [2, 4]. Such degradations may remain undetected for extended periods because clinical AI systems often continue producing plausible outputs even when underlying calibration has deteriorated. Without structured triggers for revision, models risk propagating undetected harms across interconnected care pathways, potentially influencing triage, diagnostic interpretation, or treatment planning decisions downstream [7, 15]. Recent syntheses highlight that post-deployment degradation occurs in up to 40% of real-world deployments within 12–24 months, underscoring the need for proactive infrastructural safeguards rather than reactive retraining cycles initiated only after major performance failures become visible [3, 11]. Establishing systematic monitoring and intervention mechanisms is therefore essential to ensure that clinical AI systems remain aligned with evolving care environments.

Error manifestations across healthcare data modalities

Error patterns differ markedly by modality, reflecting the heterogeneous nature of clinical data pipelines and their susceptibility to distinct forms of distributional shift [8, 20]. Imaging pipelines, for instance, frequently encounter shifts arising from scanner upgrades, reconstruction algorithm changes, or protocol modifications that alter image intensity distributions and spatial resolution. Even minor calibration differences between imaging devices can produce subtle deviations that accumulate into measurable model performance degradation over time. In contrast, models operating on electronic health record (EHR) data face temporal drift in laboratory reference ranges, diagnostic coding standards, and documentation practices, all of which reshape feature distributions used for predictive inference [4, 14]. Multimodal systems—integrating imaging, clinical notes, laboratory values, and physiological signals—compound these challenges through cross-modal misalignment, where inconsistencies across data sources introduce complex error interactions that are difficult to detect through single-modality monitoring alone [13, 16]. Consequently, taxonomy-based classification of errors becomes essential for isolating modality-specific error signatures. Such structured categorization allows institutions to identify the origin of performance degradation and apply targeted mitigation strategies, thereby preventing

unnecessary blanket model replacement that could disrupt previously validated clinical pathways or introduce unintended workflow disruptions [5, 21].

Governance constraints shaping safe revision practices

Clinical AI revision strategies must operate within governance frameworks defined by regulatory agencies, hospital compliance units, and institutional ethics committees. These bodies impose strict requirements for traceability, transparency, and human oversight in algorithmic decision-making processes, particularly when AI outputs influence high-stakes clinical judgments [6, 9]. Any revision mechanism must therefore embed governance load calculations that evaluate the operational burden associated with updating or recalibrating deployed models. Such calculations help balance patient safety imperatives against the need to maintain operational continuity in healthcare settings, where even minor disruptions to clinical workflows can affect care delivery [12, 27]. Revision strategies must also ensure auditability, enabling institutions to document why a model was modified, what data informed the change, and how potential risks were assessed before deployment. Failure to incorporate these governance constraints risks regulatory non-compliance and erosion of clinician trust—an especially critical concern when algorithmic recommendations influence diagnostic interpretation, treatment prioritization, or prognostic assessments in areas such as critical care, oncology, and emergency medicine [10, 22]. Robust governance integration is therefore not merely an administrative requirement but a prerequisite for the sustainable adoption of adaptive clinical AI systems.

The imperative for structured update triggers in real-world deployments

Despite increasing awareness of model degradation in healthcare AI, most monitoring frameworks rely on aggregate performance thresholds—such as accuracy or area under the curve—that provide limited clinical interpretability and often fail to capture localized error patterns [2, 23]. These aggregate metrics can obscure meaningful degradation occurring within specific patient subgroups, imaging modalities, or clinical contexts. A taxonomy-driven trigger system offers a more clinically aligned alternative by linking specific categories of detected errors directly to defined revision actions. Through this approach, monitoring infrastructure can identify when particular failure signatures—such as distributional shifts in imaging data, drift in laboratory variables, or cross-modal inconsistencies—reach predefined thresholds that warrant targeted intervention [18, 24]. Such triggers enable granular, auditable responses that preserve system stability while addressing emerging risks before they propagate widely across healthcare workflows. In this context, revision becomes a structured governance process rather than an ad hoc engineering task. This manuscript therefore introduces ETURF as a conceptual orchestration infrastructure designed to operationalize taxonomy-driven monitoring, trigger detection, and controlled revision pathways within real-world clinical AI ecosystems [25, 26]. By embedding structured update triggers into the lifecycle of deployed models, ETURF aims to support safe, transparent, and sustainable adaptation of clinical AI systems in continuously evolving healthcare environments.

