Morphpublishing
Research
open access academic journal

Peer reviewed

Hierarchical Outcome Construction and Latent Phenotype Inference for Anastomotic Leak in Perioperative Electronic Health Record Systems

Anish Thapa[a], **Bikash Adhikari**[b]

Abstract:
Postoperative adverse event modeling in gastrointestinal surgery increasingly relies on electronic health record data, yet the most difficult methodological task is often not model fitting but the formal specification of the event itself. Anastomotic leak is especially challenging because it does not appear in the record as a single stable datum. Instead, it emerges through delayed and heterogeneous manifestations that may include radiographic suspicion, inflammatory deterioration, procedural source control, antimicrobial escalation, operative revision, or diagnostic coding entered after the underlying process has already evolved. A clinically coherent computational system therefore requires more than a binary endpoint. It requires an explicit outcome construction framework and a phenotype representation that can distinguish between alternative expressions of the same underlying failure process. This paper develops a technical account of outcome definition and AL phenotyping using a hierarchical event model grounded in perioperative longitudinal data. The central argument is that the target of prediction should be treated as a latent postoperative state, while observed leak labels should be treated as operational reconstructions assembled from partially informative evidence streams. Building on that distinction, the paper formalizes event anchoring, temporal eligibility, hierarchical evidence fusion, graph-based phenotype representation, dynamic state estimation, uncertainty propagation, and cross-site transport. It further shows why binary AL labels often merge biologically different trajectories and why phenotype-aware modeling is better suited to surveillance, calibration, and implementation. The resulting framework supports interpretable risk estimation while preserving clinically relevant distinctions between early catastrophic failure, contained radiographic leak, management-defined leak, and ambiguous postoperative inflammatory states.

[a] **Karnali Academy of Health Sciences, Department of Health Informatics, Jumla–Chandannath Road, Khalanga, Jumla, Nepal**
[b] **Lumbini Buddhist University, Faculty of Science and Technology, Lumbini Sanskritik–Tenuhwa Road, Lumbini, Nepal**

1. Introduction

Anastomotic leak is one of the most analytically difficult postoperative outcomes because the event of interest is neither purely anatomical nor purely administrative [1]. It is anatomical in the sense that it concerns failure of an anastomotic construct, compromise of tissue integrity, and contamination of surrounding spaces. It is administrative in the sense that recognition of the event is mediated through coding practice, procedural documentation [2], imaging access, antibiotic choice, and institutional thresholds for reintervention. In a routine electronic health record, the analyst does not observe dehiscence as a transparent binary fact. The analyst observes traces of clinical suspicion and downstream management. For this reason, any effort to define AL as a machine-learning endpoint has to confront a foundational question that is often bypassed in simpler prediction settings: what exactly is being measured, at what time, and through which observation channels [3].

This difficulty is not a mere philosophical inconvenience. It changes the statistical meaning of the target. If a leak is declared only when a patient is reoperated on, then the label partly reflects surgical aggressiveness. If a leak is declared only when a diagnostic code appears, then the label partly reflects documentation and coding discipline. [4] If a leak is declared through broad-spectrum antibiotic escalation plus imaging, then the label partly reflects local care pathways. Outcome definition is therefore inseparable from ascertainment [5]. The target label cannot be presumed to exist independently of the observation system through which it is assembled [6]. In perioperative prediction work, this fact is particularly consequential because the same clinical state can produce different recorded signatures across hospitals, surgeons, and time periods [7].

