Phenotype-Resolved Active Surveillance Architectures for Early Detection of Anastomotic Leak in Postoperative Gastrointestinal Surgical Recovery Pathways

Juan Esteban[1], Carlos Andrés[2], Miguel Ángel[3]

[11]Departamento de Ingeniería de Sistemas, Universidad de Sucre, Carrera 28 No. 5-267, Barrio Puerta Roja, Sincelejo, Sucre, Colombia
[22]Departamento de Ingeniería Informática, Universidad del Tolima, Barrio Santa Helena Parte Alta, Calle 42 No. 1-02, Ibagué, Tolima, Colombia
[33]Departamento de Ciencias de la Computación, Universidad Surcolombiana, Avenida Pastrana Borrero Carrera 1, Barrio Cándido Leguízamo, Neiva, Huila, Colombia

ABSTRACT Background conditions after gastrointestinal reconstruction make anastomotic leak difficult to detect at the moment when escalation is most likely to preserve physiological reserve. In the earliest postoperative interval, tissue injury from the operation, resuscitation effects, analgesia, transient ileus, inflammatory adaptation, and ordinary hemodynamic volatility all produce signals that overlap with the initial manifestations of leak. The bedside problem is therefore not only one of forecasting a later adverse outcome, but of determining when a patient has departed from an expected recovery path in a way that warrants new diagnostic attention before overt deterioration has made the diagnosis easier but less useful. This paper develops a technical framework for early leak detection built around phenotype-resolved surveillance rather than static risk scoring. The framework treats leak as a family of latent postoperative trajectories whose observability depends on host reserve, surgical anatomy, care intensity, and intervention history. It integrates delayed-supervision learning, irregularly sampled multimodal observations, hidden phenotype dynamics, uncertainty-aware alarming, and active diagnostic allocation. Special emphasis is placed on the distinction between biological onset, inferable deviation, clinical suspicion, and formal confirmation, because those moments are often separated in practice and should not be conflated in model development. The resulting perspective recasts postoperative leak surveillance as a sequential clinical inference problem in which the main objective is to shorten the interval between the first inferable abnormal trajectory and the first effective diagnostic or therapeutic response, while containing false-alarm burden, preserving calibration, and maintaining transportability across hospitals with different postoperative workflows.

I. Introduction

Anastomotic leak presents a peculiar analytic difficulty because it begins as a local failure of healing but is usually perceived through remote and delayed manifestations [1]. The anastomosis itself is not continuously visible after the operation, and the earliest biological events that matter most for timely intervention occur within a microenvironment defined by perfusion adequacy, tissue strain, edema, bacterial burden, matrix remodeling, and the quality of local containment. The postoperative record does not provide direct access to that microenvironment. Instead, it offers a shifting collection of indirect observations. Some are numerical, such as inflammatory markers, metabolic variables, and hemodynamic trends. Some are narrative, such as nursing descriptions of abdominal change or a surgeon's qualitative concern about the trajectory of recovery. Some are procedural, such as transfer to a higher-acuity setting, repeated laboratory testing, or new imaging. Each of these channels is informative, yet none is a transparent window into the hidden local state. Early detection therefore belongs to the broader class of inference problems in which the clinically important event is latent, its manifestations are distributed across asynchronous measurements, and the measurements themselves are shaped by care decisions.

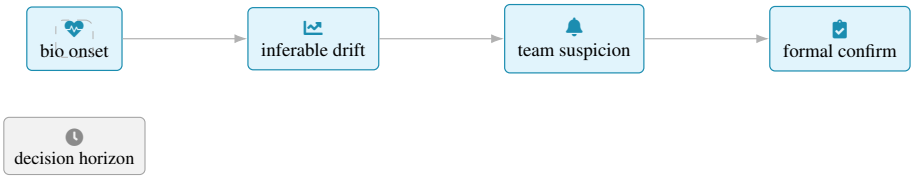

FIGURE 1: Event semantics separating biological onset, inferable deviation, clinician suspicion, and formal confirmation. The interval between inferable deviation and suspicion is the clinically useful horizon for surveillance, because a calibrated alert in this window can redirect monitoring or diagnostics before overt deterioration.

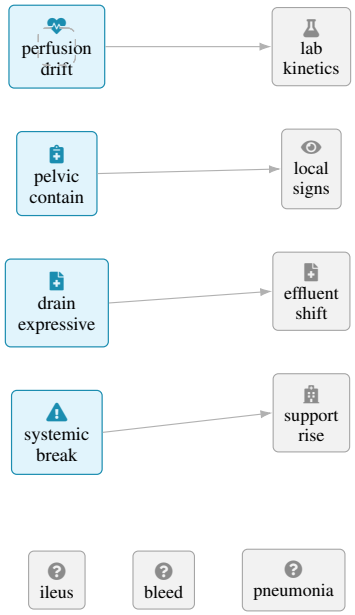

FIGURE 2: Phenotype-resolved declaration space. Distinct leak modes project onto different dominant cue channels, while nearby postoperative mimics occupy overlapping neighborhoods. A pooled classifier tends to emphasize only the highest-amplitude shared signals, whereas phenotype resolution allows earlier, mode-specific evidence to contribute.

That structure has consequences. It means that early leak detection should not be treated as a simple extension of retrospective complication prediction. In static prediction, the usual question is whether a patient belongs to the subset that will later satisfy a given outcome definition. In postoperative surveillance, the question is temporally sharper. One asks whether, at the present moment, the trajectory has become sufficiently incompatible with expected recovery that ordinary monitoring is no longer proportionate. That difference is substantial [2]. A patient can be at elevated baseline susceptibility and nevertheless recover normally. Another can be relatively unremarkable preoperatively yet begin to diverge within hours because of local ischemia, tension, perfusion failure, or contamination. A clinically relevant model must therefore do more than summarize baseline risk. It must interpret the actual path being traversed.

The field has often proceeded as if the label itself were obvious. Yet the event called anastomotic leak is usually recognized only after clinicians accumulate enough evidence to justify a radiologic, procedural, or operative judgment. Formal diagnosis is thus a late observational milestone, not a guaranteed marker of the earliest actionable state. This distinction matters because systems built on diagnosis time alone are rewarded for discovering the signs that cluster near recognition rather than the subtler deviations that arise earlier. The challenge is amplified by the fact that these earlier deviations are rarely unique to leak. They may also appear in pneumonia, ileus, bleeding, urinary infection, aspiration, pancreatitis, or ordinary inflammatory exuberance.

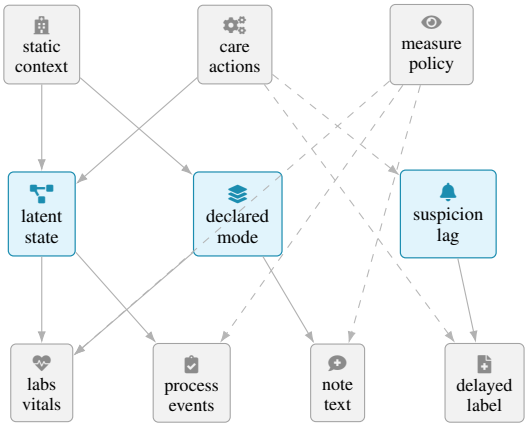

FIGURE 3: Delayed-supervision generative architecture for irregular multimodal surveillance. Static context and care actions drive a latent biological state, declaration mode, and suspicion process; the observation policy shapes what is measured; confirmation remains a delayed consequence rather than a direct proxy for onset.

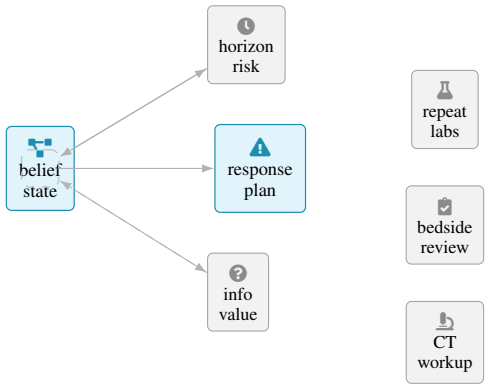

FIGURE 4: Uncertainty-aware control layer. Posterior belief is translated into next-step diagnostic allocation by combining shrinking horizon risk with expected value of information. The policy therefore distinguishes between cases that warrant immediate escalation and cases where targeted clarification is the better first move.

