Sequential Postoperative Surveillance for Anastomotic Leak Using Dynamic State Estimation

Sompong Rattanakorn[1] **and Chaiwat Lertsakul**[2]
[1]Department of Computer Engineering and Digital Technology, Buriram Rajabhat University, 439 Jira Road, Nai Mueang Subdistrict, Mueang Buriram, Buriram 31000, Thailand
[2]Faculty of Information and Communication Technology, Yala Rajabhat University, 133 Tessaban 3 Road, Sateng Subdistrict, Mueang Yala, Yala 95000, Thailand

ABSTRACT Anastomotic leak remains one of the most difficult postoperative complications to recognize early because its emergence is temporally distributed, physiologically heterogeneous, and operationally entangled with routine recovery after major gastrointestinal surgery. Clinical deterioration rarely appears as a single abrupt observation. Instead, it unfolds through a sequence of weakly informative signs whose interpretation depends on timing, co-occurrence, persistence, and response to treatment. This feature of the problem creates a mismatch with computational approaches that estimate risk only before surgery or from a single static postoperative snapshot. A surveillance perspective is more appropriate because it treats leak detection as repeated inference under evolving evidence. This paper develops a technical framework for such surveillance using structured electronic health record data. The proposed formulation models postoperative recovery as a partially observed stochastic system in which latent anastomotic integrity, inflammatory escalation, perfusion disturbance, and support intensity evolve over time while measurements arrive irregularly and with informative missingness. Risk is updated sequentially through dynamic state estimation and horizon-specific event prediction rather than through one-time classification. The paper also derives action-aware alert rules that account for false-alarm burden, timing value, intervention cost, and clinician attention, and it outlines validation strategies centered on lead time, calibration drift, and portability across institutions. The resulting argument is that postoperative surveillance should be treated as a longitudinal decision system in which preoperative information provides an initial prior but not the final answer. The goal is not maximal algorithmic complexity for its own sake, but a disciplined representation of the way postoperative evidence actually accumulates in clinical care.

KEYWORDS

1. Introduction

Anastomotic leak occupies an unusual position in perioperative medicine because it is both a surgical complication and a monitoring problem [1]. The technical failure, if it occurs, originates in tissue approximation, perfusion, tension, healing capacity, microbial contamination, and postoperative stress. Yet the moment at which clinicians can infer that failure with sufficient confidence is typically delayed relative to the operation itself. Between those two moments lies a period of uncertainty in which the patient may show minor physiologic perturbations that are expected after major surgery, subtle signs that are compatible with several benign explanations, or an initially quiet course that later turns toward systemic deterioration. A computational formulation that takes only the preoperative state or a single early postoperative snapshot as its principal input therefore captures only a fraction of the inferential problem. The more faithful formulation asks how belief about an evolving latent complication should be revised as evidence accumulates during recovery.

That distinction is not minor. A preoperative model answers whether a patient appears globally vulnerable before the incision. A postoperative surveillance system answers whether the patient who is already recovering now exhibits a trajectory more consistent with uncomplicated convalescence, nonspecific in-

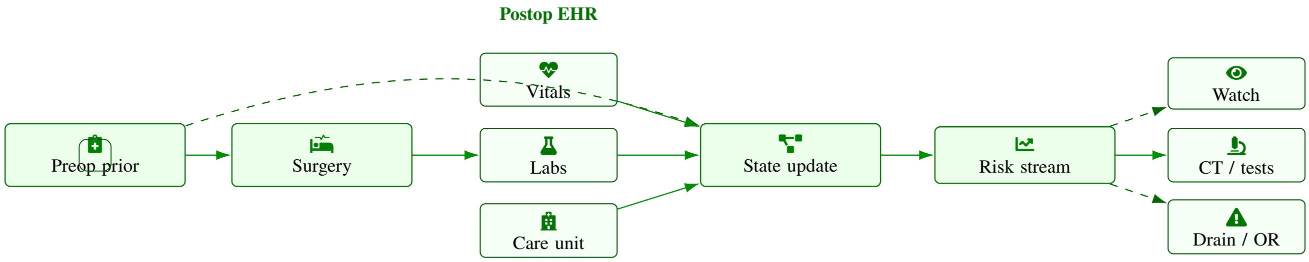


Figure 1 Sequential postoperative surveillance starts from a preoperative prior but becomes driven by continuously refreshed postoperative evidence. The model ingests ward or ICU observations, updates a dynamic state estimate, and maps that estimate to staged actions rather than a single one-time classification.

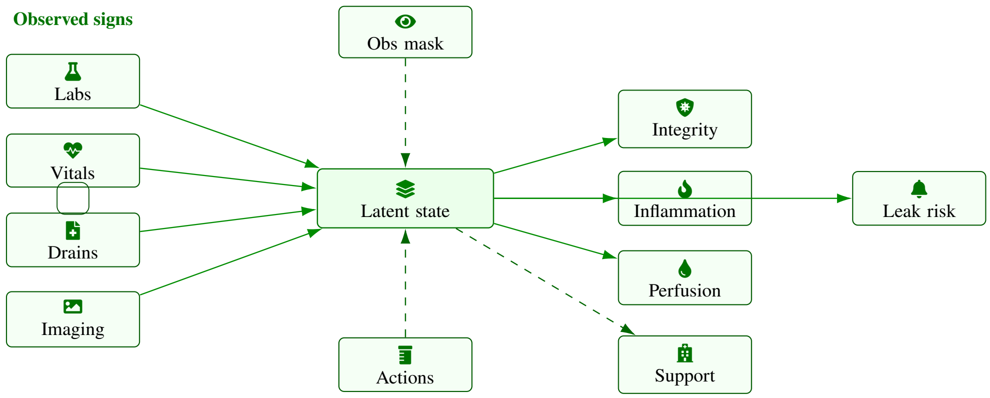


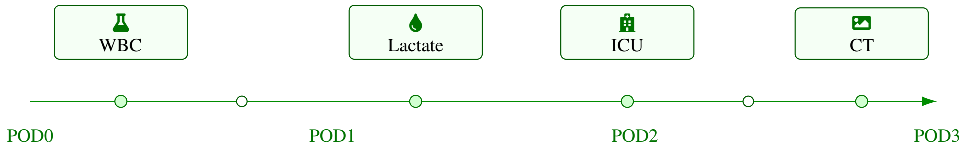
Figure 2 The postoperative period is represented as a partially observed state-space system. Noisy surrogates are fused into a compact latent state that carries hidden anastomotic integrity, inflammatory escalation, perfusion stress, and care intensity, while the observation process and treatment history remain explicit parts of the model.

flammatory stress, or emerging leak. These are related questions, but they are structurally different. The first concerns baseline propensity. The second concerns latent state identification under temporal uncertainty. In practice clinicians care about both, but the second question is the one that becomes urgent on the ward, in the ICU, and during postoperative reassessment. It is the question that governs whether laboratory tests are repeated, whether imaging is obtained, whether a drain is interrogated more aggressively, whether antibiotics are escalated, and whether the threshold for operative re-exploration should shift.

The historical tendency to privilege preoperative prediction is understandable. Preoperative variables are easy to assemble into a tabular dataset, and one outcome label per admission yields a familiar supervised learning problem. Postoperative surveillance is technically messier. Measurements are irregularly spaced, different variables are observed with different cadence, missingness carries clinical meaning, intervention timing contaminates the data generating process, and useful performance cannot be summarized by a single discrimination measure [2]. Yet methodological convenience is not a sound basis for defining the scientific problem. If the clinical task is early recognition during recovery, then the algorithmic formulation should reflect the recovery process rather than simplify it away [3].

A surveillance framing begins with the idea that the patient's postoperative condition is a latent trajectory rather than a fixed category. The anastomosis is not directly observed in continuous time. Instead, clinicians receive scattered signals from vital signs, laboratory tests, support requirements, nursing assessments, imaging, drain output, and operative context. Each signal is noisy. Some are specific but infrequent. Others are common but nonspecific. Their meaning is conditional on timing and on the preceding trajectory. A single leukocyte count, for instance, is much less informative than a sustained elevation whose expected postoperative resolution fails to occur in parallel with rising lactate and increasing support intensity. The inferential problem is therefore sequential and contextual. Computational systems that ignore that structure may still provide risk rankings, but they do not correspond closely to the way clinical knowledge is actually assembled at the bedside.

This shift in framing also clarifies why calibration and operating-point choice become so important. In a surveillance setting the output is not merely a ranking. It is an evolving probability that must be tied to thresholds for observation, diagnostic escalation, and intervention. As Sadanandan (2024) observed in an electronic health record study where postoperative white blood cell count, ICU admission, and postoperative lactate emerged among the strongest predictors, the burden of false alarms and the appropriate balance between recall and precision are implementation decisions rather than purely model-internal properties [4]. That observation is especially consequential for surveillance, because a probability that is not well calibrated over time cannot support action-tiered thresholds, and a detector that performs well only after clinicians have already become suspicious offers limited clinical advantage [5].



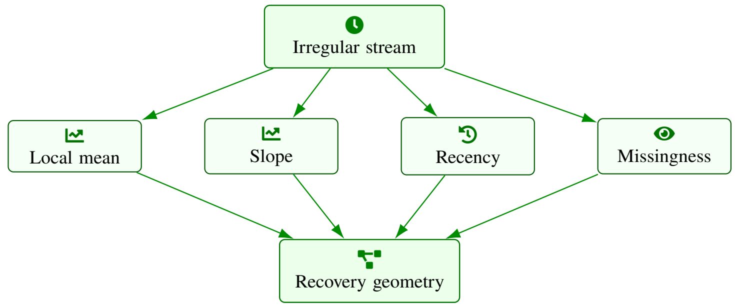


Figure 3 Electronic health record surveillance operates on irregularly sampled trajectories rather than complete evenly spaced panels. The representation therefore preserves local level, direction of change, recency, and informative missingness so that the model can distinguish benign postoperative fluctuation from persistent abnormal recovery geometry.