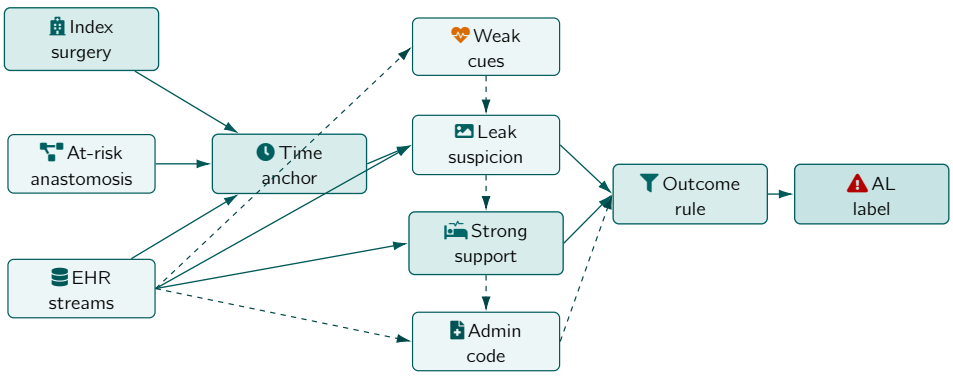

Figure 1. Hierarchical outcome construction for anastomotic leak. The diagram separates anatomical eligibility and temporal anchoring from the evidence hierarchy, then fuses weak cues, suspicion, corroboration, and delayed administrative abstraction into a single operational AL label.

A proof-of-concept EHR study operationalized AL through a composite 30-day outcome requiring diagnosis-code evidence together with corroborating clinical evidence, used features available within the first 24 postoperative hours, observed a positive rate of 7.6%, and found that post-operative white blood cell count, ICU admission, post-operative lactate, surgery type, and pre-operative albumin were among the strongest predictors, while also identifying reproducibility, external validation, and subgroup analysis as unresolved challenges [8]. That design choice is important because it acknowledges the incompleteness of any single coding signal. At the same time, it raises a more general methodological issue. Once outcome definition is composite, the event is no longer a simple label but a rule over heterogeneous evidence streams. The next logical step is to ask whether those streams should only be fused into a single binary endpoint or whether they should also be used to identify clinically meaningful leak phenotypes.

Published in CURRENT ISSUE

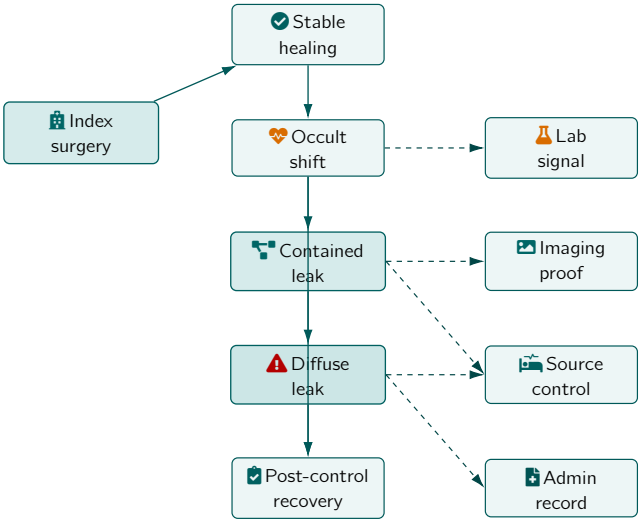

Figure 2. Latent postoperative state versus recorded manifestations. The central trajectory represents the unobserved clinical failure process, whereas laboratory changes, imaging, source-control procedures, and administrative coding appear as partial observation channels with different timing and specificity.

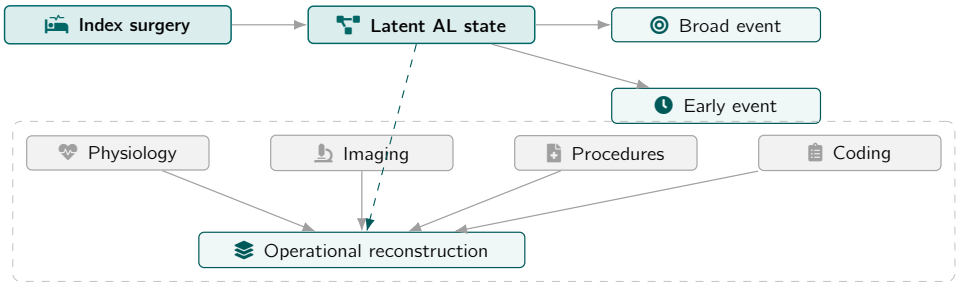

Figure 3. Event semantics for anastomotic leak. The target is framed as a latent postoperative failure state anchored to the index surgery, while physiology, imaging, procedures, and coding provide partially informative observations for operational reconstruction. Broad and early estimands share the same latent substrate but differ in temporal admissibility.