The diagnostic task is therefore not generic deterioration detection. It is discrimination among neighboring postoperative trajectories that share some features but differ in latent origin, temporal ordering, and optimal response.

A notably useful clue from recent structured-record work is methodological rather than celebratory: the most durable lesson is that early postoperative signal may exist while major obstacles remain in external validation, subgroup bias analysis, and the translation of risk output into a real bedside workflow, a set of unresolved issues that Sadanandan (2024) identified with unusual directness [3]. That point deserves emphasis because it discourages an overly narrow reading of model performance [4]. A moderate retrospective discrimination statistic may coexist with poor transportability, unexamined demographic disparity, and no defined action policy. Conversely, a model with modest raw discrimination may still be clinically worthwhile if it is calibrated, interpretable, and timed to trigger useful diagnostic escalation before routine care would have done so.

The natural temptation in a rare but serious complication is to ask for a single definitive early signal. In practice, leak behaves too variably for that expectation to be realistic. Some cases evolve through perfusion-dominant tissue failure with initially modest systemic response. Others declare themselves through contained pelvic inflammation, prolonged ileus, and subtle but persistent drift in inflammatory variables. Others progress rapidly into diffuse contamination and systemic collapse. These are not just different severities of one phenotype. They are different ways in which local failure becomes observable. The implication is that a surveillance architecture should not assume that all leak cases approach recognition along one common path. It should instead represent a space of latent phenotypes, each with its own mapping into laboratories, vital signs, local findings, care-process features, and documentation.

This phenotype-resolved view changes what counts as a good model. The objective is no longer to learn one boundary between leak and nonleak in a pooled feature space. The

objective is to estimate whether the patient is entering one of several leak-compatible trajectories and, if so, which one is most plausible under the available evidence [5]. Such a model may use different combinations of signal for different patients. In one patient, persistent mismatch between expected inflammatory decay and observed laboratory kinetics may dominate. In another, the crucial early clue may be a process variable such as unexpectedly sustained high-acuity monitoring combined with a specific local textual descriptor. In still another, the trajectory may remain ambiguous until targeted information is obtained. Under these conditions, surveillance and active diagnostic allocation become inseparable. The model should not merely report risk. It should also estimate what additional observation would most efficiently reduce uncertainty.

The organizational premise of this paper is therefore different from a standard prediction study. It begins by separating four times that are often conflated: the latent biological initiation of meaningful failure, the earliest time at which that failure leaves an inferable footprint in routine data, the time at which clinicians become suspicious enough to alter investigation, and the time of formal confirmation. Once those moments are distinguished, one can define what early detection ought to mean and why ordinary labels only imperfectly supervise it. The paper then develops a phenotype-space account of leak emergence, not to multiply categories for their own sake, but to explain why pooled early-warning systems often underperform at the bedside. From there it constructs a delayed-supervision generative model for irregular multimodal postoperative data, embeds measurement policy and intervention history inside the observation process, and formalizes active sensing through diagnostic value-of-information. The final sections turn to safety, calibration, implementation failure, and the architecture of a prospective translational study.

The argument remains deliberately technical and deliberately moderate. The aim is neither to promise that a model can replace clinical judgment nor to claim that early leak detection is solved by any single algorithmic family [6]. The ambition is narrower and more defensible. It is to specify a mathematically coherent way of treating postoperative leak surveillance as a dynamic clinical inference problem whose central burden is temporal, heterogeneous, and action dependent. Under that framing, the quality of a system depends not only on whether it predicts an eventual code or reoperation, but on whether it can locate the earliest interval in which the patient has begun to leave the domain of plausible healing and whether it can do so in a manner that clinicians can trust, audit, and act on.

II. Event Semantics and the Clinical Decision Horizon

The first step in constructing an early-detection architecture is to decide what event is actually being detected. In many studies the answer is implicit. The event is whatever the dataset says it is, often a coded complication, a radiologic statement, a drainage procedure, or a return to the operating room. Those endpoints are understandable because they are concrete, but they fuse together several distinct phenomena. A coded complication is an administrative artifact. A radiologic statement is a formalized interpretation of an image that was already ordered because someone had become concerned. A return to the operating room is a therapeutic consequence of accumulated evidence and changing physiology. None of these, by itself, is the earliest moment at which meaningful local failure might have become inferable. If one does not separate those moments, the notion of early detection becomes slippery enough to accommodate almost any retrospective classifier.

It is useful to define four times for each patient. Let t_b denote biological initiation, meaning the earliest time at which the anastomosis enters a state of clinically meaningful failure or unstable compromise [7]. Let t_i denote inferable deviation, meaning the first time at which the available data contain enough signal that a sufficiently strong model could, in principle, distinguish the patient from ordinary recovery better than chance. Let t_s denote clinician suspicion, meaning the first time at which the care team behaves as if leak is materially on the differential, through imaging, specific reassessment, extended observation, or other targeted action. Let t_c denote confirmation, meaning the time at which formal documentation or procedural evidence satisfies the institution's definition of leak. These times satisfy

$$t_b \le t_i \le t_s \le t_c,$$

but the gaps between them vary greatly across patients. In some cases t_i and t_s nearly coincide because the presentation is dramatic. In others they may be separated by many hours. The interval $[t_i, t_s)$ is the interval that matters most for a surveillance system, because it is where model assistance could plausibly change what happens next.

Once these times are distinguished, the relevant performance target changes. A system that identifies patients just before t_c may still achieve strong discrimination, especially if later imaging-related or documentation-related features are included. Yet such a system contributes little to early rescue. By contrast, a system that raises calibrated concern between t_i and t_s, even if not perfectly specific, may shorten the delay to imaging, source control, or intensified review. This observation shifts attention away from global area-under-curve metrics toward time-indexed performance [8]. The proper question becomes how often the system can enter the interval of clinical usefulness, how long before t_s that

happens, and how much false-alarm burden is required to do so.

The semantics of t_i are especially important because it is not directly observed. Unlike t_s or t_c, inferable deviation is a latent construct. It refers not to what clinicians did, but to what could have been known from the data then available. In that sense it is model relative. A richer model may place t_i earlier than a simpler one because it can exploit weak multivariate structure that humans or threshold rules do not easily notice. This does not make t_i arbitrary. It means that the early-detection problem is genuinely epistemic. One is asking when the evidence crossed from being too weak to support action to being strong enough that a rational observer should have updated meaningfully. If that boundary is not modeled explicitly, then one cannot distinguish true early detection from timely rediscovery of what the care team already suspected.

The postoperative care environment complicates this further because actions influence both t_s and the data observed afterward. A clinician may order repeat laboratories because of a vague concern, thereby creating a denser signal in the record. Another patient with similar latent physiology may remain on a sparse ward schedule and not declare themselves until much later. In one sense this means that t_i itself is partly action dependent, since inferability depends on the observations acquired. But the dependence is not complete [9]. Some trajectories are already detectable from routine information, while others require targeted information acquisition. Thus a surveillance model must reason in two layers. First, given the current routine data, how likely is it that the patient has already entered a leak-compatible course? Second, if uncertainty remains substantial, which next observation would most usefully reduce it?

The concept of a clinical decision horizon follows from these definitions. The horizon is the forward-looking interval inside which a changed diagnostic or monitoring action would plausibly alter patient management in a meaningful way. Its length depends on anatomy, host reserve, likely leak phenotype, and institutional response speed. A patient with stable physiology and suspected contained leak may have a relatively long horizon for imaging and drainage. Another with rapidly worsening systemic compromise may have a short horizon before intervention becomes unavoidable. A model that does not account for horizon length may alert too late in fast trajectories and too aggressively in slower ones. Hence the surveillance objective must be horizon aware. It should estimate not only whether leak-compatible deviation is present, but also whether the remaining time for low-burden clarification is shrinking.

This perspective also changes the role of baseline risk. Baseline covariates such as nutritional depletion, procedure complexity, or emergency status are important because they shape the prior probability that t_b will occur at all and may influence the gap between t_b and t_s. But they do not define the decision horizon on their own [10]. The horizon is determined by the evolving interplay between latent pathology, observable signal, and clinical reserve. A frail patient may require action at lower posterior risk because the cost of waiting is greater. A robust patient with ambiguous findings may justify a more information-seeking posture. Therefore baseline risk should initialize surveillance, not substitute for it.

