Mind the Gap: An Analytical Framework for Evaluation–Regulation Alignment in Medical Video AI

Anonymous ISBI 2026 Submission
Paper ID: [To be assigned]

***Abstract*—Artificial intelligence systems for real-time medical video analysis—particularly computer-aided detection (CADe) in colonoscopy—have achieved regulatory clearance and clinical deployment. However, the evaluation metrics used for regulatory submission were inherited from static-image diagnostic paradigms and fail to capture video-specific operational properties such as detection latency, temporal false-positive burden, and prediction consistency across frames. This paper formalizes a three-way misalignment between the metrics AI developers report, the metrics regulators require for device clearance, and the metrics clinicians need for safe deployment. We propose the Evaluation–Regulation Alignment (ERA) framework, which decomposes clinical video AI tasks into atomic sub-tasks, maps each to required evaluation dimensions, and assesses regulatory coverage to expose systematic gaps. Applying ERA to colonoscopy CADe, we demonstrate that no FDA-cleared device has been evaluated on temporal detection latency (time from lesion appearance to first alert), alert burden dynamics, or prediction consistency—despite their direct impact on clinical safety. We conclude with actionable metric recommendations and a phased regulatory integration pathway.**

***Index Terms*—Medical video AI, uncertainty quantification, evaluation framework, regulatory science, computer-aided detection, colonoscopy**

I. Introduction

Medical video AI has progressed from research prototypes to regulatory clearance and clinical deployment with remarkable speed. In colonoscopy alone, four computer-aided detection (CADe) systems have received U.S. Food and Drug Administration (FDA) authorization: GI Genius (Medtronic) via the De Novo pathway in 2021 [1], EndoScreener (Wision AI) and SKOUT (Iterative Health) via 510(k) clearance in 2022, and CADDIE (Odin Medical/Olympus) via 510(k) in 2024 [2], [3]. Meta-analytic evidence from randomized controlled trials confirms that CADe significantly increases adenoma detection rates (ADR) during screening colonoscopy [4].

Yet the evaluation paradigm underlying these clearances was designed for a different era. The primary metrics—per-frame sensitivity, per-procedure ADR, and false-positive rates—were developed for static diagnostic imaging or procedure-level clinical endpoints [5], [6], treating each video frame as independent and each procedure as a binary outcome. This framing systematically ignores the temporal dimension that defines video AI: how predictions evolve across frames, how quickly the system responds to a lesion entering the visual field, and how the cumulative burden of transient false alerts affects clinical decision-making.

This paper makes three contributions. First, we present a taxonomy of video-specific evaluation dimensions that current metrics fail to capture, several of which represent unmeasured sources of temporal uncertainty (Section II). Second, we propose the *Evaluation–Regulation Alignment* (ERA) framework, a structured analytical tool that maps clinical sub-tasks to required evaluation dimensions and exposes systematic gaps in regulatory coverage (Section III). Third, we provide concrete, actionable metric recommendations—including metrics that bridge evaluation and uncertainty quantification—with a phased regulatory integration pathway (Section IV).

We focus on colonoscopy CADe as the most mature clinical use case, with four FDA-cleared devices and substantial trial evidence. However, the ERA framework is designed to generalize to other medical video AI domains including capsule endoscopy, surgical video analysis, and real-time ultrasound assessment.

II. The Evaluation–Regulation Gap

A. Current Evaluation Practice

The evaluation of colonoscopy CADe systems relies on a small set of established metrics. At the *per-frame level*, standalone performance testing measures sensitivity (proportion of frames containing lesions that are correctly flagged) and false-positive rate (erroneous alerts per frame or per unit time). At the *per-procedure level*, clinical trials measure ADR (proportion of procedures in which at least one adenoma is detected), adenomas per colonoscopy (APC), and adenoma miss rate (AMR) [1], [4]. These metrics have been foundational for establishing clinical utility [14]. The landmark GI Genius trial demonstrated a 30% relative increase in ADR and a 46% relative increase in APC [1]. Subsequent meta-analysis of twelve RCTs involving over 11,000 patients confirmed a consistent benefit across systems and populations [4]. The recent EAGLE trial extended this evidence to cloud-based architectures [2].