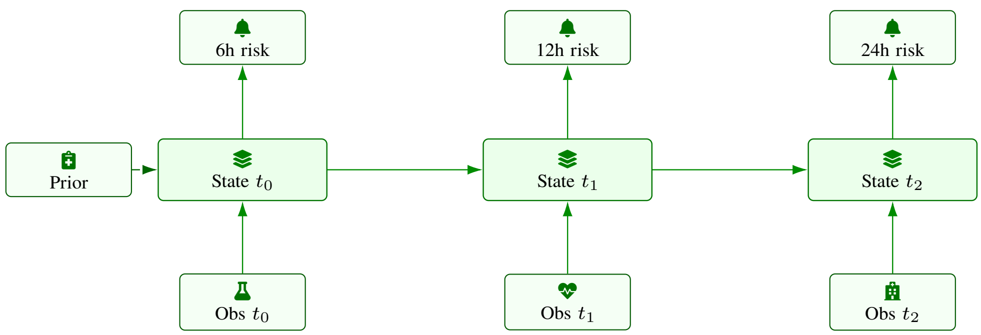


Figure 4 Sequential inference updates leak probability whenever new postoperative evidence arrives. The same patient can move from low prior concern to elevated short-horizon risk and then either stabilize or escalate further, which is why multi-horizon prediction is more faithful than a single admission-level estimate.

Once postoperative monitoring is recognized as the core task, the role of preoperative information changes. Baseline covariates do not disappear. Nutritional status, age, comorbidity burden, surgical urgency, procedure type, operative duration, and baseline renal or inflammatory reserve all matter. But they function most naturally as priors. They define the initial risk landscape from which surveillance begins. After surgery, the relevant question is how those prior beliefs should be updated as the patient generates new data. A patient with low baseline risk but a rapidly worsening postoperative course should soon outrank a patient with high baseline vulnerability but smooth recovery. Static scoring systems can only approximate that reordering. Sequential models are designed for it.

There is also a conceptual reason to favor surveillance over purely preoperative prediction. Postoperative care is already organized around repeated assessment. Nurses record vital signs at intervals conditioned by acuity. Surgeons examine patients repeatedly. Laboratory draws are ordered more frequently when concern rises. Imaging is typically triggered by threshold crossing in clinical suspicion rather than by protocol alone. The hospital, in other words, already behaves like a surveillance system. An algorithm that produces a single estimate before surgery sits somewhat outside that workflow. A dynamic monitoring model can instead be embedded into the rhythm of care, augmenting rather than replacing the serial judgments clinicians already make [6].

The technical implications of this change are extensive. The target variable is no longer a single binary outcome for the entire admission but a time-indexed event probability over clinically chosen horizons. The input is no longer one row per patient but an irregular, intervention-affected sequence. Missingness must be modeled rather than dismissed. Evaluation must quantify lead time, alert burden, and recalibration over time rather than relying only on static discrimination. Interpretation must explain why concern is increasing now, not only which predictors were influential on average in the training data. Deployment must account for clinician attention as a scarce resource. Maintenance must track drift because monitoring practices, case mix, and postoperative pathways evolve.

This paper develops a research framework that takes those implications seriously. It argues that anastomotic leak detection after surgery should be understood as a problem of partially observed dynamic inference combined with constrained decision support. The focus is not on any single architecture, since

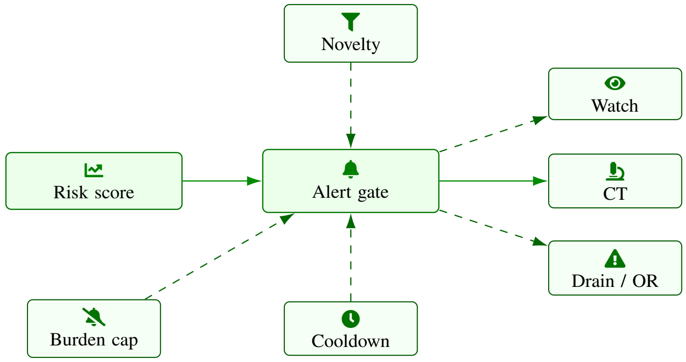


Figure 5 Risk output is translated into staged postoperative actions through an alert controller that accounts for novelty, refractory intervals, and cumulative burden. This structure reflects the fact that the acceptable false-positive rate depends on what the alert is asking clinicians to do, from intensified observation to diagnostic or procedural escalation.

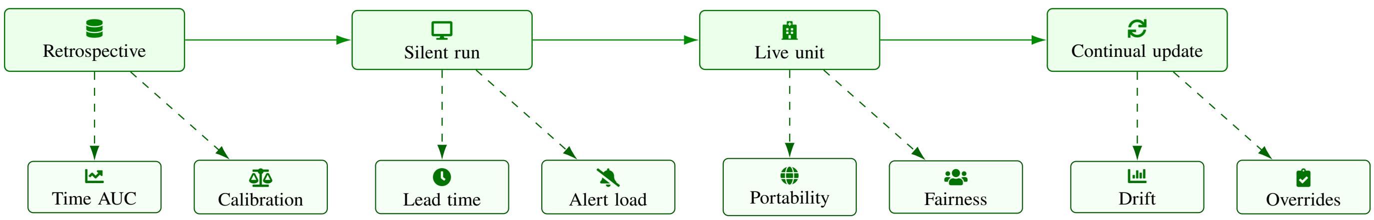


Figure 6 A surveillance model should be evaluated and maintained as part of institutional infrastructure. The progression from retrospective study to silent execution, live deployment, and continual updating emphasizes that lead time, alert burden, portability, fairness, drift, and clinician overrides are all first-class performance targets rather than secondary implementation details.

several families of models can inhabit this framework. The focus is on the structural choices that become unavoidable when the task is defined correctly: how to represent latent recovery state, how to integrate irregular measurements, how to separate biologic signal from the observation process, how to produce horizon-specific probabilities, how to transform those probabilities into staged actions, and how to validate the resulting system in a way that reflects bedside use rather than only retrospective classification convenience.

The argument proceeds by first formulating the postoperative period as a stochastic system with hidden anastomotic integrity and observable clinical surrogates. It then examines how measurement timing and missingness reshape inference, because a surveillance model is only as good as its representation of what is known and what remains unknown. The discussion then turns to sequential risk updating, decision thresholds under asymmetric costs, and performance evaluation under repeated alerting. The later sections address transportability, continual learning, and workflow integration. The underlying claim throughout is restrained but specific [7]. Preoperative prediction retains value, yet it should be nested within a broader longitudinal system in which risk is continually revised. For an event that evolves after surgery and reveals itself through changing physiology, postoperative surveillance is not an optional refinement. It is the more coherent primary formulation.

2. Postoperative Recovery as a Partially Observed Stochastic System

A surveillance-oriented formulation starts by distinguishing the event label from the physiologic process that gives rise to it. The recorded diagnosis of leak is not the complication itself. It is a time at which evidence becomes strong enough for clinical recognition, documentation, or intervention. Before that time the patient occupies an evolving latent state that may range from uncomplicated recovery to microscopic failure with limited systemic impact, then to overt inflammatory and hemodynamic disturbance. Any computational system seeking early recognition must therefore estimate not merely the eventual event, but the hidden state from which the event later becomes observable.

Let x_t denote a latent vector describing postoperative condition at time t. Its components may be interpreted as abstract coordinates for anastomotic integrity, inflammatory activation, perfusion deficit, organ stress, and intensity of supportive care. These coordinates need not be directly measurable. They serve as a compressed representation of the biologic and clinical forces generating the observed data stream. Let y_t denote the vector of observations available at time t, including laboratory values, vital signs, care setting, medication exposures, and other structured signals. The surveillance problem is then to estimate the conditional distribution of x_t and the conditional event probability over a future horizon, given all observations and interventions up to time t.

A minimal mathematical description uses a state transition equation and an observation equation. The latent state evolves according to intrinsic recovery dynamics, treatment effects, and

Table 1 Conceptual differences between preoperative prediction and postoperative surveillance

Aspect	Preoperative Models	Surveillance Systems
Timing	Before surgery	Continuous postoperative
Data structure	Static snapshot	Irregular time series
Target	Admission-level risk	Horizon-specific risk
Update mechanism	None	Sequential updating
Clinical role	Baseline stratification	Ongoing monitoring

Table 2 Latent state components and observable proxies

Latent Dimension	Interpretation	Example Observables
Integrity	Anastomotic stability	Drain output, imaging
Inflammation	Systemic response	WBC, CRP
Perfusion	Circulatory adequacy	Lactate, BP
Organ stress	Physiologic strain	Creatinine, oxygen need
Support intensity	Care escalation	ICU stay, vasopressors

random perturbations. Observed variables arise as noisy, incomplete projections of that state. The generic form is [8]

$$\begin{aligned} x_{t\Delta} &= f_\Delta x_t, a_t, u_t \ \eta_{t,\Delta} \\ y_t &= g_t x_t, m_t \ \epsilon_t \end{aligned} \tag{2.1}$$

where a_t denotes actions such as fluids, antibiotics, imaging, ICU transfer, or procedural escalation, u_t denotes exogenous context including operative factors and baseline patient characteristics, m_t denotes the measurement process governing which variables are observed and when, $\eta_{t,\Delta}$ is process noise, and ϵ_t is observation noise. The central point is that the data arriving after surgery are not a passive sample from a static condition. They are the result of a coupled system in which physiology changes, clinicians intervene, and measurement intensity adapts to suspicion.