Event semantics are equally relevant for evaluation datasets. Suppose a model is trained against a label derived from reoperation plus antibiotics plus diagnosis code within thirty days. Such a label may be reasonable for defining confirmed complications, but it is coarse for supervising t_i. If one uses it naively, the model will often learn the strongest signals near t_s or t_c, especially care-process variables that are highly informative but temporally late. To prevent this, labels should be supplemented with adjudicated timelines where possible, or the model should be built with latent onset and latent inferability as hidden variables rather than assuming that the confirmation label tells the whole story. This is not a mere technical refinement. It is what allows the architecture to target the clinically relevant interval instead of optimizing primarily for late recognition.

The distinction between t_s and t_c has another practical implication. In many institutions, the important diagnostic threshold is not formal confirmation but the moment at which the team commits to leak-directed investigation. A model that consistently acts before that moment can still be valuable even if it does not precede the eventual charted diagnosis by a large margin [11]. Accordingly, implementation studies should measure lead time against the first leak-directed action, not only against the final documented endpoint. The surveillance system enters clinical practice through workflow, not through code books.

A final consequence of this event-semantic framing is conceptual. It makes clear that early leak detection is not a binary property of a model but a relationship among the model, the available data, and the care process into which the model is inserted. A model can be early in one institution and late in another because workflows differ. It can be early for one phenotype and late for another. It can be early under dense observation and not under sparse observation. Thus the object of design is not simply an algorithm. It is an algorithm-plus-policy system tuned to the interval between inferability and ordinary recognition. That interval, not the eventual complication label alone, is where the scientific and clinical interest ought to concentrate.

III. Latent Phenotype Space of Early Leak Declaration

If leak were expressed through one stereotyped sequence, the modeling problem would be difficult but conceptually simple. A single hidden-state model could be trained to detect that sequence earlier than clinicians do. The reality is more unruly. Anastomotic failure is a family of postoperative trajectories whose early manifestations depend on anatomy, local containment, diversion status, host reserve, tissue quality, and the pace at which contamination escapes local buffering. That heterogeneity should not be treated as nuisance variance around one canonical event [12]. It is part of the event itself. The same binary endpoint may arise through markedly different prediagnostic pathways, and a surveillance system that ignores that fact will often default to late, high-amplitude cues because they are the only features shared reliably across all cases.

One can describe a latent phenotype space in which each phenotype corresponds to a distinct mode of early declaration rather than merely a different severity level. A perfusion-dominant phenotype may begin with tissue compromise and modest systemic response, generating weak but coherent abnormalities in metabolic clearance, local pain trajectory, and delayed physiologic normalization. A contained inflammatory phenotype may evolve through prolonged ileus, small temperature oscillations, slow inflammatory rise, and increasingly suspicious local findings without dramatic hemodynamic collapse. A drain-expressive phenotype may become inferable comparatively early because local effluent changes provide proximal evidence of contamination. A rapidly propagating systemic phenotype may produce abrupt tachycardia, rising lactate, organ-support requirement, and a short decision horizon. These descriptions are not intended as a fixed taxonomy to be applied rigidly. They illustrate that early leak signal is distributed across phenotypic directions in postoperative state space, not along one universal axis.

The presence of multiple phenotypes changes how measurements should be interpreted. In a pooled model, one often asks whether a variable has high feature importance overall. That question can be misleading. A variable may be highly informative for one phenotype and weak for another, and its aggregate importance can therefore obscure the way it truly contributes to early detection. Persistent lactate deviation may be strongly informative in perfusion-dominant trajectories but much less so in contained pelvic processes [13]. Unexpected monitoring intensity may be a powerful clue for some occult cases because it captures bedside concern not otherwise encoded numerically. Drain-related observations may matter greatly when a drain is well positioned and hardly at all when it is absent or remote. A phenotype-resolved architecture does not insist that every variable serve the same purpose in every case. Instead, it learns conditional roles.

It is helpful to think of phenotypes as latent explanatory states rather than hard labels. Let $\phi_t \in \{1, \dots, K\}$ denote the active latent declaration phenotype at time t, with $\phi_t = 0$ representing ordinary recovery or nonleak postoperative trajectories. The phenotype need not be constant. A patient may begin in a low-amplitude occult phenotype and later evolve toward a more systemic one. Early surveillance therefore benefits from modeling phenotype transitions, not only static assignment. Such transitions can encode the biological fact that local processes propagate outward and the clinical fact that observability changes with time. A model that knows which phenotype is likely can weight modalities accordingly and produce more specific alarms. A model that ignores phenotype is forced to treat all deviations as evidence toward one undifferentiated leak class, which often leads either to undercalling subtle cases or to overcalling every unstable patient.

The geometry of this phenotype space can be represented in several ways. One may use a mixture model in which each phenotype defines a different distribution over postoperative trajectories. One may use a switching dynamical system in which the transition matrix between latent states depends on procedure class, baseline reserve, and interventions [14]. Or one may embed trajectories into a continuous latent manifold and interpret phenotypes as regions rather than discrete labels. The appropriate choice is partly technical and partly practical. What matters conceptually is that the early-warning system is allowed to believe that several forms of leak-compatible evolution exist and that the mapping from latent pathology to observed variables is phenotype dependent.

A useful consequence of phenotype resolution is sharper differential diagnosis against competing postoperative syndromes. Bleeding, aspiration, pneumonia, urinary infection, ileus, pancreatitis, and routine inflammatory excess overlap with leak in some dimensions and diverge in others. If leak is modeled as a single class, the decision boundary must cut broadly through an already crowded state space. If leak phenotypes are resolved, the system can distinguish, for example, a perfusion-dominant leak pattern from a hemorrhagic pattern, or a contained inflammatory pelvic pattern from uncomplicated ileus. This does not eliminate ambiguity, but it makes it tractable by replacing one coarse discrimination with several finer ones. The false-alarm burden can thereby be reduced without postponing detection until gross systemic deterioration appears.

Phenotype modeling also permits more coherent use of textual and local signals. Narrative documentation often contains clues that are weak in a pooled classifier because they are rare, semantically variable, or institution specific. Yet these clues may be quite informative within a phenotype. Terms implying unusual drain character, escalating lower abdominal tenderness, or persistent unexplained distension

can become significant when coupled to the right physiological context. In a phenotype-agnostic model, such language can either be ignored as noise or overfit to documentation habits. In a phenotype-resolved model, text contributes by updating the probability that the patient occupies one region of declaration space rather than another [15]. This is a subtler and often more stable role.

The same logic applies to missingness. Absence of certain measurements does not mean absence of information. Sparse monitoring in a seemingly stable patient may be consistent with ordinary recovery, but in a case where subtle local concerns would otherwise raise suspicion, that sparsity increases uncertainty and may alter which phenotype remains plausible. Conversely, dense repeat testing without explicit documentation of concern may shift weight toward occult or perfusion-dominant phenotypes because it implies some latent trigger in the care process. Missingness is therefore phenotype informative when interpreted inside the appropriate latent model. Treating it merely as a nuisance to be imputed leaves useful structure on the table.

Mathematically, phenotype dependence can be introduced by allowing each phenotype to define its own observation model and transition law. If x_t denotes latent biological severity and y_t the observed multimodal record, then

$$\begin{aligned} x_{t+\Delta} &= f_{\phi_t}(x_t, u_t, c) + \eta_t \\ y_t &= g_{\phi_t}(x_t, m_t, c) + \epsilon_t \\ \phi_{t+\Delta} &\sim \Pi(\phi_{t+\Delta} \mid \phi_t, x_t, u_t, c), \end{aligned} \tag{1}$$

where u_t denotes interventions, m_t denotes measurement policy state, and c denotes static context. The phenotype-specific functions f_ϕ and g_ϕ allow the same hidden severity to project differently depending on how the leak is expressing itself. A low-to-moderate severity state under one phenotype may produce little systemic signal, while the same severity under another phenotype may already perturb several routine variables. This is closer to the clinical reality than a single observation equation shared by all leaks.

A more refined representation lets the model infer a distribution over phenotypes rather than choosing one. Let π_t be a simplex-valued vector with components $\pi_t^{(k)} = \mathbb{P}(\phi_t = k \mid \mathcal{F}_t)$. Then the system can report not only total leak probability but also phenotype uncertainty [16]. This matters because action choice can depend on phenotype mix. A case believed to be in a contained phenotype with high uncertainty might trigger imaging for clarification. A case believed to be in a rapidly propagating phenotype with shrinking reserve might justify more urgent senior review. Thus phenotype inference is not just a descriptive embellishment. It directly informs sequential decision making.