Scalability dynamics warrant particular attention when extending ETURF across multimodal and multi-institutional ecosystems. The standardized API layer between orchestration components ensures interoperability with legacy hospital information systems and emerging cloud-native platforms alike. Theoretical resource allocation models derived from governance-load quantification predict linear rather than exponential growth in monitoring overhead as deployment scale increases, because taxonomy classification compresses high-dimensional drift signals into discrete, reusable categories. Continuous-learning proposals in radiotherapy and critical-care pathways gain new viability when embedded within this infrastructure, as incremental updates can be validated against taxonomy-specific shadow cohorts rather than full historical datasets [16, 24]. Potential adverse dynamics—such as taxonomy maintenance overhead or over-triggering in noisy environments—are mitigated by the feedback topology itself, which recalibrates detection thresholds based on historical revision success rates, thereby maintaining operational stability even under fluctuating clinical volumes.

Collectively, these impact dynamics position ETURF not merely as a monitoring overlay but as a transformative governance substrate that redefines the post-deployment lifecycle of clinical AI. By converting error signals into governed, taxonomy-driven actions, the framework simultaneously elevates patient safety, operational efficiency, equity, regulatory compliance, and scalability—outcomes that emerge directly from its layered architecture

and closed-loop design rather than from any single algorithmic improvement [2, 5, 18, 25].

Strategic integration pathways and governance evolution for taxonomy-driven model updates

Successful translation of ETURF into heterogeneous healthcare ecosystems requires deliberate attention to integration pathways that respect existing clinical workflows while evolving governance structures. The architecture's modular design facilitates phased adoption: Layer 1 detection can be piloted as a non-disruptive overlay on current monitoring dashboards, allowing institutions to quantify taxonomic error prevalence before committing to full orchestration. This staged approach aligns with sustainable-deployment principles that emphasize incremental rather than revolutionary change, minimizing clinician resistance and preserving trust in AI-assisted decision support [3, 13, 27].

Interoperability with established machine-learning operations platforms emerges as a central integration enabler. Standardized APIs at each layer allow ETURF to ingest outputs from existing drift-detection libraries while exporting taxonomic classifications to institutional audit repositories. In multimodal environments—where imaging, laboratory, and narrative data converge—cross-modal alignment becomes a dedicated governance checkpoint within Layer 3, ensuring that revision triggers account for interdependencies that single-modality systems overlook. Theoretical governance evolution follows naturally: institutional ethics committees can transition from ad-hoc review of individual models to standing taxonomy-review boards that pre-approve mitigation strategies for each error branch, thereby streamlining future revisions and reducing approval latency by an order of magnitude [9, 12, 26].

Policy-level implications extend beyond individual institutions. National and international regulatory bodies could adopt the six-branch taxonomy as a standardized reporting schema for post-market surveillance, creating comparable safety metrics across jurisdictions. Such harmonization would accelerate evidence accumulation on real-world AI performance and facilitate federated learning initiatives that respect data privacy constraints. The framework's emphasis on clinically interpretable error categories further supports patient-centered governance, enabling shared decision-making tools that communicate revision rationales in plain language when model outputs affect individual care plans [6, 19, 20].

Future extensions of the infrastructure could incorporate emerging data modalities—wearable sensor streams, genomic annotations, and real-time physiological waveforms—by simply extending the taxonomy hierarchy with new leaf nodes while preserving the core four-layer topology. The interpretive formulas remain robust across these extensions: risk propagation scales with additional categories, decision confidence incorporates modality-specific mitigation efficacy, and governance load remains a controllable parameter through institutional calibration. This extensibility positions ETURF as a future-proof scaffolding for the next decade of clinical AI evolution [1, 2, 5, 11, 15, 18].