This state-space perspective resolves a persistent ambiguity in postoperative prediction studies. When a model uses ICU admission, repeated laboratory ordering, or imaging occurrence as predictive features, is it learning leak biology or clinician concern? The answer is usually both. ICU transfer is partly a marker of severity and partly an institutional process variable. Repeated lactate measurement may indicate metabolic instability, but it also reflects the decision to monitor more closely. A surveillance framework does not eliminate this ambiguity, yet it makes it explicit. The observation process m_t becomes an object of modeling rather than an unexamined nuisance. That distinction is important for transportability because care process features can be highly predictive within one institution and poorly portable across settings with different monitoring norms.

A second advantage of the latent-state formulation is that it accommodates delayed observability. Suppose microscopic disruption begins at time $t^\dagger$, but the patient does not show convincing clinical evidence until $t^\star > t^\dagger$. The recorded event time may be closer to $t^\star$, while the hidden state begins deviating at $t^\dagger$. A surveillance model should ideally become increasingly concerned during the interval $t^\dagger, t^\star$. In classification terms, this interval is problematic because the patient is technically labeled positive for the admission but may not yet be distinguishable from uncomplicated recovery at early times. In state-space terms, this interval is precisely where the model should operate. The task is to identify growing posterior probability of a deteriorating latent state before formal recognition occurs.

To make this concrete, define the future leak event over horizon h as $E_{t,h} = \mathbf{1}\{T \leq t \ h\}$, where T is the time of clinically meaningful recognition. The quantity of interest is not the unconditional admission-level risk $\mathbb{P}T < \infty$, but the horizon-specific conditional probability

$$\pi_t h = \mathbb{P} E_{t,h} = 1 \mid \mathcal{H}_t,$$

where $\mathcal{H}_t$ denotes the complete history up to time t. This probability changes as $\mathcal{H}_t$ expands. It is perfectly coherent for a patient to begin with modest prior risk, move into a period of elevated short-horizon risk, then return toward lower risk after intervention or spontaneous stabilization [9]. Static models cannot express that temporal modulation except through repeated rerunning on ad hoc snapshots. Surveillance models treat it as their natural output.

The postoperative system is also nonstationary in a clinically meaningful sense. The same laboratory value can have different significance on different postoperative days. Moderate leukocytosis may be expected immediately after surgery but more concerning if it persists into a period when inflammatory markers should decline. Tachycardia on emergence from anesthesia

Table 3 Types of postoperative data and their roles in inference

Data Type	Characteristics	Informational Role
Vital signs	Frequent, noisy	Trend detection
Laboratories	Intermittent	Biochemical state
Interventions	Action-dependent	Severity proxy
Imaging	Sparse, specific	Confirmation signal
Missingness	Informative	Encodes suspicion

Table 4 Feature construction for temporal representation

Feature Class	Description	Example
Levels	Recent values	Current lactate
Slopes	Temporal trends	Rising WBC
Volatility	Variability	Fluctuating HR
Response	Post-action change	Lactate after fluids
Recency	Time since last obs	Hours since labs

differs from tachycardia several days later in a patient whose bowel function has not recovered and whose oxygen requirement is rising. This means the observation map g_t is time dependent. Clinical meaning is not solely in the value of y_t, but in the comparison between the current observation and the time-indexed expected recovery path. A robust surveillance model should therefore encode both absolute values and deviations from expected postoperative trajectories.

One way to express this formally is through residualization against stage-specific baselines. Let $\mu_k t \mid u$ denote the expected postoperative course for variable k given baseline and operative context u. Then define residualized observations

$$r_{k,t} = y_{k,t} - \mu_k t \mid u.$$

The state estimator can operate on $r_{k,t}$ rather than raw values, allowing the model to focus on abnormal deviation rather than normal postoperative physiology. This approach also clarifies why surveillance should not be reduced to simple threshold rules on single variables. What matters is not merely whether a marker is elevated, but whether it is elevated relative to the patient's expected stage-specific trajectory and whether the deviation is persistent, accelerating, or treatment resistant.

Latent-state modeling further allows explicit representation of uncertainty. In routine postoperative care uncertainty is not uniform. It is lower when measurements are frequent and concordant, and higher when data are sparse, noisy, or conflicting [10]. A model that emits only a point estimate of risk but conceals its uncertainty can mislead clinicians into treating weak evidence as strong evidence. State-space methods provide a natural mechanism for uncertainty propagation. If the posterior over x_t is broad because measurements are missing or because recent interventions have destabilized interpretation, then the predicted event probability should be accompanied by lower confidence or by a recommendation to gather additional information rather than to act definitively.

Another important consequence of this formulation is that actions become endogenous. Once a surveillance alert is triggered, clinicians change treatment and measurement behavior. The future trajectory is therefore conditioned on the very decisions the model influenced. This creates a feedback loop that makes retrospective performance harder to interpret. Nevertheless, the state-space framework accommodates this by treating actions a_t as part of the transition equation. The system is not only predicting an event. It is estimating how the patient may evolve under current management. Later sections will return to the decision-theoretic consequences of that coupling.

The surveillance problem thus differs from traditional preoperative prediction in three fundamental ways. First, the object of inference is time indexed and conditional on evolving history. Second, the data generating process includes adaptive observation and intervention. Third, the value of prediction depends on when and how the prediction changes, not only on whether the final ranking is correct. These differences are not minor technicalities. They determine the representation, the learning objective, and the evaluation scheme. A research program that aims to improve clinical leak detection should therefore begin from the partially observed stochastic system itself rather than from the convenience of a single-row dataset [11].

3. Measurement Timing, Informative Missingness, and Recovery Geometry

A surveillance system built on electronic health record data inherits a difficult property of clinical documentation: information is not merely incomplete, it is selectively observed. Some mea-

Table 5 Multi-horizon risk prediction and associated actions

Horizon	Risk Focus	Typical Action
6-hour	Immediate instability	Close monitoring
12-hour	Early deterioration	Repeat labs
24-hour	Short-term risk	Imaging
48-hour	Progressive decline	Intervention planning

Table 6 Action tiers and decision considerations

Action Tier	Description	Key Considerations
Routine	Standard care	Low cost, low sensitivity
Observation	Increased monitoring	Attention burden
Diagnostics	Imaging/tests	Cost vs benefit
Intervention	Surgery/drainage	High risk, high impact

surements are scheduled routinely, some are repeated because clinicians are worried, and some are absent precisely because the patient appears stable. Missingness is therefore not a random nuisance that can be erased by imputation without conceptual loss. It is part of the signal. Postoperative monitoring models that treat measurement absence as blank noise or that collapse trajectories into single imputed snapshots risk confusing low concern with low risk and high scrutiny with high biologic danger.

To examine this issue carefully, suppose there are p possible variables and that at time t_j only a subset is observed. Let $r_{j,k} \in \{0, 1\}$ indicate whether variable k is measured, and let $y_{j,k}$ be its value when available. A surveillance representation should include both the values and the observation pattern. The simplest reason is pragmatic. A lactate level unavailable for sixteen hours means something different from a lactate redrawn every few hours. Yet the deeper reason is epistemic. Observation itself communicates clinician suspicion, location of care, and perceived instability. A model that ignores $r_{j,k}$ discards information about the institutional process that generated the data.

This idea can be formalized by assigning an intensity model to observation times. Let $N_k t$ be the counting process for variable k. The probability that the variable is measured in the next short interval can depend on the hidden state and on recent actions:

$$\begin{aligned}\lambda_k^{\mathrm{obs}} t = \exp\big(&\alpha_k \;\, \beta_k^\top x_t \\ &\gamma_k^\top a_t \;\, \delta_k^\top u\big).\end{aligned} \tag{3.1}$$

The coefficients β_k encode how strongly true physiologic severity increases monitoring, while γ_k and δ_k encode the influence of treatment status and baseline context. Jointly modeling $\lambda_k^{\mathrm{obs}} t$ with the latent clinical state allows the surveillance system to distinguish between deterioration-driven measurement and protocol-driven measurement. That separation is never perfect, but even an approximate model reduces the risk that the algorithm simply learns to imitate local ordering behavior.

Representation of postoperative recovery also requires geometry [12]. The raw observation space is high dimensional, heterogeneous in scale, and difficult to interpret directly. Yet much of postoperative physiology lies on lower-dimensional manifolds corresponding to meaningful clinical processes. Inflammation, perfusion deficit, organ stress, respiratory compromise, nutritional reserve, and support intensity are not single variables, but coordinated directions in data space. A useful surveillance system therefore constructs latent coordinates aligned with those directions rather than operating only on individual laboratory values in isolation.

Let $z_t \in \mathbb{R}^d$ be a lower-dimensional embedding of the patient's history up to time t. One can obtain z_t through a structured encoder that ingests recent levels, slopes, variability measures, and observation masks. A linearized expression is

$$\begin{aligned} z_t &= W_b b \;\, W_o o \;\, W_h \phi \mathcal{H}_t \\ \phi \mathcal{H}_t &= \big[\ell_t, s_t, v_t, q_t, r_t\big], \end{aligned} \tag{3.2}$$

where b contains baseline patient factors, o contains operative context, ℓ_t contains recent local levels of measured variables, s_t contains temporal slopes, v_t contains volatility or inconsistency metrics, q_t contains treatment-response summaries, and r_t contains observation-process features such as recency and count of measurements. Although nonlinear encoders are often useful, the structural decomposition itself matters because it reflects distinct kinds of evidence. A persistently abnormal level is different from a rapidly rising trend. Either one is different again from a variable that has not been checked in an unusually long interval.