The latent phenotype space further clarifies why early-warning performance often varies across procedures. Esophageal, gastric, colorectal, and bariatric reconstructions differ not only in baseline incidence or clinical consequences, but in how early leak states become observable. The same binary endpoint therefore contains procedure-specific phenotype mixtures. A pooled model can struggle because it implicitly averages over those mixtures. A phenotype-resolved architecture can preserve some shared structure while allowing the prevalence and transition logic of phenotypes to depend on anatomy and operative context. This is a principled way to combine generalization with specificity, which is preferable to either naive pooling or complete separation into tiny procedure-specific datasets.

Perhaps the most important conceptual payoff of phenotype resolution is that it dislodges the assumption that the ideal early-warning signal should look the same in every patient. It need not. What should be consistent is the inferential framework: the system should estimate whether the patient is moving into one of the leak-compatible modes of postoperative evolution and whether the evidence is strong enough, persistent enough, and urgent enough to justify altering care [17]. That is a more realistic target than a universal leak signature, and it accommodates the fact that surgical complications are shaped by local anatomy, host biology, and monitoring architecture in ways that no single scalar score can fully absorb.

IV. A Delayed-Supervision Generative Model for Irregular Multimodal Surveillance

The combination of latent onset, phenotype heterogeneity, irregular observations, and treatment-dependent measurement makes a purely discriminative classifier an awkward tool. A generative or partially generative formulation is better aligned with the structure of the problem because it can represent what is hidden, what is observed, and how observation depends on actions already taken. The main challenge is to build a model that is expressive enough for clinical reality without losing the interpretability and temporal discipline required for deployment.

Let x_t denote latent postoperative state, capturing hidden biological progression toward or away from leak-compatible trajectories. Let ϕ_t denote latent declaration phenotype. Let y_t denote observed multimodal data at time t, including structured measurements, process events, and encoded text. Let a_t denote clinician actions or interventions, and let m_t denote observation policy state, meaning the current regime of surveillance density and workflow response. Let d_t denote the observed diagnosis or confirmation process, which is delayed relative to hidden progression. The joint model aims to estimate

$$p(x_{0:T}, \phi_{0:T}, m_{0:T}, d_{0:T} \mid y_{0:T}, a_{0:T}, c),$$

where c collects static context such as procedure class, age, baseline reserve, and operative descriptors.

A convenient formulation is a hidden semi-Markov switching state-space model. The hidden state evolves continuously or in fine discrete time, while the phenotype state changes more rarely and with duration dependence. This is appropriate because biological severity may drift gradually even while the manner of expression remains stable for some interval. One can write [18]

$$\begin{aligned} x_{t+\Delta} &= A_{\phi_t} x_t + B_{\phi_t} a_t + \Gamma_{\phi_t} c + \eta_t \\ \phi_{t+\Delta} &\sim \Pi_\Delta(\phi_{t+\Delta} \mid \phi_t, x_t, a_t, c, \tau_t) \\ \tau_{t+\Delta} &= \tau_t + \Delta, \end{aligned} \tag{2}$$

where τ_t tracks dwell time in the current phenotype and allows semi-Markov duration effects. This dwell-time term matters because contained inflammatory trajectories often exhibit prolonged ambiguous phases, whereas rapidly propagating trajectories may transition quickly to overt instability. Duration dependence therefore carries real clinical content.

The observation model must accommodate heterogeneous modalities and informative missingness. Let $y_t = \{y_t^{(s)}, y_t^{(p)}, y_t^{(n)}\}$ denote structured physiological data, process events, and narrative encodings, respectively. The observation density can be factorized as

$$\begin{aligned} p(y_t \mid x_t, \phi_t, m_t, c) = p(y_t^{(s)} \mid x_t, \phi_t, m_t, c) \\ \times p(y_t^{(p)} \mid x_t, \phi_t, m_t, c) \\ \times p(y_t^{(n)} \mid x_t, \phi_t, m_t, c). \end{aligned} \tag{3}$$

The factorization need not imply conditional independence in practice; latent shared components can still couple these channels. Its usefulness lies in making explicit that different modalities have different noise structures, different susceptibilities to workflow artifacts, and different relevance at different phases of recovery.

Measurement-policy dynamics should not be hidden inside arbitrary missing-value heuristics. The system needs an explicit model for when data are acquired. Let o_t denote a vector of observation indicators and sampling intensities. One may write

$$\begin{aligned} m_{t+\Delta} &= H m_t + J x_t + K a_t + L c + \zeta_t \\ o_t &\sim p(o_t \mid m_t, x_t, c). \end{aligned} \tag{4}$$

This component captures the practical truth that repeated testing, note density, or transfer to a monitored setting may carry information about latent concern even before formal suspicion is documented. Modeling m_t does not remove confounding by itself, but it prevents the architecture from pretending that measurements appeared exogenously. It also allows missingness-aware calibration: a risk estimate based on sparse routine data can be accompanied by larger uncertainty than one based on dense multimodal evidence [19].

The most delicate part of the model is the diagnosis process. Observed confirmation d_t is not identical to latent onset or even to latent inferability. It is a delayed stochastic observation generated after enough evidence has accumulated or after enough concern has triggered targeted testing. A useful way to encode this is to let diagnosis depend on hidden severity, phenotype, and current investigation intensity. Let q_t denote latent clinician suspicion. Then

$$\begin{aligned} q_{t+\Delta} &= \rho q_t + \alpha^\top y_t + \beta^\top a_t + \gamma^\top c + \xi_t \\ d_t &\sim \text{Bernoulli}(\sigma(\lambda_1 x_t + \lambda_2 q_t + \lambda_3 \mathbf{1}\{a_t \in \mathcal{A}_{\text{diag}}\})) . \end{aligned} \tag{5}$$

This does not claim that suspicion is directly measurable; it merely introduces a latent process mediating between hidden pathology and observed confirmation. Such mediation is necessary if one wants the model to learn earlier precursors rather than to overfit the proximate signatures of explicit workup.

Because labels are weak and delayed, the training objective should be built around latent-variable inference rather than direct cross-entropy on fixed time points. Let Θ denote parameters and Z the collection of hidden states. Then a variational or expectation-maximization style objective can be written as

$$\begin{aligned} \mathcal{L}(\Theta) = \mathbb{E}_{q(Z)}[\log p_\Theta(Y, D, Z \mid A, C)] \\ - \mathbb{E}_{q(Z)}[\log q(Z)] \\ + \omega_1 \mathcal{R}_{\text{cal}} + \omega_2 \mathcal{R}_{\text{early}} + \omega_3 \mathcal{R}_{\text{fair}}, \end{aligned} \tag{6}$$

where $\mathcal{R}_{\text{cal}}$ encourages probability calibration, $\mathcal{R}_{\text{early}}$ rewards evidence appearing before overt recognition, and $\mathcal{R}_{\text{fair}}$ penalizes subgroup disparity in clinically relevant operating regions. The presence of an explicit early-detection regularizer is not cosmetic. Without it, the optimizer has every incentive to rely on late, high-confidence signals adjacent to t_s and t_c.

One way to construct $\mathcal{R}_{\text{early}}$ is to weight positive evidence by a function of lead time relative to observed suspicion. If $\hat{t}_i$ denotes the model-implied first inferable deviation and t_s observed suspicion time, then a reward term might increase when $\hat{t}_i$ occurs sufficiently before t_s while avoiding pathological incentives to fire implausibly early on low-quality evidence. A smooth formulation avoids the brittleness of hard thresholding and keeps the latent onset posterior flexible.

The generative formulation also supports patient-specific baselines without requiring a separate model [20]. A baseline uncomplicated trajectory can be represented as the $\phi_t = 0$ regime with context-conditioned dynamics. Leak-compatible regimes are then departures from that baseline family. Inference becomes a question of whether the posterior mass is shifting from $\phi_t = 0$ into one or more positive phenotypes. This has two advantages. First, it ties early warning to expected recovery rather than to global thresholds. Second, it makes clear that not all deviations from baseline imply leak; a patient may temporarily move into a nonleak complication

subspace if the model is extended to include alternative postoperative syndromes.

Irregular timing requires either continuous-time latent dynamics or event-based discretization. The latter is often more practical in hospital data because events are naturally asynchronous. One can represent the record as a sequence $\{(t_j, e_j, v_j)\}_{j=1}^{N}$, where e_j is event type and v_j the associated value or embedding. Continuous-time transition operators then evolve the hidden state between events, and event-specific observation densities update it when an event occurs. This avoids arbitrary resampling to hourly grids and preserves the informative content of event intervals.