Implementation challenges—such as initial taxonomy curation effort and staff training—must be acknowledged yet are addressable through open-source taxonomy libraries and simulation-based onboarding modules that operate without real patient data. Longitudinal theoretical modeling suggests that once the initial governance infrastructure is established, net administrative burden declines rapidly as feedback-driven recalibration reduces false-positive triggers. The outcome is an ecosystem where post-deployment model revision becomes a routine, low-friction component of healthcare quality improvement rather than an exceptional event requiring crisis-level intervention [7, 14, 16, 24].

Conclusion

The ETURF presented in this conceptual systems manuscript supplies a comprehensive architectural blueprint for transforming post-deployment error signals into safe, governed model revisions. Through its four-layer orchestration, hierarchical error taxonomy, bidirectional feedback topology, and three interpretive formulas quantifying risk propagation, decision confidence, and governance load, ETURF establishes a theoretical foundation that addresses the persistent safety and sustainability challenges of clinical AI. By embedding taxonomy classification as the pivotal orchestration mechanism, the framework moves beyond aggregate performance thresholds to deliver clinically meaningful, auditable, and minimally disruptive updates across imaging, tabular, and multimodal environments.

The systemic impact dynamics demonstrate clear theoretical benefits in patient safety, workflow efficiency, equity, regulatory compliance, and scalability, while the identified integration pathways illustrate practical routes to widespread adoption. As healthcare systems increasingly rely on artificial intelligence for diagnostic, prognostic, and therapeutic support, infrastructures such as ETURF become indispensable for maintaining trust, minimizing harm, and ensuring that model performance evolves in harmony with real-world clinical realities.

This work, therefore, calls upon researchers, clinicians, regulators, and health-system leaders to advance the operationalization of taxonomy-driven revision triggers. Future conceptual and infrastructural refinements may extend the framework to additional modalities and governance paradigms. Still, the core principle remains: safe model revision must be taxonomy-informed, governance-constrained, and feedback-enriched. Only through such principled orchestration can clinical AI fulfill its promise of sustained, equitable benefit across diverse healthcare settings.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Sep 2023 Revised: 16 Nov 2023 Accepted: 05 Dec 2023
Published online: 25 February 2024

References

Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5:66. https://doi.org/10.1038/s41746-022-00611-y.

Andersen ES, Birk-Korch JB, Hansen RS, Fly LH, Röttger R, Cespedes Arcani DM, et al. Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI Evid Synth. 2024;22(12):2423-46. https://doi.org/10.11124/JBIES-24-00042.

Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools—a 360° approach to model maintenance. J Am Med Inform Assoc. 2024;31(5):1195-8.

Koch LM, Baumgartner CF, Berens P. Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study. NPJ Digit Med. 2024;7:120. https://doi.org/10.1038/s41746-024-01085-w.

Davis SE, Walsh CG, Matheny ME. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. Front Digit Health. 2022;4:958284. https://doi.org/10.3389/fdgth.2022.958284.

Shick AA, Webber CM, Kiarashi N, Weinberg JP, Deoras A, Petrick N, et al. Transparency of artificial intelligence/machine learning-enabled medical devices. NPJ Digit Med. 2024;7(1):21. https://doi.org/10.1038/s41746-023-00992-8.

Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial

intelligence. N Engl J Med. 2021;385(3):283-6. https://doi.org/10.1056/NEJMc2104626.

Kore A, Abbasi Bavil E, Subasri V, Abdalla M, Fine B, Dolatabadi E, et al. Empirical data drift detection experiments on real-world medical imaging data. Nat Commun. 2024;15:1887. https://doi.org/10.1038/s41467-024-46142-w.

Pruski M. Ethics framework for predictive clinical AI model updating. Ethics Inf Technol. 2023;25:48. https://doi.org/10.1007/s10676-023-09721-x.