This paper argues for the second path. A patient with diffuse contamination requiring urgent source control, a patient with a small contained pelvic collection managed by drainage, and a patient with persistent inflammatory deterioration who never undergoes definitive radiographic confirmation are not equivalent data objects even when each lies within the broader semantic territory of anastomotic failure. Treating them as exchangeable positive cases simplifies the prediction task but compresses heterogeneity that matters for calibration, interpretability, and clinical action. A surveillance model trained on such compressed labels may perform well in rank ordering while remaining opaque about which kind of leak it has learned. The same predicted probability can then correspond to different latent clinical regimes, making downstream interpretation unstable.

The central proposal of this paper is to formalize AL in two layers [9]. The first layer is a hierarchical outcome construction mechanism that maps longitudinal perioperative records into an event indicator with explicit temporal and evidentiary logic. The second layer is a phenotype inference mechanism that embeds positive and near-positive admissions into a structured latent space defined by onset pattern, containment, physiologic burden, intervention

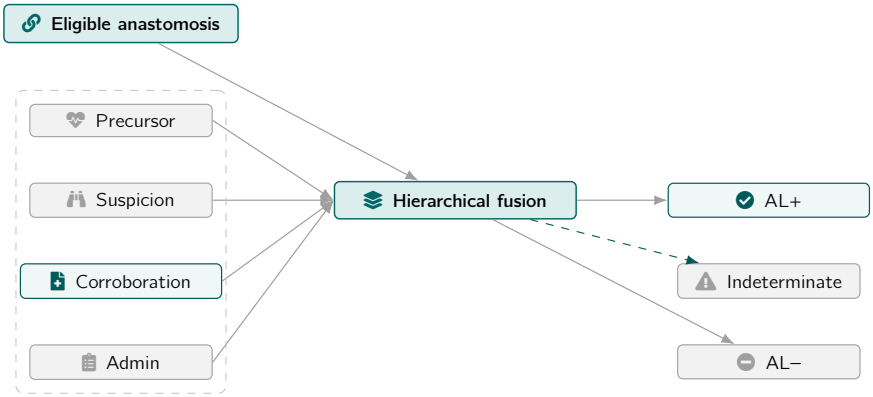

Figure 4. Hierarchical label construction. Eligibility is separated from evidence assembly, and evidence layers contribute with different credibility rather than as exchangeable signals. The fusion stage naturally supports positive, indeterminate, and negative operational states.

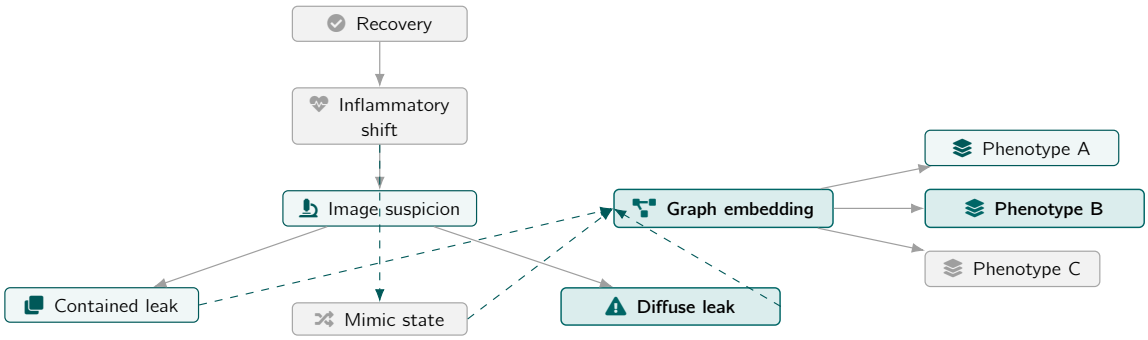

Figure 5. Latent phenotype inference on a clinical event graph. The event graph separates recovery, inflammatory deviation, suspicion, contained failure, diffuse failure, and non-leak mimics, while the embedding stage maps these structured activations into softer phenotype neighborhoods rather than forcing rigid subtype labels.