Text encodings deserve special handling because they can easily leak late recognition. The generative model should time censor note content and, where possible, factor narrative representation into descriptive and judgmental components. Descriptive content includes local findings, symptoms, and observed changes. Judgmental content includes explicit suspicion or decision language [21]. If these are not separated, the model may learn to trigger primarily when clinicians have already begun to name the problem. Censoring or disentangling text is therefore part of model validity, not simply a preprocessing preference.

A final strength of the generative approach is its compatibility with uncertainty decomposition. The posterior over x_t, ϕ_t, and m_t naturally expresses uncertainty from sparse data, phenotype ambiguity, and observation-policy confounding. This can be surfaced to clinicians in a clinically meaningful way. A posterior leak probability of 0.35 concentrated on one phenotype with dense corroborating evidence is different from a posterior of 0.35 spread thinly across multiple phenotypes under sparse observation. The former may justify immediate imaging. The latter may justify better information first. Pure discriminative classifiers often compress these cases into the same scalar output, thereby losing an important determinant of action.

In summary, delayed-supervision generative modeling fits the problem because it honors the separation between latent progression, phenotype expression, measurement policy, and observed recognition. It provides a principled home for irregular multimodal data, a route to earlier evidence than confirmation labels alone permit, and a probabilistic substrate upon which an active diagnostic policy can be built. The next step is to specify that policy.

V. Active Diagnostic Allocation and Uncertainty-Aware Control

A surveillance system that merely announces risk is incomplete. In clinical reality the important question is what should happen next. Once the posterior over hidden state and phenotype has been updated, the system must decide whether to continue passive observation, gather more information, intensify bedside review, request imaging, or trigger a more urgent escalation pathway [22]. This is not a postprocessing detail. The same posterior can justify different actions depending on the patient's reserve, the likely phenotype, the density of existing observations, and the local cost of delay. Accordingly, early leak detection should be cast as a sequential control problem under partial observability.

Let b_t denote the belief state, summarizing posterior distributions over leak presence, phenotype, suspicion gap, and current uncertainty. Let $\mathcal{A}$ denote possible actions, including continued routine monitoring, short-interval laboratory repetition, focused abdominal reassessment, senior surgical review, computed tomography, contrast study, interventional consultation, or urgent operative evaluation. A rational policy chooses action a_t to minimize expected downstream loss. Formally,

$$a_t^{\star} = \arg\min_{a \in \mathcal{A}} \Big[c(a, b_t) + \mathbb{E}[V(b_{t+\Delta}) \mid b_t, a] \Big], \tag{7}$$

where $c(a, b_t)$ is the immediate cost of action at the current belief state and V is the value function for future expected loss. Although solving the full problem exactly may be impractical, the equation is useful because it states the principle clearly: action should depend not only on current risk but on what the action changes about future knowledge and future harm.

The structure of $c(a, b_t)$ is asymmetrical. Missing a leak-compatible trajectory while waiting can be costly in a nonlinear way because local contamination, organ dysfunction, and the difficulty of source control may worsen with time. On the other hand, false-positive escalation is not negligible. Imaging burdens the patient, consumes resources, and may initiate cascades of low-yield intervention. Therefore the cost of observation under true hidden leak and the cost of escalation under true nonleak should not be treated as mirror images. A useful form is [23]

$$\begin{aligned} c(a, b_t) = c_a + \mathbb{E}_{b_t}\Big[& \alpha\, L_t\, h_t\, \mathbf{1}\{a = \text{observe}\} \\ & + \beta\,(1 - L_t)\, \mathbf{1}\{a \in \mathcal{A}_{\text{high}}\} \Big], \end{aligned} \tag{8}$$

where L_t is hidden leak state, h_t encodes urgency or shrinking decision horizon, and $\mathcal{A}_{\text{high}}$ denotes high-cost escalation actions. The urgency term allows the same leak probability to imply different action as the expected cost of further delay increases.

A practical surveillance system should usually separate diagnostic from therapeutic escalation. In the early interval the most appropriate intervention is often information acquisition rather than treatment itself. This is where value-of-information becomes central. If a is a candidate information-gathering action, such as repeat lactate, formal bedside reassessment, or imaging, then its expected utility is not

reducible to how often it confirms leak. It is better described by how much it decreases expected future loss through better discrimination. One may define

$$\mathrm{VOI}(a \mid b_t) = V(b_t) - \mathbb{E}\Big[V(b_{t+\Delta}^{(a)}) \mid b_t, a\Big]. \tag{9}$$

An action with high value-of-information can be worthwhile even when current leak probability is only moderate, provided uncertainty is materially reduced and the remaining decision horizon is not long enough to justify passive waiting. This is precisely the regime in which early surveillance can make the most difference.

Phenotype resolution changes the action policy substantially. Different phenotypes imply different diagnostic yields from the same action. Imaging may have high expected value in a contained inflammatory phenotype but lower marginal value in a trajectory already dominated by severe systemic instability where bedside review and source control planning matter more immediately. Repeat laboratory sampling may be useful when the posterior is driven by uncertainty in kinetic trends but less useful when local narrative findings already dominate concern. Thus the optimal policy conditions on phenotype mix, not merely on total leak probability [24]. In belief-state terms, the posterior over ϕ_t is part of the sufficient statistic for action.

Uncertainty itself should be decomposed. Epistemic uncertainty arises from limited data, unusual procedure contexts, or distribution shift. Aleatoric uncertainty arises from overlap between leak and nonleak complications despite adequate data. These have different implications. High epistemic uncertainty may justify caution in acting on an alert or prompt a preference for additional data collection before costly escalation. High aleatoric uncertainty may persist even after further measurement and should instead influence calibration of expectations and communication. A surveillance system that compresses both into one scalar "confidence" loses an opportunity to tailor action more intelligently.

Because clinical teams respond poorly to oscillating alerts, instantaneous posterior thresholding is usually inferior to evidence accumulation. Let $r_t = \mathbb{P}(L_t = 1 \mid b_t)$ be the current posterior leak probability. Define a persistence-weighted evidence score

$$\begin{aligned} E_t &= \int_0^t \kappa(t-s)\, r_s \, ds \\ U_t &= \int_0^t \kappa(t-s)\, u_s \, ds, \end{aligned} \tag{10}$$

where u_s is uncertainty and κ a decay kernel. Then action can depend on (r_t, E_t, U_t) jointly. A new abrupt spike in r_t with low accumulated persistence may deserve verification. A sustained moderate r_t with high E_t may merit imaging even if the instantaneous posterior is not extreme. High U_t may tilt the policy toward information-seeking rather than definitive escalation [25]. This triad is more faithful to how clinicians reason than a static score crossing a single line.

One can also encode expected burden on the human system. Alert fatigue is a dynamic cost because repeated false or nonactionable warnings reduce future adherence. Let g_t denote trust state or cumulative alert burden. Then the effective cost of a new alert can be increased as a function of g_t. The control problem becomes

$$\begin{aligned} V(b_t, g_t) = \min_{a \in \mathcal{A}} \Big[& c(a, b_t, g_t) \\ & + \mathbb{E}[V(b_{t+\Delta}, g_{t+\Delta}) \mid b_t, g_t, a] \Big]. \end{aligned} \tag{11}$$

This addition is not cosmetic. In real hospitals, the value of a surveillance system depends partly on preserving attention for the warnings that matter most. A mathematically elegant detector can fail pragmatically if it spends that attention too freely.

Recipient-specific routing should also be part of the control policy. A recommendation to nursing staff might emphasize the need for renewed abdominal examination and drain assessment. A recommendation to the surgical team might emphasize the posterior shift toward a contained inflammatory phenotype and the high expected value of imaging. A recommendation to the intensivist might frame the issue as a divergence between expected recovery and current organ-support dependence. These are different actions operationally, even if they arise from the same latent model state [26]. The architecture should therefore map belief states not directly to one abstract alarm, but to role-specific intervention suggestions.

Control policies must further respect resource and timing constraints. Imaging availability, overnight staffing, transport risk, and patient stability all affect what can rationally be recommended. A policy that ignores those constraints may optimize a theoretical loss while remaining unusable. The belief state should therefore include logistical context, or the action mapping should be localized by institution. This is one reason why transportability cannot be assessed at the model level alone; policy transport matters as much as score transport.