One practical challenge is that postoperative measurements are asynchronous. Vital signs may be available every few hours, arterial blood gas variables only in higher-acuity settings, and

Table 7 Evaluation dimensions for surveillance models

Metric	Purpose	Limitation
Dynamic AUC	Rank discrimination	Ignores timing
Calibration	Probability accuracy	Horizon-specific
Lead time	Early detection	Needs adjudication
Alert precision	Alarm quality	Trade-off sensitive
Alert burden	Workflow impact	Context dependent

Table 8 Sources of bias and confounding in EHR-based surveillance

Source	Mechanism	Impact
Measurement bias	Selective testing	Inflated associations
Action confounding	Clinician response	Feedback loops
Site variation	Practice differences	Poor transportability
Label delay	Late documentation	Distorted timing
Missingness	Nonrandom absence	Misleading signals

chemistry panels once or twice daily unless concern rises. A naive approach that bins all measurements onto a fixed clock and forward-fills aggressively can create false continuity. A better approach constructs time-decayed sufficient summaries that preserve recency and uncertainty. For a variable k, define a recency-weighted mean and slope at surveillance time t:

$$\begin{aligned} \mu_k t &= \frac{{}_{j:t_j \leq t}\, w_j t r_{j,k} y_{j,k}}{{}_{j:t_j \leq t}\, w_j t r_{j,k}} \\ \mu_k t &= \frac{{}_{j:t_j \leq t}\, w_j t r_{j,k} t_j - \bar{t}_k t y_{j,k}}{{}_{j:t_j \leq t}\, w_j t r_{j,k} t_j - \bar{t}_k t^2}, \end{aligned} \tag{3.3}$$

with $w_j t = \exp\{-\lambda_k t - t_j\}$ and $\bar{t}_k t$ the weighted time centroid. This representation preserves the intuition that a recent value counts more than a remote one and that a worsening trend should matter even if the current absolute level remains within a nominal range.

The recovery geometry should also incorporate stage-specific expectations. The postoperative period is not homogeneous [13]. The first hours after surgery are characterized by fluid shifts, analgesic effects, transient hemodynamic perturbations, and inflammatory response. Later days should show stabilization or improvement in uncomplicated recovery. Consequently the same direction in observed data space may mean different things at different times. One way to incorporate this is to define time-varying normal trajectories conditioned on operative context and then represent the patient by residual deviation from those trajectories. For variable k,

$$\tilde{y}_{k,t} = y_{k,t} - \mathbb{E} y_{k,t} \mid b, o, t,$$

so that the latent encoder operates on abnormality relative to expected postoperative recovery rather than on raw magnitude alone. This helps reduce the tendency of models to overreact to normal immediate postoperative perturbations and underreact to persistent late abnormalities.

Another underused source of information is response to treatment. Postoperative surveillance is not only about the values observed before clinicians act, but also about what happens after action. A patient whose lactate normalizes after modest resuscitation differs materially from a patient whose lactate remains elevated despite fluids and support. Similarly, transient tachycardia relieved by analgesia carries a different implication from tachycardia that persists across interventions. These distinctions can be encoded through action-conditioned deltas. If action a occurs at time t_a, then for variable k,

$$\begin{aligned} \Delta_k^a &= y_{k,t_a\delta} - y_{k,t_a^-} \\ \rho_k^a &= \frac{y_{k,t_a\delta} - y_{k,t_a^-}}{\delta}, \end{aligned} \tag{3.4}$$

where t_a^- denotes the latest time before action and δ is a clinically chosen response window. Such features are often more informative than pre-action values because they reveal the system's dynamic resistance or responsiveness.

The geometry of surveillance should further respect heterogeneity across procedure types and care pathways. A low anterior resection with pelvic dissection, a colectomy with proximal diversion, and an upper gastrointestinal anastomosis do not occupy the same baseline risk landscape or the same recovery manifold [14]. Even within a single broad category, emergency surgery differs from elective surgery, and ICU recovery differs from ward recovery. Rather than fitting a completely separate model for every context, one can encode context-specific affine shifts or gating functions so that the latent space is shared but

Table 9 Strategies for transportability and maintenance

Strategy	Goal	Approach
Feature separation	Reduce bias	Bio vs process split
External validation	Generalization	Multi-site testing
Drift monitoring	Stability	Distribution tracking
Recalibration	Maintain accuracy	Update probabilities
Audit logging	Feedback analysis	Track actions/outcomes

warped according to operative and pathway variables. This preserves sample efficiency while allowing the meaning of observed trajectories to vary rationally across contexts.

The question of imputation remains unavoidable. Some variables must be carried forward or interpolated to compute continuous-time state updates. Yet imputation in surveillance should be interpreted as uncertainty management, not as completion of a fully known dataset. A carried-forward value does not mean the physiology remained unchanged; it means the last observed value is the best available anchor, subject to growing uncertainty as time passes. This idea can be represented by inflating variance with elapsed time. If the imputed estimate of variable k at time t is $\hat{y}_{k,t}$, then its uncertainty may scale as

$$\mathrm{Var}\hat{y}_{k,t} = \sigma_k^2\big(1\ \omega_k \Delta t\big),$$

where Δt is time since last observation. A state estimator can then treat stale information as less authoritative without discarding it entirely.

These considerations collectively define a richer notion of postoperative recovery geometry. The patient is not a point in a static feature space. The patient is a path through a time-varying latent landscape, observed along irregular coordinates and subject to action-dependent perturbations. Surveillance quality depends on representing that path faithfully enough that the system can detect meaningful departures from expected recovery before they are clinically obvious. Much of the later modeling and evaluation burden follows from this representational commitment.

4. Sequential Inference, Dynamic Hazard Updating, and Multi-Horizon Prediction

Once a latent representation of postoperative recovery has been defined, the next task is to update belief about future leak in real time. This requires a bridge between state estimation and event prediction. Static classifiers can be rerun repeatedly on refreshed features, but that strategy lacks an explicit model of temporal dependence and offers little guidance on how to combine old and new information coherently [15]. Sequential inference provides a principled alternative. It treats each new observation as modifying a current posterior over latent state, which then implies a revised distribution over future event times.

Let $px_t \mid \mathcal{H}_t$ denote the posterior distribution over latent state after assimilating data up to time t. The arrival of a new observation $y_{t\Delta}$ changes this posterior through prediction and correction. In generic form,

$$\begin{aligned} px_{t\Delta} \mid \mathcal{H}_t &= px_{t\Delta} \mid x_t, a_t, u\, px_t \mid \mathcal{H}_t\, dx_t \\ px_{t\Delta} \mid \mathcal{H}_{t\Delta} &\propto py_{t\Delta} \mid x_{t\Delta}, m_{t\Delta}\, px_{t\Delta} \mid \mathcal{H}_t. \end{aligned} \tag{4.1}$$

This pair of equations expresses the basic surveillance logic. Before the next measurement arrives, the model projects the patient forward under current management and uncertainty. When the measurement appears, the model corrects its belief according to how compatible the observation is with each possible latent state. The output that clinicians see is not the hidden state itself, but a function of that posterior, typically a short-horizon leak probability.

A natural event model is a discrete-time hazard over clinically meaningful intervals. Suppose the postoperative course is partitioned into intervals $I_m = t_m, t_{m1}$. Conditional on survival without recognized leak up to t_m, define the hazard

$$\begin{aligned} h_m &= \mathbb{P}T \in I_m \mid T > t_m, \mathcal{H}_{t_m} \\ &= \sigma\big(\alpha_m\ \theta^\top \bar{x}_{t_m}\ \bar{x}_{t_m}^\top Q \bar{x}_{t_m}\big), \end{aligned} \tag{4.2}$$

where $\bar{x}_{t_m}$ is the posterior mean or another summary of $px_{t_m} \mid \mathcal{H}_{t_m}$, Q captures interactions, and α_m is a baseline interval effect that accommodates changing background risk across postoperative time. The horizon-specific risk over the next H intervals then follows from the product of survival complements:

$$\pi_{t_m} H = 1 - \prod_{\ell=m}^{mH-1} 1 - h_\ell.$$

This expression highlights why horizon matters. A model tuned for immediate deterioration may underperform for broader horizons, while a model optimized for 72-hour discrimination may blur clinically important short-term escalation. Surveillance should therefore estimate several horizons simultaneously, for example 6-hour, 12-hour, and 24-hour risks, because those horizons correspond to different actions.

The latent state itself may evolve in multiple regimes. Uncomplicated recovery, delayed normalization, systemic inflammatory escalation, post-intervention stabilization, and overt sepsis are not merely different points on a single continuous axis. They can function as regimes with qualitatively different dynamics. A switching-state model represents this by introducing a discrete mode variable s_t [16]. The latent evolution then depends on both x_t and s_t, and the mode itself changes over time. In simple form,

$$x_{t\Delta} = A_{s_t} x_t \;\; B_{s_t} a_t \;\; \eta_{t,\Delta}$$
$$\mathbb{P} s_{t\Delta} = j \mid s_t = i, x_t, a_t = \Gamma_{ij} x_t, a_t. \tag{4.3}$$

The event hazard may then depend on both the continuous state and the discrete regime. This is clinically sensible because the same level of laboratory abnormality can mean different things in different modes. A modestly elevated white blood cell count in a steadily improving patient does not imply the same risk as the identical count in a patient whose support needs have recently increased and whose acid-base profile is worsening.