However, these metrics share a critical limitation: they are *temporally agnostic*. Per-frame sensitivity treats each frame as independent, discarding the sequential structure of video. Per-procedure ADR captures only the final clinical outcome, not the operational behavior of the AI system during the procedure. Between the frame and the procedure, an entire stratum of video-specific operational properties goes unmeasured.

B. Unmeasured Video-Specific Dimensions

Table I presents a taxonomy of six evaluation dimensions specific to medical video AI that are absent from current regulatory evaluation. These dimensions directly impact clinical

TABLE I
TAXONOMY OF UNMEASURED VIDEO-SPECIFIC EVALUATION DIMENSIONS IN MEDICAL VIDEO AI

Dimension	Definition	Clinical Impact	Eval'd?
Detection latency	Frames from lesion entering FOV to first alert	Missed lesions during rapid scope movement	No
Temporal FP burden	Rate and clustering of transient false alerts	Alert fatigue, cognitive overload	No
Prediction consistency	Stability of outputs across consecutive frames	Clinician trust and distraction	No
Lesion persistence	Maintaining identity of same lesion across frames	Double-counting, inaccurate APC	No
Recovery time	Frames to re-detect after brief occlusion	Robustness to real-world scope motion	No
Inference latency	End-to-end processing delay per frame	Real-time usability	Partial

safety and efficacy but have no standardized measurement protocol.

Detection latency measures the temporal delay between a lesion becoming visible and the system producing an alert. A system with high per-frame sensitivity may still exhibit clinically significant detection latency if it requires multiple frames of stable lesion presence before triggering. During rapid scope withdrawal—common in clinical practice—this delay translates directly into missed lesions that technically appeared in the visual field.

Temporal false-positive burden extends the static false-positive rate into the time domain. A system producing one brief false alert per minute has a fundamentally different clinical impact than one producing a sustained five-second false alert every ten minutes, even if both report identical per-frame false-positive rates. The *clustering* and *duration* of false positives determine their cognitive cost [8].

Prediction consistency captures whether the system's output remains stable when viewing the same region across frames. Flickering predictions—alternating rapidly between positive and negative on the same mucosal region—erode clinician trust and create distraction during the procedure, even if the aggregate sensitivity is high.

Lesion persistence tracking refers to the system's ability to recognize that successive detections across frames correspond to the *same* anatomical lesion. Without persistence tracking, a single polyp viewed across thirty frames could be counted as thirty separate detection events, inflating per-frame sensitivity while providing no information about the system's clinical utility.

Recovery time measures how quickly the system re-detects a lesion after temporary occlusion—by luminal fluid, scope rotation, or instrument passage. This is a common clinical scenario that existing metrics ignore entirely.

Rex et al. [7] have independently noted that current CADe systems exhibit a persistent weakness in detecting large flat lesions, suggesting that aggregate sensitivity metrics mask clinically significant detection failures that temporal analysis could reveal.

C. The Three-Way Misalignment

The gap is not merely between "what exists" and "what is ideal." It is a *structural* three-way misalignment among three stakeholder metric sets. Let $\mathcal{M}_{\text{dev}}$ denote the set of metrics AI developers typically report, $\mathcal{M}_{\text{reg}}$ the metrics regulators require for clearance, and $\mathcal{M}_{\text{clin}}$ the metrics clinicians need for safe deployment.

In current practice: $\mathcal{M}_{\text{dev}} \supseteq \mathcal{M}_{\text{reg}}$ (developers measure at least what is required), and this condition is typically satisfied. However, $\mathcal{M}_{\text{reg}} \subset \mathcal{M}_{\text{clin}}$ (regulators require less than clinicians need), and critically, $\mathcal{M}_{\text{dev}} \cap \mathcal{M}_{\text{video}} \approx \varnothing$ (developers rarely measure video-specific properties), where $\mathcal{M}_{\text{video}}$ denotes the six dimensions in Table I. Since regulators do not require these metrics, developers have no incentive to measure them; since developers do not measure them, regulators have no evidence base to incorporate them. This creates a self-reinforcing gap.

A recent cross-sectional analysis of 903 FDA-cleared AI devices found that clinical performance studies were reported for only approximately half, and study designs were frequently unspecified [9]. The FDA's evolving guidance on AI lifecycle management [11], [12] acknowledges the need for continuous monitoring but does not specify temporal evaluation dimensions for video-based systems. The EU Medical Device Regulation and AI Act similarly lack video-specific requirements [13].