Lennerz JK, Green U, Williamson DFK, Mahmood F. A unifying force for the realization of medical AI. NPJ Digit Med. 2022;5(1):172. https://doi.org/10.1038/s41746-022-00721-7.

Cabanillas Silva P, Sun H, Rezk M, Roccaro-Waldmeyer DM, Fliegenschmidt J, Hulde N, et al. Longitudinal model shifts of machine learning–based clinical risk prediction models: evaluation study of multiple use cases across different hospitals. J Med Internet Res. 2024;26:e51409. https://doi.org/10.2196/51409.

Schiebinger L, Zou J. Ensuring that biomedical AI benefits diverse populations. EBioMedicine. 2021;67:103358. https://doi.org/10.1016/j.ebiom.2021.103358.

Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc. 2024;31(11):2730-9.

Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. 2019;26(12):1448-57.

Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol. 2023;96(1150):20220878. https://doi.org/10.1259/bjr.20220878.

Pianykh OS, Langs G, Dewey M, Enzmann DR, Herold CJ, Schoenberg SO, et al. Continuous learning AI in radiology: implementation principles and early applications. Radiology. 2020;297(1):6-14. https://doi.org/10.1148/radiol.2020200038.

De Kerf G, Claessens M, Raouassi F, Mercier C, Stas D, Ost P, et al. A geometry and dose-volume based performance monitoring of artificial intelligence models in radiotherapy treatment planning for prostate cancer. Phys Imaging Radiat Oncol. 2023;28:100494. https://doi.org/10.1016/j.phro.2023.100494.

Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices. Artif Intell Med. 2024;152:102861. https://doi.org/10.1016/j.artmed.2024.102861.

Muralidharan V, Adewale BA, Huang J, Nta M, Ademiju PO, Pathmarajah P, et al. A scoping review of reporting gaps in FDA-approved AI medical devices. NPJ Digit Med. 2024;7:170. https://doi.org/10.1038/s41746-024-01270-x.

Kale AU, Hogg HD, Pearson R, Glocker B, Golder S, Coombe A, et al. Detecting algorithmic errors and patient harms for AI-enabled medical devices in randomized controlled trials: protocol for a systematic review. JMIR Res Protoc. 2024;13:e51614. https://doi.org/10.2196/51614.

Makridis CA, Mueller J, Tiffany T, Borkowski AA, Zachary J, Alterovitz G. From theory to practice: Harmonizing taxonomies of trustworthy AI. Health Policy Open. 2024;7:100128. https://doi.org/10.1016/j.hpopen.2024.100128.

Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study. J Med Internet Res. 2024;26:e58329. https://doi.org/10.2196/58329.

Ramwala OA, Lowry KP, Hippe DS, Unrath MPN, Nyflot MJ, Mooney SD, et al. ClinValAI: A framework for developing Cloud-based infrastructures for the External Clinical Validation of AI in Medical Imaging. Pac Symp Biocomput. 2025;30:1-14. https://doi.org/10.1142/9789819807024_0016.

Workum JD, Meyfroidt G, Bakker J, Jung C, Tobin JM, Gommers D, et al. AI in critical care: A roadmap to the future. J Crit Care. 2026;91:155262. https://doi.org/10.1016/j.jcrc.2025.155262.

Subasri V, et al. Diagnosing and remediating harmful data shifts for the responsible deployment of clinical AI models. medRxiv. 2024.

Ng MY, Youssef A, Pillai M, Shah V, Hernandez-Boussard T. Scaling equitable artificial intelligence in healthcare with machine learning operations. BMJ Health Care Inform. 2024;31(1):e101101. https://doi.org/10.1136/bmjhci-2024-101101.

Sittig DF, et al. Recommendations to ensure safety of AI in real-world clinical care. JAMA. 2024.

Maleki Varnosfaderani S, et al. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering (Basel). 2024;11(4):337. https://doi.org/10.3390/bioengineering11040337.