Sequential surveillance also benefits from explicit smoothing in time. Real physiologic trajectories are noisy, but clinical concern should not oscillate wildly in response to small fluctuations when the broader trajectory is unchanged. A learning objective can encourage temporal coherence without suppressing genuine abrupt changes. Let $\hat{\pi}^h_{t_m}$ denote the predicted risk for horizon h. A regularized loss may combine survival likelihood with a smoothness penalty:

$$\mathcal{L} = - \sum_{i} \sum_{m} \ell^{\text{surv}}_{i,m}$$
$$\lambda \sum_{i} \sum_{m} \left(\hat{\pi}^h_{i,t_{m1}} - \hat{\pi}^h_{i,t_m}\right)^2, \tag{4.4}$$

where $\ell^{\text{surv}}_{i,m}$ is the appropriate interval log-likelihood. The penalty discourages unstable jitter while allowing sustained rises or falls in risk. In practice one may refine this further by making the penalty weaker when a genuinely informative new observation arrives, so that the model remains sensitive to abrupt deteriorations.

Multi-horizon prediction introduces a calibration challenge. A risk estimate that is well calibrated for 24 hours may still be miscalibrated for 6 hours. Since different actions attach to different horizons, miscalibration cannot be ignored. One approach is to learn a shared latent state encoder and separate horizon-specific heads, each calibrated by isotonic or spline-based postprocessing on held-out data. Another is to model the full conditional event-time distribution and derive horizon probabilities by integration. The latter is more coherent but can be harder to estimate robustly in small samples [17]. The choice is less important than the principle: surveillance output should preserve meaningful probabilistic interpretation at the horizons where actions occur.

Because postoperative events are relatively uncommon, repeated interval-level training can generate extreme class imbalance. Yet not all negative intervals are equally informative. Intervals far from any event during smooth recovery contribute less to early detection than intervals that occur near the transition from latent deterioration to overt clinical recognition. A useful weighting scheme emphasizes imminence. If $d_{i,m}$ denotes time from interval m to event for case i, then intervals can receive weights $w_{i,m} = \exp{-\kappa d_{i,m}}$ before the event and lower weights for remote benign intervals. This concentrates learning on the portion of the trajectory where early warning is both possible and clinically valuable.

An additional complication is intervention confounding. Suppose clinicians order imaging when subtle signs already raise concern. The event is then recognized sooner than it would have been without imaging. Retrospective models trained on such data may learn that imaging occurrence precedes leak and may therefore use imaging as a powerful but partly circular predictor. In a surveillance framework this is hazardous because it ties prediction to local behavior rather than to underlying physiology. A cleaner formulation separates two targets: latent leak state and recognized leak event. Imaging may legitimately increase the probability that the event will be formally recognized soon, but it does not by itself create the latent failure. Joint models that distinguish these targets can better support early biological surveillance rather than merely echoing current clinical suspicion.

One way to formalize this distinction is with competing hazards. Let $T^\star$ be the time of latent state transition into clinically meaningful failure and T the time of formal recognition. Then $T^\star \leq T$, and recognition depends on both latent state and observation intensity. The hazard for T can be decomposed conceptually into a biological component and an ascertainment component [18]. Although full identification is difficult, even a partial decomposition helps explain why surveillance models need careful interpretation. A detector that becomes accurate only after observation intensity has already increased may still be useful as a confirmer, but it is different from a detector that senses biological change earlier.

At the algorithmic level, several architectures can implement the sequential inference program. Recurrent neural networks with time gaps, neural controlled differential equations, latent ordinary differential equations, hidden Markov models, particle filters, and generalized additive state-space models are all viable. The choice should be guided by event count, need for interpretability, computational constraints, and calibration requirements. In many surgical datasets the sample size and event rate do not support extremely high-capacity models without strong regularization. A hybrid architecture often provides the best compromise: a transparent temporal feature encoder feeding a moderately flexible hazard model with explicit calibration. This retains the advantages of nonlinear interaction modeling without fully obscuring the pathway from data to prediction.

The essential point is not the superiority of one architecture. It is that postoperative leak detection becomes conceptually cleaner when framed as sequential belief updating over latent recovery state. That framing makes explicit why the model's job is ongoing rather than one-time, why horizon-specific probabilities matter, and why prediction quality must be assessed according to temporal behavior as well as final discrimination. In the surveillance context, a model that updates coherently, remains calibrated across horizons, and provides usable lead time is more valuable than one that merely achieves a slightly better admission-level AUC.

5. Action-Aware Alert Design and Sequential Decision Rules

A surveillance system that only estimates risk and leaves thresholding unspecified is incomplete. The clinical value of prediction emerges only when probability is mapped to action, and postoperative actions differ substantially in burden. Intensifying bedside monitoring has a low direct cost but consumes attention. Ordering a contrast-enhanced CT scan creates logistical burden, radiation exposure, contrast risk, and financial cost. Initiating invasive drainage or returning to the operating room carries much greater consequences [19]. It follows that there is no single universally correct alarm threshold. The appropriate threshold depends on which action the alert is supposed to trigger and on how the institution values missed detections relative to unnecessary workup.

Let the action set be $\mathcal{A} = \{a_0, a_1, a_2, a_3\}$, corresponding respectively to routine follow-up, intensified observation, diagnostic escalation, and definitive invasive escalation. Suppose that at time t the model outputs one or more horizon-specific probabilities $\pi_t h$. A rational decision rule should maximize expected utility rather than applying a universal classification threshold. If Ua, π_t, z_t denotes the utility of action a given estimated risk π_t and latent state summary z_t, then the recommended action is

$$a_t^\star = \arg\max_{a\in\mathcal{A}} Ua, \pi_t, z_t.$$

To make this operational, utility can be decomposed into direct action cost, false-negative loss, false-positive loss, and the informational value of further observation. A simple expression is

$$\begin{aligned} Ua, \pi_t, z_t = &-c_a z_t \\ &- \pi_t\, \ell_a^{\mathrm{miss}} z_t \\ &- 1 - \pi_t\, \ell_a^{\mathrm{fp}} z_t \;\; v_a^{\mathrm{info}} z_t. \end{aligned} \tag{5.1}$$

The term c_a is the immediate burden of action. The term ℓ_a^{miss} quantifies harm if leak is present but action a is insufficiently aggressive. The term ℓ_a^{fp} quantifies harm if leak is absent but action a is unnecessarily intense. The term v_a^{info} represents the value of information generated by the action, such as the benefit of repeated examination or imaging in reducing posterior uncertainty. This last term is what makes surveillance fundamentally sequential. Some actions are worth taking not because they directly treat the leak, but because they improve the future information state.

Under this framework different thresholds emerge naturally. The probability needed to justify intensified nursing surveillance is lower than the probability needed to justify CT imaging, and lower again than the probability needed to support operative re-exploration. Fixed-threshold classification hides this fact and therefore obscures the true implementation problem. A surveillance system should instead expose multiple action bands. Low but rising risk may prompt closer observation and repeat measurement. Moderate risk may prompt diagnostic testing. High risk in an already unstable patient may justify immediate invasive escalation. The exact numbers depend on institutional capacity and clinical philosophy, but the principle is general [20].

Repeated evaluation also raises the problem of alert fatigue. If probabilities are updated every hour, then naive thresholding may produce clusters of nearly identical notifications. Such behavior reduces trust and overwhelms clinicians with redundant messages. The decision layer should therefore incorporate hysteresis, novelty, and cooldown logic. Let $\hat{\pi}_t$ be the relevant predicted probability, $\Delta\hat{\pi}_t = \hat{\pi}_t - \hat{\pi}_{t-\delta}$ its recent change, and t_{last} the time of the most recent alert. An action-triggering notification can be governed by

$$\begin{aligned} A_t = \mathbf{1}\{\hat{\pi}_t > \tau_a, \quad &\Delta\hat{\pi}_t > \kappa_a, \\ t - t_{\mathrm{last}} > \rho_a, \quad &C_t < c_a^{\max}\}, \end{aligned} \tag{5.2}$$

where τ_a is the threshold for action a, κ_a is a minimum novelty requirement, ρ_a is a cooldown interval, and C_t is a recent cumulative alert burden constrained by $c_a^{\max}$. This logic ensures that alerts represent new clinically relevant information rather than repeated restatements of an unchanged risk state.

Action-aware thresholding should also account for postoperative stage. Early after surgery, mild deviations are common and often nonspecific. Later in the course, the tolerance for persistent abnormality should decrease because uncomplicated recovery ought to show normalization. A time-varying threshold can encode this. For action tier a,

$$\tau_a t = \tau_{a,0}\;\; \tau_{a,1} e^{-\omega_a t}\;\; \tau_{a,2}\mathbf{1}\{t > t_a^\dagger\}.$$

Here the effective threshold may start higher during the immediate postoperative window and then fall as the expected range narrows. Such adaptive thresholds align more closely with clinical reasoning than a single cutoff applied at all times.