III. THE ERA FRAMEWORK

A. Framework Overview

We propose the Evaluation–Regulation Alignment (ERA) framework as a structured analytical tool for identifying, categorizing, and prioritizing evaluation gaps in medical video AI. The framework operates across three layers, illustrated in Fig. 1.

Layer 1: Clinical Task Decomposition. The target video AI system is decomposed into atomic clinical sub-tasks. For colonoscopy CADe, these are: (1) real-time detection of candidate lesions, (2) spatial localization via bounding box overlay, (3) temporal tracking of detected regions across frames, (4) aggregation of frame-level detections into lesion-level events, and (5) clinical decision support at the procedure level.

Layer 2: Metric–Task Mapping. For each sub-task, the framework identifies the evaluation dimensions required for comprehensive assessment. This mapping draws from the taxonomy in Table I and from existing clinical quality indicators [6]. Each sub-task may require multiple dimensions; each dimension may be relevant to multiple sub-tasks.

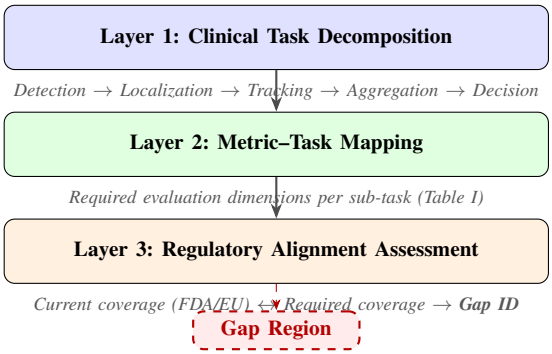

Fig. 1. The Evaluation–Regulation Alignment (ERA) framework. Each clinical sub-task is mapped to required evaluation dimensions, which are then assessed against current regulatory requirements to identify gaps.

TABLE II
ERA MAPPING FOR COLONOSCOPY CADE SYSTEMS

Sub-task	Required Metric	Existing Proxy	Reg. Status	Gap
Detection	Detection latency (ms)	None	Not req'd	High
Local'n	Spatial accuracy (IoU)	Bounding box	Partial	Low
Tracking	Temporal consistency	None	Not req'd	High
Tracking	Lesion persistence	None	Not req'd	High
Aggreg'n	Alert burden (alerts/min)	Per-frame FP	Static only	Med
Decision	Procedure-level confidence	ADR, APC	Covered	Low

Layer 3: Regulatory Alignment Assessment. Each required evaluation dimension is mapped against current regulatory requirements across jurisdictions (FDA, EU MDR/AI Act). Dimensions are classified as: *covered* (required and routinely reported), *partially covered* (acknowledged but not standardized), or *uncovered* (absent from regulatory requirements). Uncovered dimensions with high clinical impact constitute the gap region.

B. Application to Colonoscopy CADe

Table II presents the ERA mapping for colonoscopy CADe, applying the framework to the five clinical sub-tasks identified in Layer 1.

The mapping reveals a clear pattern: sub-tasks at the temporal boundaries of the pipeline—detection onset, cross-frame tracking, and alert aggregation—exhibit the most severe evaluation gaps. The sub-tasks at the spatial (localization) and procedural (decision) boundaries are better served by existing metrics. This is not coincidental: the existing evaluation paradigm was designed for spatial analysis of static images and for procedure-level clinical endpoints. The temporal middle layer of the pipeline—where video AI fundamentally differs from image AI—is the blind spot.

The GI Genius De Novo submission included standalone testing on 150 colonoscopy videos comprising 5.8 million frames, measuring per-lesion sensitivity and false-positive object rates. It also included an RCT measuring ADR as the primary endpoint. Both the standalone testing and the clinical trial produced robust evidence for their respective dimensions. But neither addressed detection latency, temporal consistency, or alert burden dynamics. Notably, the cloud-based CADDIE system's EAGLE trial [2] reported network latency as a system performance metric—demonstrating that developers *can* measure temporal properties when architecturally motivated, even though regulators do not require them. The 510(k) submissions for subsequent devices inherited this metric framework via the predicate device pathway, perpetuating the gap [3].