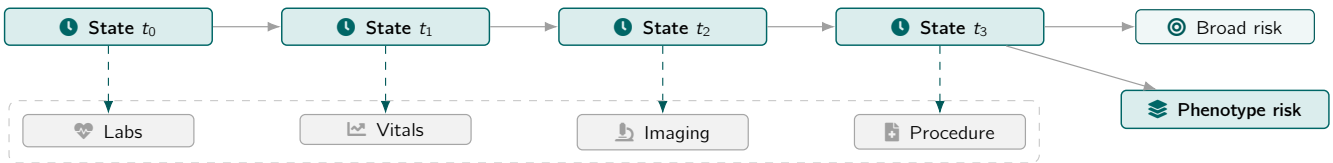

Figure 6. Dynamic state-space surveillance. Hidden postoperative states evolve over time, irregular observations provide partial updates, and the most recent posterior state emits both a broad leak risk and a phenotype-specific forward risk. This separates progression from measurement and preserves the time structure of surveillance.

intensity, and evidentiary modality. This layered approach treats the binary endpoint as necessary but insufficient. The event indicator is preserved for benchmarking and decision thresholds, while the phenotype representation captures heterogeneity that the event indicator necessarily discards.

A hierarchical formulation is preferable to flat classification for several reasons. First, the evidence for leak is not homogeneous. Operative revision, drain placement for an anastomotic collection, radiographic extravasation, broad-spectrum antimicrobial escalation, laboratory deterioration, and diagnostic coding do not contribute equally to the

Published in CURRENT ISSUE

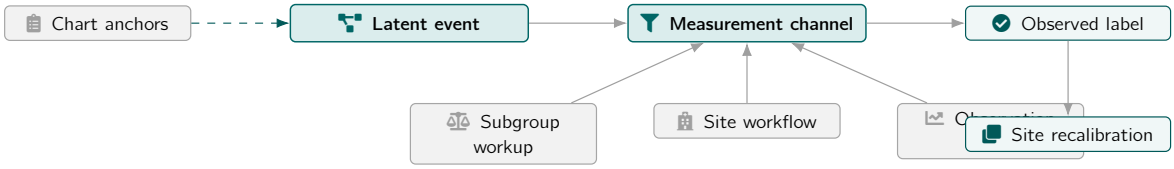

Figure 7. Identifiability, fairness, and distribution shift. The latent event is only partially anchored by chart evidence, while subgroup workup, site workflow, and observation intensity reshape the measurement channel that produces the observed label. Transport therefore benefits from recalibration at the observation layer rather than assuming direct portability of raw labels.

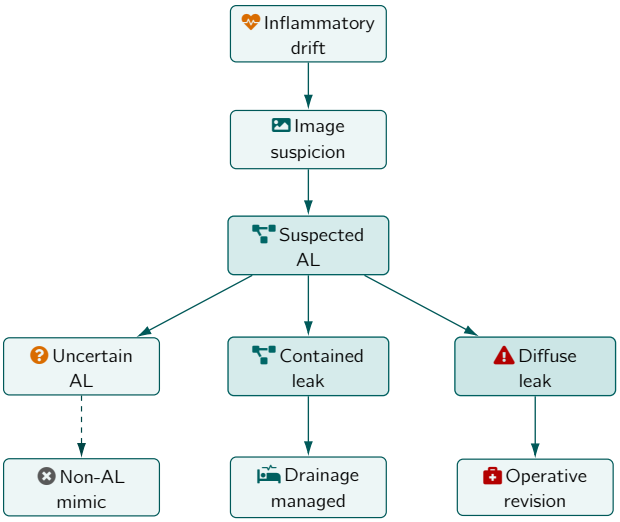

Figure 8. Clinical event graph for latent phenotype inference. Positive and near-positive admissions are organized as neighboring graph regions rather than rigid classes, allowing ambiguous inflammatory states, contained leak, and severe diffuse failure to remain clinically distinct while still connected through a shared suspicion node.

credibility or semantics of the event. Second, those evidence channels are temporally ordered in ways that matter. Some are antecedent physiologic signatures, some are confirmatory signals, and some are delayed administrative abstractions. Third, heterogeneity in leak expression is not random residual variance. It is part of the clinical object. A contained leak is not just a milder version of a catastrophic leak [10]. It may differ in detectability, time course, physiologic signature, and management consequences. Fourth, deployment in real perioperative settings requires outputs that are intelligible enough to support action before overt decompensation, not merely after it.