A more sophisticated view casts alert design as a partially observed control problem. The state x_t is hidden, the posterior $px_t \mid \mathcal{H}_t$ summarizes current knowledge, and the chosen action affects both the patient and the future data stream. The optimal strategy can be written through a Bellman recursion on the belief state b_t, where b_t denotes the current posterior:

$$\begin{aligned} Vb_t = \max_{a\in\mathcal{A}} \{&Rb_t, a \\ &\mathbb{E}\big[Vb_{t\Delta} \mid b_t, a\big]\}. \end{aligned} \tag{5.3}$$

Although solving the full partially observed Markov decision process exactly is usually unrealistic in clinical deployment, the formulation is still valuable conceptually. It clarifies that alerts should be judged not only by immediate classification accuracy but by the downstream trajectory they induce [21]. An intermediate action such as repeat laboratories or expedited examination may be optimal because it reduces uncertainty enough to improve later decisions. The surveillance problem is therefore closer to staged control under uncertainty than to ordinary binary classification.

Competing postoperative explanations must also be considered. Fever, tachycardia, ileus, leukocytosis, and mild acidosis can arise from several causes other than leak. A detector that estimates only leak versus no leak may overreact to broadly nonspecific postoperative stress. One solution is to model several latent complication states simultaneously, for example uncomplicated recovery, pulmonary complication, nonspecific inflammatory response, delayed bowel recovery, and suspected leak. The decision rule can then compare posterior mass across these states rather than focusing on leak in isolation. If the data are more consistent with a pulmonary process than with leak, the appropriate response differs even if overall risk of deterioration is rising. This multi-state view is especially useful when surveillance is intended to guide diagnostic testing rather than merely to emit a binary alarm.

An additional design principle is abstention under severe uncertainty. There are periods during which the model may lack enough information to support a confident recommendation. This can happen when measurement density is unusually sparse, when recent interventions make trajectories hard to interpret, or when the patient's profile lies outside the training distribution. Instead of forcing a probability that will be treated as precise, the system may issue an uncertainty flag recommending data acquisition rather than a direct clinical action. Formally, if posterior entropy or predictive variance exceeds a threshold, the recommendation may move from treatment space into information-gathering space. This is often safer than overconfident guessing.

Finally, action-aware design requires attention to human factors. Clinicians do not follow probabilities mechanically. They integrate them with examination, operative memory, drain character, radiology interpretation, and contextual details that may not be present in the structured data [22]. For this reason an alert should not be framed as a command. It should be framed as an updated estimate with a transparent rationale and a suggested next-best action tier. The system is most likely to support practice when it sharpens attention at the right moment rather than attempting to replace clinical judgment altogether.

In sum, postoperative surveillance requires more than accurate prediction. It requires disciplined translation of prediction into staged responses under asymmetric loss, temporal dependence, and constrained attention. Models that ignore this layer may perform well in retrospective tables yet fail in real units because their alerts are either too frequent, too late, too opaque, or attached to the wrong action. Decision-aware design therefore belongs at the core of the research agenda rather than at the margin.

6. Evaluation, Lead Time, and the Limits of Static Performance Metrics

A surveillance system should be evaluated according to the temporal and operational questions it is meant to answer. Conventional metrics such as area under the receiver operating characteristic curve remain informative, but they do not identify whether the system produces usable warning before overt recognition, whether the probability estimates can support action-tiered thresholds, or whether alert burden is tolerable in real workflows. A broader evaluation framework is therefore necessary, one that emphasizes the sequential character of the task.

The first evaluation dimension is horizon-specific discrimination. At any surveillance time t, the model estimates the probability of leak within a future horizon h. Two patients with the same admission-level outcome may differ radically in near-term risk at a given time point. Time-dependent discrimination therefore compares patients in the same risk set, meaning those who remain event free and under observation at time t. Dynamic AUC or concordance measures conditioned on the risk set are more meaningful than a single admission-level AUC because they ask whether the model ranks imminent deterioration appropriately among patients currently being monitored.

The second dimension is calibration over time and across horizons. A surveillance probability must mean what it says [23]. If the model reports 0.30 risk over the next 12 hours, clinicians must be able to interpret that as an approximate 30% chance in similar circumstances. Miscalibration is particularly problematic in surveillance because thresholding strategies often assume probability semantics rather than mere rank ordering. Calibration should therefore be assessed separately for each horizon and at multiple postoperative stages. Formally, for horizon h and prediction bins B_b,

$$\mathrm{Cal}h = {}_{b}\,\frac{n_b}{N_h}\left(\bar{y}_{b,h} - \bar{\pi}_{b,h}\right)^2, \tag{6.1}$$

where $\bar{y}_{b,h}$ is empirical frequency of event within horizon h among predictions in bin b, and $\bar{\pi}_{b,h}$ is the mean predicted probability in that bin. Because monitoring density and care setting affect both data availability and event prevalence, calibration should also be reported after stratifying by observation intensity and postoperative day.

A third and distinctly surveillance-oriented dimension is lead time. A true positive alert delivered after the patient is already clinically obvious offers little additional value. Let A_i denote the first actionable alert time for patient i, and T_i the clinically adjudicated recognition time. Then patient-level lead time is $L_i = T_i - A_i$ whenever $A_i < T_i$. The distribution of L_i among detected events is more informative than its mean alone. One should report the fraction of events detected at least 6 hours, 12 hours, or 24 hours before recognition, since these windows correspond to realistic opportunities for intervention. Systems with similar discrimination may differ sharply in this distribution, and the one with earlier usable warning is often more valuable even if its pointwise metrics are only modestly better.

Repeated prediction creates another evaluative distinction between event sensitivity and alert precision. Event sensitivity asks what proportion of leak cases receive at least one actionable alert before recognition. Alert precision asks what proportion of emitted alerts are followed by a leak within the specified horizon. Both matter because a system can achieve high event sensitivity by issuing many alarms, thereby overwhelming clinicians with false positives. Conversely a system can achieve high alert precision by alerting only on very late, obvious cases, thereby missing opportunities for early rescue [24]. A balanced assessment must therefore report both event-level and alert-level performance.

Alert burden deserves explicit quantification rather than informal discussion. Useful measures include alerts per patient-day, median time between alerts, proportion of patients receiving at least one false-positive alert, and number of alerts required per detected event. These quantities represent the cost in clinician attention. In many operational settings attention is the scarcest resource, and a small gain in discrimination may not justify a large increase in burden. Surveillance evaluation that omits these measures is incomplete because it cannot answer whether the detector can coexist with routine care.

The retrospective nature of most available datasets introduces further complications. Outcomes are observed under the care that clinicians actually provided, not under the care that would have occurred had the model been available. This means that strong retrospective performance may arise partly from patterns of clinician suspicion already embedded in the data. For example, if imaging was ordered promptly when subtle signs appeared, then the recorded recognition time may be earlier than it would have been without those decisions. A model may appear to offer little lead time not because the signals were absent, but because clinicians were already acting optimally. The reverse problem also occurs. If the model exploits observation-process artifacts such as repeated laboratory ordering or ICU transfer, it may look highly accurate retrospectively while providing little independent value prospectively. These issues do not make retrospective evaluation useless, but they require humility in interpretation.

Several design choices can improve retrospective validity. Landmark analysis is one. At predefined postoperative times the available information is frozen and future risk is predicted from that point, thereby reducing leakage from late information. Another is to stratify performance by whether diagnostic escalation had already occurred at the time of prediction [25]. If the model becomes strong only after imaging is ordered or ICU transfer occurs, then it may be serving more as a confirmer than as an early detector. A silent-run prospective study offers an even better intermediate step. In such a study the model updates in real time but its alerts are hidden from clinicians, allowing realistic measurement of calibration, alert burden, and potential lead time under current practice without changing care.

Ground truth timing also requires special attention. Administrative codes or discharge labels are inadequate for surveillance because they often lag behind the true onset of clinical suspicion and can distort lead-time analysis. Ideally the event time should be adjudicated from chart review as the earliest time at which the evidence would reasonably support leak-oriented escalation. This standard is harder to obtain but far more relevant. Without it a system may seem to provide generous lead time simply because the recorded event time reflects late documentation rather than clinical reality.

External validation is indispensable because surveillance models are especially vulnerable to local process artifacts. Different hospitals vary in laboratory cadence, ICU admission thresholds, postoperative pathways, documentation norms, and thresholds for imaging. These factors shape both the observation process and the relationship between observed data and latent leak state. A model that implicitly depends on frequent early laboratory draws may degrade in a setting where such tests are ordered more selectively. External validation should therefore report not only discrimination and calibration but also comparative summaries of measurement density, missingness patterns, and care setting distribution. Without that context, performance differences across sites are hard to interpret.

Subgroup analysis is equally important. Procedure type, diversion status, emergency status, age, sex, comorbidity burden, race, care unit, and baseline nutritional profile can influence both true risk and observability. A surveillance system should be assessed for disparity not only in AUC but in lead time, calibration, and alert burden across these groups [26]. A model that detects leaks equally well overall may still be inequitable if one subgroup receives systematically later alerts because it is monitored less intensively or because missingness is interpreted differently.

One additional metric deserves emphasis: decision utility. Because the ultimate purpose of surveillance is action, model evaluation can be tied directly to net benefit under staged thresholds. If an alert for imaging is issued whenever $\pi_t h > \tau$, then one can estimate the clinical utility of that policy relative to alternatives by weighting true positives and false positives according to the costs attached to imaging and missed detection. Decision curves adapted for repeated time-dependent predictions are methodologically demanding, but they provide the closest approximation to the real question clinicians ask: does this surveillance policy improve the trade-off between catching leaks and overworking patients and staff?