The implications extend beyond colonoscopy. The FDA's Predetermined Change Control Plan (PCCP) guidance [11] now permits pre-authorized modifications to AI algorithms post-clearance, provided changes remain within pre-specified bounds. But if the evaluation dimensions used to define those bounds are temporally agnostic, algorithm updates that improve per-frame sensitivity while degrading temporal consistency would fall within the permitted change envelope—undetected by the regulatory framework.

IV. RECOMMENDATIONS

Based on the ERA analysis, we propose four concrete metrics and a phased regulatory integration pathway.

R1: Temporal Detection Latency Protocol. Standardized measurement of time-to-first-alert from lesion FOV entry, computed across varied scope movement speeds. Report as median and 95th percentile latency (in frames and milliseconds), not mean values, to capture tail behavior that impacts clinical safety.

R2: Alert Burden Index (ABI). A composite metric that captures the *frequency*, *duration*, and *temporal clustering* of false-positive alerts per unit procedure time. Unlike per-frame FP rate, the ABI distinguishes between a single sustained alert and many brief flickering alerts—a distinction with significant implications for cognitive load and alert fatigue [8].

R3: Temporal Consistency Score (TCS). A frame-to-frame prediction stability metric computed as the normalized autocorrelation of detection confidence over sliding temporal windows. TCS distinguishes genuine uncertainty (confidence varies because the clinical ground truth is ambiguous) from artifactual inconsistency (confidence varies due to model instability). Crucially, TCS operationalizes a form of temporal uncertainty quantification: a low TCS signals that the system's predictions carry high temporal uncertainty, even if per-frame confidence scores appear high. Reporting TCS alongside per-frame sensitivity provides a two-dimensional characterization of system behavior that bridges evaluation and UQ.

R4: Lesion-Event Correspondence Accuracy. A metric assessing the system's ability to map frame-level detections to discrete lesion events, measured against temporally annotated

ground truth. This metric directly addresses the gap between frame-level evaluation and clinically meaningful lesion-level assessment.

Integration Pathway. We propose a phased approach: (1) *Immediate*: incorporate temporal metrics into standalone performance testing as supplementary reporting, using existing benchmark video datasets; (2) *Near-term*: integrate temporal metric bounds into PCCPs, ensuring that algorithm updates preserve temporal performance; (3) *Long-term*: establish temporal metrics as required evaluation dimensions in regulatory submissions, informed by post-market evidence from phases 1 and 2.

V. Discussion

The ERA framework reveals that the evaluation–regulation gap in medical video AI is not a matter of insufficient evidence but of *misaligned evaluation dimensions*. The clinical trials supporting CADe clearance are methodologically rigorous [1], [2], [4]. The regulatory guidance is evolving thoughtfully [10], [12]. The gap lies in the *mapping* between what trials measure and what video AI systems actually do during clinical use.

This disconnect has direct implications for uncertainty quantification. Video AI systems operate under compound uncertainty—arising from data noise, model limitations, and temporal dynamics—yet existing UQ methods remain anchored to static prediction settings [15], [16]. The six unmeasured dimensions in Table I each represent a source of temporal uncertainty invisible to current UQ frameworks: detection latency uncertainty (will the system detect within a clinically acceptable window?), prediction consistency uncertainty (is frame-to-frame variability epistemic or artifactual?), and alert burden uncertainty (what is the cognitive cost of the system's uncertainty expression to clinicians?). Developing UQ methods that address these temporal dimensions—and embedding them within regulatory evaluation—is a natural extension of this work.

This structural gap has practical consequences. A CADe system that achieves high ADR in a trial may do so through sustained, reliable detection of easily visible polyps while exhibiting high detection latency for briefly visible flat lesions—a failure mode that ADR alone cannot reveal, but that Rex et al. [7] have identified clinically. Similarly, two systems with identical per-frame false-positive rates may impose radically different cognitive burdens depending on the temporal distribution of their false alerts.

The broader context reinforces urgency. As of mid-2025, over 1,250 AI-enabled medical devices have received FDA authorization [9], [10], yet generalizability evidence remains limited. The video AI domain compounds this challenge because the temporal dimension introduces failure modes invisible to static evaluation.

Limitations. This work is an analytical framework, not an empirical validation. The proposed metrics require standardized benchmarking datasets with temporal ground-truth annotations; such datasets exist for colonoscopy (e.g., SUN Colonoscopy Database, LDPolypVideo) but lack the temporal annotation granularity needed for metrics R1–R4. The ERA framework has been applied here to a single domain; validation across capsule endoscopy, surgical video, and ultrasound AI is needed.