Another subtle point concerns patients approaching discharge. Observation density is about to fall, and opportunities for immediate rescue may narrow. A moderate but persistent posterior in that context may justify extending surveillance or obtaining definitive clarification before release, even when the same posterior would not trigger action in a densely monitored inpatient environment. Thus the control layer should not view time since surgery in isolation. It should also account for anticipated future observability. The value of an information-gathering action rises when the next chance to observe the patient closely will be much later.

By treating early detection as uncertainty-aware control rather than static classification, one recovers the clinical meaning of surveillance. The system is no longer trying merely to name a complication [27]. It is trying to decide, under partial knowledge, when to convert growing concern into low-burden clarification, when to escalate to high-specificity tests, and when the horizon for passive waiting has become unacceptable. That is a more demanding target, but it is exactly the target that matters at the bedside.

VI. Failure Taxonomy, Calibration, and Distributional Instability

No surveillance architecture is useful unless it fails in understandable ways. The rarity of leak, the ambiguity of early postoperative physiology, and the dependence of observed data on workflow together create several distinct failure modes. Lumping these together as "model error" obscures what can be corrected, what must be monitored, and what should disqualify a system from action in certain settings. A rigorous framework therefore needs an explicit taxonomy of failure.

One category consists of label-semantic failure. This occurs when the supervision used for training does not align with the clinical target. If confirmation times are treated as onset, or if outcomes are defined mainly by codes and procedures without regard to latent inferability, the model may become exquisitely tuned to late-stage evidence. Retrospective metrics may still look respectable because the outcome definition rewards such behavior. Yet the system will underperform as an early-warning tool. Label-semantic failure can be mitigated by adjudicated timelines, latent onset modeling, and evaluation against suspicion time as well as confirmation time.

A second category is proxy failure. Here the model does not primarily learn leak-compatible physiology but instead learns proxies for clinician concern, measurement density, or local workflow. Examples include overreliance on ICU transfer, repeat testing frequency, note templates, or procedure-specific documentation style [28]. Proxy features are not useless; some genuinely contain information about latent concern. The failure occurs when they dominate to such an extent that transport to a new site or a changed workflow causes sharp degradation. Proxy failure can be diagnosed through explanation audits and feature ablation under distribution shift. It is best prevented by explicit modeling of observation policy and by regularizing against excessive reliance on workflow variables alone.

A third category is calibration failure. A system may rank patients correctly yet produce probabilities that are not trustworthy in the operating region where actions are taken. This is especially dangerous in rare-event surveillance because a seemingly small miscalibration near action thresholds can translate into large changes in imaging burden or missed detections. Calibration must therefore be assessed dynamically and conditionally. The model can be well calibrated overall and still overestimate risk in the immediate postoperative phase, underestimate risk in certain procedure classes, or become overconfident under sparse observations. Calibration is not a one-time statistic but an operating characteristic that changes with time, phenotype mix, and institution.

A fourth category is representation failure. Here the model lacks sufficient expressive capacity or the wrong latent structure, so it cannot represent some leak phenotypes without waiting for large global abnormalities to appear. This often manifests as good performance for dramatic systemic presentations and poor performance for contained or slow-burning trajectories. Representation failure is less obvious than calibration failure because global metrics can hide it [29]. Phenotype-specific evaluation and cluster-based error analysis are more revealing. If certain phenotype families show systematically shorter lead time or lower sensitivity, the representation likely needs revision rather than mere recalibration.

A fifth category is equity failure. Because surveillance systems allocate attention, they can amplify existing disparities even if they do not explicitly use sensitive attributes. If one subgroup has sparser observations, different documentation styles, or different baseline access to early imaging, then a model may produce later or less certain alerts in that subgroup. Aggregate performance can remain acceptable while clinically important disparities in lead time, alarm burden, or false reassurance persist. Equity evaluation therefore needs to include subgroup calibration, subgroup time-to-alert, false-negative concentration, and action yield, not merely pooled discrimination. It also requires asking whether the disparity arises from biology, from measurement inequality, or from representation choices.

A sixth category is postdeployment drift. A successful model changes clinician behavior. Earlier imaging, more frequent repeat laboratories, and altered documentation patterns all feed back into the data-generating process. The model may then encounter a world in which the patterns it learned are no longer distributed as before. Drift of this kind is not incidental. It is expected [30]. Silent rollout, ongoing recalibration, and drift monitors in latent space are therefore essential. Static retrospective validation without adaptation planning is insufficient for a surveillance model intended to influence care.

These failure modes suggest that calibration should itself be multidimensional. Let $\hat{p}_t$ denote risk prediction at time t. Conventional calibration asks whether $\mathbb{P}(L_t = 1 \mid \hat{p}_t \approx p) \approx p$. In this domain one also needs phase calibration, phenotype calibration, and action calibration. Phase calibration asks

whether probabilities mean the same thing early after surgery as later. Phenotype calibration asks whether the predicted phenotype mix corresponds to what later unfolds. Action calibration asks whether a thresholded recommendation at a given posterior band produces approximately the diagnostic yield that users were told to expect. This last form is especially important because bedside trust depends more on action yield than on abstract probability properties.

Mathematically, one can regularize calibration and disparity jointly. Suppose G indexes subgroups and B prediction bands. A penalty term of the form

$$\mathcal{R}_{\text{fair}} = \sum_{g \in G} \sum_{b \in B} w_{g,b} \left| \text{freq}_{g,b} - \text{conf}_{g,b} \right| + \lambda \sum_{g \in G} \left| \text{Lead}_g - \text{Lead}_{\text{all}} \right| \tag{12}$$

encourages subgroup calibration and narrows disparities in average lead time. Such regularization must be applied carefully to avoid suppressing genuine heterogeneity in risk, but it is preferable to discovering substantial subgroup underperformance only after deployment.

Distributional instability also argues for model introspection beyond scalar performance. One should monitor which latent variables dominate current alerts, whether phenotype allocation has shifted in implausible ways, and whether the observation-policy component is exerting more influence than the physiological component [31]. If a site transfer causes a surge in alerts driven mainly by note frequency or laboratory density rather than by residual physiological structure, the problem is not mere recalibration. It is a warning that the model has drifted toward workflow imitation. Conversely, if physiological residual structure remains stable but baseline prevalence changes, threshold adjustment may suffice.

A mature surveillance program should therefore include a failure dashboard. Such a dashboard would track dynamic calibration, alert yield, subgroup lead time, latent phenotype distribution, contribution of observation-policy features, and divergence between current and training distributions in the hidden-state space. This is not bureaucratic overhead. It is part of what makes the system scientifically interpretable after installation. Without it, a decline in utility may not be recognized until trust has already eroded.

The broader point is that early leak surveillance is fragile for reasons that are understandable in advance. The complication is rare, labels are delayed, measurements are action dependent, and workflows differ. A design that does not anticipate proxy failure, calibration drift, phenotype underrepresentation, and subgroup disparity will not become robust merely by accumulating more data. What is needed is a framework in which these forms of failure are visible, measurable, and correctable. Only then can a technically impressive model become a clinically responsible one.

VII. Translational Study Architecture and Prospective Evaluation

If the objective is to move beyond proof of concept, the study design must reflect the actual clinical burden of early detection [32]. Retrospective development alone is not enough, and even strong silent prospective performance does not settle whether the system changes care beneficially. A translational architecture should therefore proceed in stages, each aligned to a specific question.

The first stage is retrospective event reconstruction. Here the goal is not simply to fit a model, but to recover the fine-grained timeline of postoperative trajectories. That requires assembling event streams with accurate timestamps for surgery end, unit transfers, laboratory arrivals, note creation and finalization, imaging orders, imaging interpretations, bedside reassessments where documented, antibiotic starts, drainage procedures, reoperation, and formal leak coding. A subset of cases should undergo chart review sufficient to approximate t_s and, where feasible, to narrow plausible windows for t_i. The purpose is not perfection in every chart. It is to define an evaluation scaffold that distinguishes early model action from late rediscovery.

The second stage is latent-model development and internal-external validation across multiple hospitals or services. Multi-center data are important not because they are fashionable, but because observation policies differ in ways that can dominate model behavior. If possible, centers should vary not only in volume but also in monitoring architecture, postoperative pathways, and documentation styles. Training should proceed with site-aware regularization and held-out center evaluation. The primary outputs at this stage should include time-dependent sensitivity before first leak-directed investigation, subgroup lead-time distributions, calibration within action bands, and explanation audits showing whether alerts remain physiologically anchored across sites.