The broader message is that static discrimination is at best a partial summary of surveillance quality. A postoperative detector should be judged on when it becomes correct, whether its probabilities remain interpretable across horizons, how much attention it consumes, whether it generalizes across institutions, and whether it improves the action trade-offs relevant to postoperative care. Evaluation that ignores these dimensions risks selecting models that are statistically impressive but clinically mismatched.

7. Transportability, Continual Learning, and Institutional Integration

Even a well-evaluated surveillance model exists within a living clinical system that changes over time. Monitoring practices shift, postoperative pathways are redesigned, laboratory technologies evolve, case mix drifts, and clinician behavior adapts to the presence of the tool itself. Transportability and maintenance are therefore not secondary issues. They determine whether a model remains useful after the moment of publication or local pilot deployment.

Transportability begins with the observation that postoperative surveillance is partly about biology and partly about institution. Laboratory values, vital signs, and broad indicators of organ stress have cross-site physiologic meaning, but the timing and frequency with which they are collected vary. ICU admission is one of the clearest examples. In one hospital it may signal substantial instability; in another it may be routine for certain operations. A model that leans heavily on such variables can perform well internally while failing elsewhere. One solution is explicit decomposition of the predictor space into biologic and process channels [27]. Let $z_t = z_t^{\mathrm{bio}}, z_t^{\mathrm{proc}}$, where the first block reflects physiologic quantities and the second reflects observation intensity or care pathway markers. During development the model can be penalized to avoid excessive dependence on z_t^{proc} unless those variables add clear incremental value after accounting for physiology.

This idea can be encoded with a regularized loss. If $f z_t^{\mathrm{bio}}, z_t^{\mathrm{proc}}$ is the risk model, then one may train by minimizing

$$\begin{aligned} \mathcal{J} = \mathcal{L}_{\mathrm{surv}} \; \lambda_1 \|\theta_{\mathrm{proc}}\|_2^2 \\ \lambda_2 \big(\mathbb{E} f z^{\mathrm{bio}}, z^{\mathrm{proc}} - \mathbb{E} f z^{\mathrm{bio}}, 0\big)^2, \end{aligned} \tag{7.1}$$

where θ_{proc} denotes parameters associated with process features. The second penalty discourages large shifts in average prediction when process variables are ablated, thereby promoting reliance on biology where possible. Such regularization does not guarantee transportability, but it makes the model's dependencies more inspectable.

Continual learning presents a separate challenge. Suppose the surveillance system is deployed and clinicians begin to respond to its alerts. Imaging may be ordered earlier, repeat laboratories drawn more frequently, and recognition times shortened. The postdeployment data distribution is therefore altered by the model itself. If retraining is performed naively on these data, the algorithm may start learning a world in which its own interventions are already embedded. This is a familiar feedback problem in decision-support systems, but it is especially acute in surveillance because the observation process is central to the model. Responsible updating thus requires logging of model scores, displayed alerts, clinician overrides, subsequent actions, and outcomes. Without that audit trail it becomes impossible to distinguish true improvement from self-reinforcing artifact.

Drift detection should occur at multiple levels. Input drift concerns changes in marginal distributions of observed variables and missingness. Representation drift concerns changes in the latent state distribution learned by the encoder. Calibration drift concerns divergence between predicted and observed event frequencies. Policy drift concerns changes in clinician response to alerts. A practical monitoring architecture might track all four [28]. If P_t denotes the distribution of latent states during deployment window t and P_0 the development reference, a divergence statistic such as

$$D_t = \mathrm{KL} P_t \parallel P_0$$

can serve as one representation-level signal, while rolling calibration plots can track probability reliability. Large discrepancies need not imply immediate failure, but they should trigger review, possible recalibration, and investigation into changing workflows.

Institutional integration requires careful choice of where the surveillance model lives in the clinical process. If it is buried in a dashboard that clinicians seldom check, even a strong model will have little effect. If it emits too many pager-level alerts, it will rapidly lose credibility. The most promising placement is usually at structured reassessment points: postoperative handoff, morning review, change in care setting, arrival of new laboratory panels, or nursing escalation pathways. At those moments the algorithm can augment an already active decision cycle rather than trying to interrupt one.

The interface should show more than a single probability. Because surveillance is inherently temporal, the output should display recent risk trajectory, the main drivers of change, and uncertainty. A rising risk curve accompanied by persistent leukocytosis, unresolved metabolic disturbance, and escalating support is easier to interpret than an isolated number. Likewise a moderate probability with high uncertainty may appropriately prompt data acquisition rather than aggressive treatment. Interpretability in this context means preserving a clinically legible explanation of why risk changed, not simply listing globally important variables from the training phase.

Fairness in institutional integration extends beyond traditional subgroup performance. Different patient groups may receive different monitoring intensity, be admitted to different care settings, or experience different thresholds for escalation. A surveillance system can inadvertently intensify these disparities. For example, a group with sparser laboratory data may receive later alerts not because the model is biased in the usual coefficient sense, but because the information environment is poorer. Fairness analysis should therefore examine whether the model's benefit, in terms of lead time or avoided delay, is distributed unevenly across groups [29]. If so, institutional remedies may include standardized postoperative monitoring protocols rather than only algorithmic adjustment.

Another integration issue is the boundary between recommendation and autonomy. Clinicians may disagree with model output because of physical examination findings, operative details, or qualitative cues not present in the structured data. The system should support override and encourage documentation of reasons. These override patterns are valuable. Persistent suppression of certain alerts without adverse outcomes may reveal oversensitivity in a specific context. Frequent clinician action before the model alerts may reveal missing features or insufficient update cadence. Deployment should thus be treated as an iterative observation period in which human and algorithmic reasoning are compared, not as a final proof of correctness.

Continual calibration is likely to be more feasible than continual full retraining in many sites. Because postoperative pathways and event prevalence can shift gradually, recalibrating the mapping from latent score to horizon probability may restore useful probability semantics without changing the underlying representation. Full retraining should be undertaken more cautiously, ideally after enough audited postdeployment data have accumulated to support causal interpretation of model-induced workflow changes.

The transportability problem also suggests a staged research agenda. First, models should be designed with explicit separation of physiologic and process channels. Second, they should undergo multi-site retrospective validation with detailed reporting of monitoring differences. Third, they should enter silent-run prospective testing. Fourth, they should be introduced with action-tiered thresholds in selected units while collecting override and utilization data. Only after those steps should claims about improved outcomes be made [30]. This sequence is slower than the conventional predictive modeling cycle, but it is more aligned with the realities of surveillance, where feedback loops and local workflow are inseparable from performance.

The overall implication is that postoperative leak surveillance is not a product that can be trained once and exported unchanged. It is a maintained component of institutional decision infrastructure. Success depends on a stable relationship between physiology, measurement, and action. Because all three evolve, the surveillance system must be designed for adaptation from the start.

8. Conclusion

Postoperative surveillance offers a more coherent primary framing for anastomotic leak detection than purely preoperative prediction because the complication reveals itself through an evolving recovery trajectory rather than through a single static condition present before incision. The most informative evidence often arises after surgery, is observed irregularly, and acquires meaning only through timing, persistence, co-occurrence, and response to intervention. A useful computational system must therefore do more than assign an admission-level risk score. It must maintain a dynamic estimate of latent postoperative state, update short-horizon event probabilities as new information arrives, and translate those probabilities into staged actions that reflect the real costs of false alarms, delayed recognition, and clinician attention.

This paper has argued for a framework built around partially observed stochastic dynamics, explicit treatment of informative missingness, multi-horizon risk estimation, and action-aware alert policies. Under that framework, preoperative variables remain relevant, but they serve as priors that initialize surveillance rather than as sufficient statistics for the entire problem. The critical methodological consequences follow naturally: evaluation should prioritize lead time, calibration, and alert burden; deployment should distinguish physiologic from process-based signal; and maintenance should anticipate drift and feedback once the system enters practice.

The central benefit of this reframing is not rhetorical. It is that it better matches the clinical task. Surgeons and postoperative teams already reason sequentially as evidence accumulates. A surveillance-oriented model formalizes that sequential reasoning, quantifies uncertainty more consistently, and has a clearer path to integration with actual postoperative workflows. Progress in this area is therefore likely to come not from ever more elaborate one-time prediction scores, but from disciplined longitudinal systems that recognize postoperative care as an ongoing inferential process. For a complication whose observability is delayed, whose physiology is dynamic, and whose management depends heavily on timing, that longitudinal formulation is the more appropriate foundation for research and eventual clinical deployment [31].

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] F. J. Hüttner, S. Tenckhoff, K. Jensen, L. Uhlmann, Y. Kulu, M. Büchler, M. K. Diener, and A. Ulrich, ''Meta-analysis of reconstruction techniques after low anterior resection for rectal cancer,'' *The British journal of surgery*, vol. 102, pp. 735–745, 3 2015.

[2] D. B. Shaul, M. Z. Schwartz, C. C. Marr, and K. R. Tyson, ''Primary repair without routine gastrostomy is the treatment of choice for neonates with esophageal atresia and tracheoesophageal fistula.,'' *Archives of surgery (Chicago, Ill. : 1960)*, vol. 124, pp. 1188–1191, 10 1989.

[3] P. Yimcharoen, H. M. Heneghan, N. Tariq, S. A. Brethauer, M. Kroh, and B. Chand, ''Endoscopic stent management of leaks and anastomotic strictures after foregut surgery,'' *Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery*, vol. 7, pp. 628–636, 5 2011.

[4] 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.

[5] R. K. Singh, T. H. Pham, B. S. Diggs, S. Perkins, and J. G. Hunter, ''Minimally invasive esophagectomy provides equivalent oncologic outcomes to open esophagectomy for locally advanced (stage ii or iii) esophageal carcinoma,'' *Archives of surgery (Chicago, Ill. : 1960)*, vol. 146, pp. 711–714, 6 2011.