Future Work. Empirical validation of the proposed metrics using temporally annotated public datasets. Multi-stakeholder consensus process involving regulators, clinicians, and AI developers to prioritize metric adoption. Extension of the ERA framework to other medical video AI domains and to emerging regulatory frameworks including the EU AI Act.

VI. Conclusion

The evaluation paradigm for medical video AI was inherited from static diagnostic imaging and has not kept pace with the temporal complexity of deployed systems. The ERA framework provides a systematic tool for identifying where evaluation metrics, regulatory requirements, and clinical needs diverge. Applied to colonoscopy CADe, it reveals that the temporal middle layer of the AI pipeline—detection onset, cross-frame tracking, and alert aggregation—constitutes a regulatory blind spot. Closing this gap requires not new clinical trials but new evaluation dimensions: temporal detection latency, alert burden dynamics, prediction consistency, and lesion-event correspondence. These metrics are technically feasible, clinically meaningful, and regulatorily actionable. We invite the medical video AI community to adopt, validate, and refine them.

Acknowledgment

This work received no external funding. We thank the organizers of the 1st Workshop on Medical Video AI Assessment and Uncertainty Quantification at ISBI 2026 for creating a forum for this critical discussion.

Compliance with Ethical Standards

This work is a theoretical analysis involving no human subjects, patient data, or animal experimentation. No ethical approval was required.

References

[1] A. Repici *et al.*, "Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial," *Gastroenterology*, vol. 159, no. 2, pp. 512–520, 2020.

[2] R. Kader *et al.*, "A novel cloud-based artificial intelligence for real-time detection of colorectal neoplasia—a randomized controlled trial (EAGLE)," *npj Digit. Med.*, 2025.

[3] A. El-Sayed, L. B. Lovat, and O. F. Ahmad, "Clinical implementation of artificial intelligence in gastroenterology: Current landscape, regulatory challenges, and ethical issues," *Gastroenterology*, vol. 169, no. 3, pp. 518–530, 2025.

[4] C. Hassan *et al.*, "Real-time computer-aided detection of colorectal neoplasia during colonoscopy: A systematic review and meta-analysis," *Ann. Intern. Med.*, vol. 176, no. 9, pp. 1209–1220, 2023.

[5] M. F. Kaminski *et al.*, "Quality indicators for colonoscopy and the risk of interval cancer," *N. Engl. J. Med.*, vol. 362, pp. 1795–1803, 2010.

[6] D. K. Rex *et al.*, "Quality indicators for colonoscopy," *Gastrointest. Endosc.*, vol. 100, no. 3, pp. 352–381, 2024.

[7] D. K. Rex *et al.*, "Detection of large flat colorectal lesions by artificial intelligence: A persistent weakness and blind spot," *Gut*, vol. 74, pp. 881–883, 2025.

[8] M. Heinlein, "Artificial intelligence and the clinical gaze: Visual practices of AI-assisted colonoscopy," *Sociology of Health & Illness*, 2025.

[9] D. Windecker *et al.*, "Generalizability of FDA-approved AI-enabled medical devices for clinical use," *JAMA Netw. Open*, vol. 8, no. 4, e258052, 2025.

[10] H. J. Warraich, T. Tazbaz, and R. M. Califf, "FDA perspective on the regulation of artificial intelligence in health care and biomedicine," *JAMA*, vol. 333, no. 3, pp. 241–247, 2025.

[11] U.S. FDA, "Marketing submission recommendations for a predetermined change control plan for AI-enabled device software functions," Final guidance, Dec. 2024.

[12] U.S. FDA, "AI-enabled device software functions: Lifecycle management and marketing submission recommendations," Draft guidance, Jan. 2025.

[13] M. Aboy, T. Minssen, and E. Vayena, "Navigating the EU AI Act: Implications for regulated digital medical products," *npj Digit. Med.*, vol. 7, 237, 2024.

[14] W. Y. Lai *et al.*, "Artificial intelligence in colonoscopy: Where are we now in 2024?" *Digestion*, 2025.

[15] L. J. Lechuga López and F. E. Shamout, "Uncertainty quantification for machine learning in healthcare: A survey," *arXiv preprint arXiv:2505.02874*, 2025.

[16] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, 2021.