The third stage is silent prospective deployment [33]. The system runs in real time but does not alert clinicians. This stage answers operational questions that retrospective analysis often misses: whether timestamps behave as assumed, whether notes arrive with clinically relevant latency, whether imaging orders or medication administrations are reliably captured, and whether calibration remains stable under live data flow. Silent deployment also provides a baseline estimate of what alerts would have occurred before actual clinician actions. That estimate is crucial when designing later interventional stages, because it reveals likely alert burden and the degree to which the system would genuinely precede routine escalation.

The fourth stage is assistive deployment with audit. Here the system becomes visible to clinicians, but actions remain discretionary and are logged carefully. The primary endpoints are not yet patient outcomes. They are usability and behavioral ones: alert acknowledgment, concordance between alert and subsequent diagnostic action, override patterns, reasons for dismissal, and clinician assessment of explanation quality. This is the stage at which role-specific routing, wording, and burden become visible as determinants of success. A technically sound posterior estimate can still fail here if it arrives at the wrong moment, goes to the wrong recipient, or presents its rationale opaquely.

Only after these stages should one proceed to interventional evaluation. A pragmatic design such as a stepped-wedge or cluster-randomized rollout is often more realistic than individual randomization because surveillance systems affect team behavior and workflow. The main outcomes should be chosen to match what early detection can plausibly influence. Time to leak-directed imaging, time to senior surgical review, severity at recognition, organ-support requirement at diagnosis, emergency reoperation burden, and length of stay may be more sensitive and more mechanistically appropriate than crude leak incidence alone. It is entirely possible that a system improves rescue without changing the biological occurrence rate of every leak [34]. Study endpoints should be open to that possibility.

The interventional trial should also include a concurrent burden analysis. Imaging frequency, repeat laboratory orders, additional consultations, and prolonged observation all impose costs. A system that shortens time to diagnosis but doubles low-yield imaging may not be acceptable. Therefore outcome evaluation should pair rescue metrics with resource-use metrics. The ratio between additional diagnostic activity and clinically meaningful earlier detections is an indispensable measure of value.

A translational study should further prespecify subgroup analyses. Procedure type, anatomical level, emergency versus elective presentation, baseline nutritional status, age band, and where available broader demographic categories should all be examined separately for calibration, lead time, and alert burden. Without this step, a system can appear successful while systematically under-serving certain populations. Prespecification matters because subgroup findings discovered only after deployment are harder to interpret and easier to discount.

Ethical evaluation must include how uncertainty is communicated. A bedside alert that expresses unwarranted certainty may accelerate unnecessary invasive action. An alert that communicates uncertainty too diffusely may be ignored. The output should therefore be phrased as a change in surveillance stance, not as a declaration of hidden truth [35]. In prospective studies, one should capture whether clinicians understand this distinction. Misinterpretation of probabilistic surveillance as deterministic diagnosis is itself a deployment hazard.

Another component of translational design is model maintenance. Since the system is expected to alter care patterns, the protocol should specify how recalibration will occur, what latent drift thresholds trigger review, and whether threshold changes are allowed during the trial. Freezing the model indefinitely may preserve trial purity but can be clinically irresponsible if severe drift emerges. Conversely, unrestricted adaptation can obscure attribution. The protocol must balance these concerns explicitly.

A final translational issue is governance of responsibility. When the system fires, who owns the next step? If this is left vague, alerts may produce diffusion of responsibility rather than faster action. Study protocols should therefore specify recipient hierarchy, expected response windows, and documentation pathways. These are often treated as implementation details, yet they are central to whether a surveillance model improves care or simply adds noise to an already complex environment.

Taken together, these stages define a study architecture that matches the true demands of early leak detection. The goal is not only to show that a model can fit the data, but to show that the model can survive live workflow, remain calibrated under site variation, direct attention usefully, and improve the interval between inferable deviation and effective clinical response. Without such an architecture, the field risks repeating a familiar cycle in which retrospective promise outpaces prospective utility [36].

VIII. Conclusion

Early detection of anastomotic leak is best understood as a problem of sequential clinical inference under delayed supervision, phenotype heterogeneity, and action-dependent observability. The complication is not hidden because the postoperative record is empty. It is hidden because the relevant evidence emerges unevenly across modalities, is entangled with treatment and workflow, and often becomes clinically decisive only after a period of ambiguity that retrospective labels do not represent well. A surveillance system intended for this setting therefore has to do more than rank patients by eventual outcome risk. It has to estimate when inferable deviation has begun, how that deviation is likely to be expressing itself, how much uncertainty remains, and what next action would most efficiently reduce harmful delay.

The framework developed here addresses those requirements by reorganizing the problem around event semantics and latent declaration phenotypes. Separating biolog-

ical initiation, inferable deviation, clinician suspicion, and confirmation clarifies what early detection ought to mean. Treating leak as a family of postoperative trajectories rather than as one homogeneous endpoint allows the model to interpret modality-specific evidence conditionally rather than forcing one signal architecture onto every case. A delayed-supervision generative model then provides a principled way to combine irregular multimodal observations with phenotype switching, informative missingness, and diagnosis lag. Active diagnostic allocation extends the system from passive prediction to uncertainty-aware control, while calibration analysis, failure taxonomy, and translational study design ensure that technical performance remains tethered to clinical use.

This approach does not presume that every leak can be made obvious in the earliest postoperative interval. Some cases will remain intrinsically ambiguous until targeted information is obtained. What it does propose is a stricter and more realistic standard for progress. A useful system should identify patients who have likely entered a leak-compatible trajectory before ordinary care would have done so, should signal why that belief is changing, should recommend proportionate next steps rather than indiscriminate escalation, and should maintain those properties across institutions and patient groups. Progress in early leak surveillance will likely come not from a single better classifier, but from models and study designs that take phenotype diversity, delayed labels, workflow feedback, and decision burden as constitutive features of the problem rather than as inconvenient afterthoughts [37].

REFERENCES

[1] J. D. Urschel, C. J. Blewett, W. F. Bennett, J. D. Miller, and J. E. M. Young, "Handsewn or stapled esophagogastric anastomoses after esophagectomy for cancer: meta-analysis of randomized controlled trials.," *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, vol. 14, pp. 212–217, 10 2001.

[2] A. A. Berty, "Manual and mechanical cervical esophagogastric anastomosis after esophageal resection: A comparison," *International Journal of Surgery Science*, vol. 5, pp. 98–101, 4 2021.

[3] B. Sadanandan, "Data-driven risk stratification for anastomotic leak: A proof-of-concept study using electronic health records," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 6, pp. 1–27, 2024.

[4] J. Muniandy, S. R. N. Palaniappan, S. Muthusamy, M. A. Bujang, T. J. Huei, Y. Y. Goh, K. T. T. Long, S. D. Thevendran, and S. S. Sivalingam, "Retrospective cohort study of non-traumatic jejunum and ileum perforation: A multi-center study," *Turkish Journal of Colorectal Disease*, vol. 33, pp. 1–6, 3 2023.

[5] I. P. Doulamis, G. Michalopoulos, V. Boikou, D. Schizas, E. Spartalis, E. Menenakos, and K. P. Economopoulos, "Concomitant cholecystectomy during bariatric surgery: The jury is still out.," *American journal of surgery*, vol. 218, pp. 401–410, 2 2019.

[6] A. R. Jensen, A. S. Wright, L. K. McIntyre, A. E. Levy, H. M. Foy, D. J. Anastakis, C. A. Pellegrini, and K. D. Horvath, "Laboratory-based instruction for skin closure and bowel anastomosis for surgical residents.," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 143, pp. 852–859, 9 2008.

[7] W. H. Allum, E. C. Smyth, J. M. Blazeby, H. I. Grabsch, S. M. Griffin, S. Rowley, F. H. Cafferty, R. E. Langley, and D. Cunningham, "Quality assurance of surgery in the randomized st03 trial of perioperative chemotherapy in carcinoma of the stomach and gastro-oesophageal junction.," *The British journal of surgery*, vol. 106, pp. 1204–1215, 7 2019.

[8] M. Kılıç, S. Madendere, A. B. Eden, F. B. Tekkalan, E. Köseoğlu, and M. D. Balbay, "Impact of urethrovesical anastomotic leakage after robotic radical prostatectomy on early postoperative continence," *Yeni Üroloji Dergisi*, vol. 18, pp. 70–77, 2 2023.