[6] B. Protyniak, A. Dinallo, W. P. Boyan, R. M. Dressner, and M. L. Arvanitis, ''Intraoperative indocyanine green fluorescence angiography—an objective evaluation of anastomotic perfusion in colorectal surgery:,'' *The American surgeon*, vol. 81, pp. 580–584, 6 2015.

[7] O. Civil, N. Okkabaz, T. T. Sahin, C. Tiryaki, M. B. Yazıcıoğlu, and M. Kement, ''Long-term results and prognostic significance of non-anatomic liver resection for colorectal liver metastasis: Single center experience,'' *Indian Journal of Surgery*, vol. 82, pp. 197–204, 12 2019.

[8] K. Darji, ''Outcomes of lumen apposing metal stents (lams) in the management of benign colonic strictures and fistulae,'' *Journal of Medical Science And clinical Research*, vol. 9, 7 2021.

[9] D. Perez-Smith, M. Wagels, and D. R. Theile, ''Jejunal free flap reconstruction of the pharyngolaryngectomy defect: 368 consecutive cases.,'' *Journal of plastic, reconstructive & aesthetic surgery : JPRAS*, vol. 66, pp. 9–15, 9 2012.

[10] C. Gao, G. Xu, C.-Y. Wang, and D. Wang, ''Evaluation of preoperative risk factors and postoperative indicators for anastomotic leak of minimally invasive mckeown esophagectomy: a single-center retrospective analysis.,'' *Journal of cardiothoracic surgery*, vol. 14, pp. 46–46, 2 2019.

[11] I. M. Paterson and J. W. H. Wong, ''Anastomotic leakage: An avoidable complication of lewis-tanner oesophagectomy,'' *The British journal of surgery*, vol. 76, pp. 127–129, 2 1989.

[12] P. S. Waters, O. Peacock, S. K. Warrier, C. Wakeman, T. Eglinton, A. C. Lynch, F. A. Frizelle, A. G. Heriot, and J. J. McCormick, ''Evolution of pelvic exenteration surgery- resectional trends and survival outcomes over three decades.,'' *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 45, pp. 2325–2333, 7 2019.

[13] M. Daboos, M. Abdelmaboud, M. Hussein, A. Salama, and A. Elshamy, ''Azygos vein preservation revisited: impact on early outcomes after repair of esophageal atresia/tracheoesophageal fistula in newborns.,'' *Updates in surgery*, vol. 75, pp. 2305–2311, 11 2023.

[14] M. R. Ard, P. Lagergren, I. Rouvelas, J. Lagergren, M. V. D. Schaaf, J. Lagergren, P. Lagergren, L. Martin, S. Swisher, W. Hofstetter, A. Correa, R. Mehran, D. Rice, M. Bhat, M. Dar, G. Lone, A. Dar, J. Dai, Z. Zhang, J. Min, X. Huang, J. Wang, Y. Yuan, Y. Hu, T. Xie, Y. Zhao, D. Dindo, N. Demartines, P.-A. Clavien, M. Kitano, K. Zhang, and Y. Yang, ''Discussion.,'' *The Journal of thoracic and cardiovascular surgery*, vol. 144, pp. 1150–1151, 8 2012.

[15] M. Koëter, M. J. van der Sangen, C. W. Hurkmans, M. D. P. Luyer, H. J. T. Rutten, and G. A. P. Nieuwenhuijzen, ''Radiation dose does not influence anastomotic complications in patients with esophageal cancer treated with neoadjuvant chemoradiation and transhiatal esophagectomy,'' *Radiation oncology (London, England)*, vol. 10, pp. 59–59, 3 2015.

[16] D. Henne-Bruns and K. Kramer, ''Incidence of, risk factors for, and prevention of intestinal leakage,'' *Zentralblatt fur Chirurgie*, vol. 138, pp. 301–306, 5 2011.

[17] B. Struecker, S. Chopra, A.-C. Heilmann, J. Spenke, C. Denecke, I. M. Sauer, M. Bahra, J. Pratschke, A. Andreou, and M. Biebl, ''Routine radiologic contrast agent examination after gastrectomy for gastric cancer is not useful,'' *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, vol. 21, pp. 801–806, 2 2017.

[18] M. A. Abozeid, A. A. Abou-Zeid, H. M. Hussein, and M. A. Omar, ''Totally versus planned conversion laparoscopic colectomy for colorectal cancers: the beginning of the learning curve,'' *SVU-International Journal of Medical Sciences*, vol. 4, pp. 240–245, 8 2021.

[19] Y. Kurokawa, H. Katai, H. Fukuda, and M. Sasako, ''Phase ii study of laparoscopy-assisted distal gastrectomy with nodal dissection for clinical stage i gastric cancer: Japan clinical oncology group study jcog0703,'' *Japanese journal of clinical oncology*, vol. 38, pp. 501–503, 6 2008.

[20] E. Licht, A. J. Markowitz, M. S. Bains, H. Gerdes, E. Ludwig, R. B. Mendelsohn, N. P. Rizk, P. Shah, V. E. Strong, and M. A. Schattner, ''Endoscopic management of esophageal anastomotic leaks after surgery for malignant disease.,'' *The Annals of thoracic surgery*, vol. 101, pp. 301–304, 9 2015.

[21] Y. L. Verschoor, D. M. J. Lambregts, J. van den Berg, B. A. Grotenhuis, A. Aalbers, B. V. Triest, R. G. Beets-Tan, M. van de Belt, S. Dokter, S. Balduzzi, E. E. Voest, J. B. A. G. Haanen, M. E. van Leerdam, G. Beets, and M. Chalabi, ''Radiotherapy, atezolizumab, and bevacizumab in rectal cancers with the aim of organ preservation: The tarzan study.,'' *Journal of Clinical Oncology*, vol. 41, pp. 158–158, 2 2023.

[22] S. Mejía-Rivera, S. Pérez-Marroquín, R. Cortés-González, and H. Medina-Franco, ''Contrast-enhanced swallow study sensitivity for anastomotic leak detection in post-esophagectomy patients,'' *Revista de gastroenterologia de Mexico (English)*, vol. 83, pp. 400–404, 3 2018.

[23] S. K. Kamarajah, A. Madhavan, J. Chmelo, M. Navidi, S. Wahed, A. Immanuel, N. Hayes, S. M. Griffin, and A. W. Phillips, ''Impact of smoking status on perioperative morbidity, mortality, and long-term survival following transthoracic esophagectomy for esophageal cancer,'' *Annals of surgical oncology*, vol. 28, pp. 4905–4915, 3 2021.

[24] C. Jarry, L. Cárcamo, J. J. González, F. Bellolio, R. Miguieles, G. Urrejola, Álvaro Zúñiga, F. Crovari, M. E. Molina, and J. T. Larach, ''Implementation of intracorporeal anastomosis in laparoscopic right colectomy is safe and associated with a shorter hospital stay,'' *Updates in surgery*, vol. 73, pp. 93–100, 6 2020.

[25] J. P. S. Guzman, L. L. Resurreccion, M. L. R. Suntay, and R. G. Bernaldez, ''Comparison between hepaticojejunostomy and hepaticoduodenostomy after excision of choledochal cyst in children: a cohort study,'' *World Journal of Pediatric Surgery*, vol. 2, pp. e000029–, 6 2019.

[26] C. A. Galvani, M. V. Gorodner, F. Moser, G. R. Jacobsen, C. Chretien, N. J. Espat, P. Donahue, and S. Horgan, ''Robotically assisted laparoscopic transhiatal esophagectomy.,'' *Surgical endoscopy*, vol. 22, pp. 188–195, 10 2007.

[27] A. Holmén, M. Hayami, E. Szabo, I. Rouvelas, T. Agustsson, and F. Klevebro, ''Nutritional jejunostomy in esophagectomy for cancer, a national register-based cohort study of associations with postoperative outcomes and survival,'' *Langenbeck's archives of surgery*, vol. 406, pp. 1415–1423, 11 2020.

[28] V. N. Rohith, S. V. Arya, A. Rani, R. K. Chejara, A. Sharma, J. K. Arora, D. S. Kalwaniya, A. Tolat, P. G, and A. Singh, ''Preoperative serum albumin level as a predictor of abdominal wound-related complications after emergency exploratory laparotomy.,'' *Cureus*, vol. 14, pp. e31980–e31980, 11 2022.

[29] P. Kandagatla, A. H. Ghandour, A. Amro, A. M. Popoff, and Z. Hammoud, ''Long-term outcomes after robotic-assisted ivor lewis esophagectomy.,'' *Journal of robotic surgery*, vol. 16, pp. 1–7, 2 2021.

[30] O. Peacock, P. S. Waters, J. C. Kong, S. K. Warrier, C. Wakeman, T. Eglinton, D. G. Murphy, A. G. Heriot, F. A. Frizelle, and J. J. McCormick, ''Complications and 5-year survival after radical resections which include urological organs for locally advanced and recurrent pelvic malignancies: analysis of 646 consecutive cases,'' *Techniques in coloproctology*, vol. 24, pp. 181–190, 1 2020.

[31] S. Srivastava, S. S. Sikora, C. M. Pandey, A. Kumar, and R. Saxena, ''Determinants of pancreaticoenteric anastomotic leak following pancreaticoduodenectomy.,'' *ANZ journal of surgery*, vol. 71, pp. 511–515, 12 2001.