[9] M. S. Abdelhamid, A. Rashad, M. A. Negida, A. Garib, S. Soliman, and T. Elgaabary, "Emergency laparoscopic left sided colonic resection with primary anastomosis: Feasibility and safety," *Archives of Surgery and Clinical Research*, vol. 2, pp. 031–038, 11 2018.

[10] D. A. Telem, E. H. Chin, S. Q. Nguyen, and C. M. Divino, "Risk factors for anastomotic leak following colorectal surgery: a case-control study.," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 145, pp. 371–376, 4 2010.

[11] R. J. Egan, H. Jones, R. W. Radwan, M. Davies, M. Evans, J. Beynon, and D. A. Harris, "Pth-332 preoperative inflammatory cell ratios (nlr and plr) predict post-operative complications and poor outcomes following surgery for rectal cancer," *Colon and Anorectum – Cancer Including Diagnosis, Prevention and Screen*, vol. 64, pp. A555.1–A555, 6 2015.

[12] A. Sharma and B. T. Chinn, "Preoperative optimization of crohn disease.," *Clinics in colon and rectal surgery*, vol. 26, pp. 75–79, 6 2013.

[13] M.-H. Kam, C.-L. Tang, E. S. Chan, J.-F. Lim, and K. W. Eu, "Systematic review of intraoperative colonic irrigation vs. manual decompression in obstructed left-sided colorectal emergencies.," *International journal of colorectal disease*, vol. 24, pp. 1031–1037, 5 2009.

[14] Q. Yin, S. Zhou, Y. S. Song, X. X. Xun, N. L. Liu, and L. Liu, "Treatment of intrathoracic anastomotic leak after esophagectomy with the sump drainage tube," 10 2020.

[15] S. Persson, I. Rouvelas, K. Kumagai, H. Song, M. Lindblad, L. Lundell, M. Nilsson, and J. A. Tsai, "Treatment

of esophageal anastomotic leakage with self-expanding metal stents: analysis of risk factors for treatment failure," *Endoscopy international open*, vol. 4, pp. E420–6, 3 2016.

[16] R. Manta, A. Caruso, C. Cellini, M. Sica, A. Zullo, V. G. Mirante, H. Bertani, M. Frazzoni, M. Mutignani, G. Galloro, and R. Conigliaro, "Endoscopic management of patients with post-surgical leaks involving the gastrointestinal tract: A large case series.," *United European gastroenterology journal*, vol. 4, pp. 770–777, 1 2016.

[17] S. D. Armbruster and F. A. Valea, *Should Routine Mechanical Bowel Preparation be Performed before Primary Debulking Surgery?*, pp. 4–5. Cambridge University Press, 7 2023.

[18] C. K. McCarthy, P. McGaha, N. Rozich, N. A. Yokell, J. S. Lees, and W. L. Berry, "Creation of colonic anastomosis in mice.," *Journal of visualized experiments : JoVE*, 1 2019.

[19] S. B. Bon, M. Rapi, R. Coletta, A. Morabito, and L. Valentini, "Plasticised regenerated silk/gold nanorods hybrids as sealant and bio-piezoelectric materials.," *Nanomaterials (Basel, Switzerland)*, vol. 10, pp. 179–, 1 2020.

[20] , , and , "Preoperative risk factors for anastomotic leaks in colorectal surgery," . , pp. 215–222, 8 2022.

[21] S. I. Jang and D. K. Lee, "Anastomotic stricture after liver transplantation: It is not achilles' heel anymore!," *International Journal of Gastrointestinal Intervention*, vol. 7, pp. 57–66, 7 2018.

[22] P. Waterland, K. Goonetilleke, D. N. Naumann, M. Sutcliff, and F. Soliman, "Defunctioning ileostomy reversal rates and reasons for delayed reversal: Does delay impact on complications of ileostomy reversal? a study of 170 defunctioning ileostomies," *Journal of clinical medicine research*, vol. 7, pp. 685–689, 7 2015.

[23] D. T. Mueller, R. Stier, B. Babic, S.-H. Chon, S. Brunner, L. M. Schiffmann, W. Schröder, C. J. Bruns, and H. F. Fuchs, "Time to endoscopic vacuum therapy: Lessons learned after 150 robotic-assisted minimally invasive esophagectomies at a german high-volume center," *Journal of the American College of Surgeons*, vol. 235, pp. S29–S30, 10 2022.

[24] R. Tyler, H. Foss, L. Phelan, S. Radley, I. Geh, and S. Karandikar, "Impact of surgeon volume on 18-month unclosed ileostomy rate after restorative rectal cancer resection.," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 25, pp. 253–260, 10 2022.

[25] J. S. Lim, N. Brant, J. M. Downs, W. Apple, R. Stadler, and D. R. Jeyarajah, "Minimally invasive lower anterior resections - better than open but not all the same.," *The American surgeon*, vol. 89, pp. 5270–5275, 12 2022.

[26] A. Caycedo-Marulanda and C. P. Verschoor, "Experience beyond the learning curve of transanal total mesorectal excision (tatme) and its effect on the incidence of anastomotic leak," *Techniques in coloproctology*, vol. 24, pp. 309–316, 2 2020.

[27] U. Waheed, R. A. Quraishi, M. Patel, L. Worthington, D. Fidler, C. Heal, and B. Alkhaffaf, "Ogc p48 the relationship between hospital acquired pneumonia (hap) and synchronous post-operative complications following oesophagectomy: An assessment of the impact on mortality," *British Journal of Surgery*, vol. 110, 11 2023.

[28] S. P. T. Tanny, M. Yoo, J. M. Hutson, J. C. Langer, and S. K. King, "Current surgical practice in pediatric ulcerative colitis: A systematic review," *Journal of pediatric surgery*, vol. 54, pp. 1324–1330, 9 2018.

[29] R. B. Rizk, M. A. Mahmoud, H. S. Mostafa, and A. S. Ahmed, "Comparative study between early versus late enteral nutrition after gastrointestinal anastomosis operations," *The Egyptian Journal of Surgery*, vol. 42, pp. 573–583, 10 2023.

[30] M. Z. Chen, J. Cartmill, and A. Gilmore, "Natural orifice specimen extraction for colorectal surgery: Early adoption in a western population.," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 23, pp. 937–943, 12 2020.

[31] M. Zawadzki, M. Krzystek-Korpacka, A. Gamian, and W. Witkiewicz, "Serum cytokines in early prediction of anastomotic leakage following low anterior resection.," *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques*, vol. 13, pp. 33–43, 1 2018.

[32] W. Bohle, I. Louris, A. Schaudt, J. Koeninger, and W. G. Zoller, "Predictors for treatment failure of self-expandable metal stents for anastomotic leak after gastro-esophageal resection.," *Journal of gastrointestinal and liver diseases : JGLD*, vol. 29, pp. 145–149, 6 2020.

[33] G. Köhler, G. Spaun, R.-R. Luketina, S. A. Antoniou, O. O. Koch, and K. Emmanuel, "Early protective ileostomy closure following stoma formation with a dual-sided absorbable adhesive barrier," *European Surgery*, vol. 46, pp. 197–202, 6 2014.

[34] P. J. Nisar, K. A. Appau, F. H. Remzi, and R. P. Kiran, "Preoperative hypoalbuminemia is associated with adverse outcomes after ileoanal pouch surgery.," *Inflammatory bowel diseases*, vol. 18, pp. 1034–1041, 8 2011.

[35] Y. Liu, G. Wang, X. Wan, Y. Cheng, Y. Wang, and X. Liu, "Use of pedicled omentum in the prevention of anastomotic leakage after resection of obstructive colorectal cancer," *Chinese Journal of General Surgery*, vol. 32, pp. 23–25, 1 2017.

[36] G. B. Hanna, P. R. Boshier, A. L. Knaggs, R. D. Goldin, and M. Sasako, "Improving outcomes after gastroesophageal cancer resection: Can japanese results

be reproduced in western centers?," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 147, pp. 738–745, 8 2012.

[37] H. Sohail, T. S. N. Babar, E. Sarwar, F. Sadaqat, M. A. Saleem, and M. M. Ali, "Outcome of colostomy closure without prior conventional bowel preparation," *Pakistan Journal of Medical and Health Sciences*, vol. 16, pp. 596–598, 2 2022.