The present discussion is intentionally technical and designed for data scientists and clinical informaticians working with longitudinal surgical records. It does not assume that a unique universal AL definition exists. Instead, it treats outcome construction as an estimand-design problem. The estimand may be broad postoperative anastomotic failure, clinically significant leak requiring intervention, or early occult leak detectable before overt sepsis. Each estimand induces a different label rule, different eligible predictor window, and different phenotype geometry. The methodological task is to keep these elements aligned. A mismatch between estimand and label rule produces hidden bias. A mismatch between label rule and phenotype representation produces semantic instability. A mismatch between phenotype representation and deployment objective produces clinically unhelpful outputs.

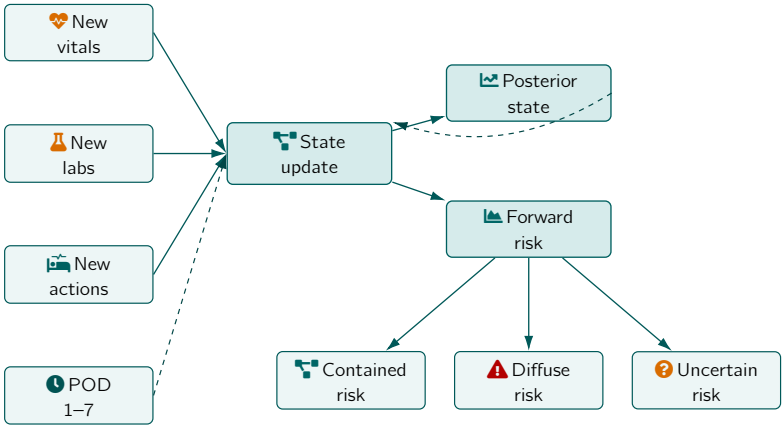

Figure 9. Dynamic surveillance as a partially observed state-space system. Irregular postoperative inputs update a latent state estimate, which then produces phenotype-specific forward risks for contained leak, diffuse leak, and uncertain leak-like deterioration over the next clinical horizon.

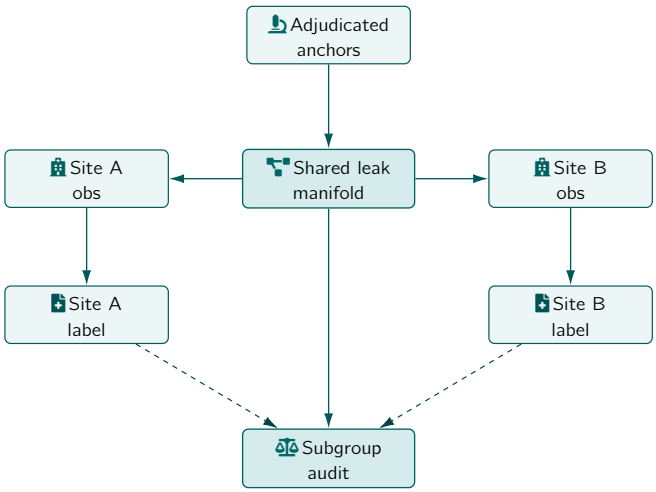

Figure 10. Transport and fairness structure under changing observation systems. A shared latent clinical manifold is constrained by adjudicated anchors, while site-specific observation channels generate different local labels and must be audited for subgroup-dependent ascertainment and calibration drift.

The paper proceeds by first defining the event semantics of AL in a longitudinal record and by distinguishing latent failure from observed evidence [11]. It then develops a hierarchical label construction strategy in which evidence channels are arranged according to clinical and temporal priority rather than simply pooled. After that, it introduces a graph-based phenotype model for leak states and near-leak states, followed by a dynamic state-space formulation for time-updated surveillance. The later sections examine identifiability, fairness, and transport under changing observation systems, and then turn to validation and clinical use. The overall objective is modest but technically demanding: to specify an AL endpoint and phenotype system that can be audited, transported, and interpreted without pretending that the postoperative failure process is directly observable.

Published in CURRENT ISSUE

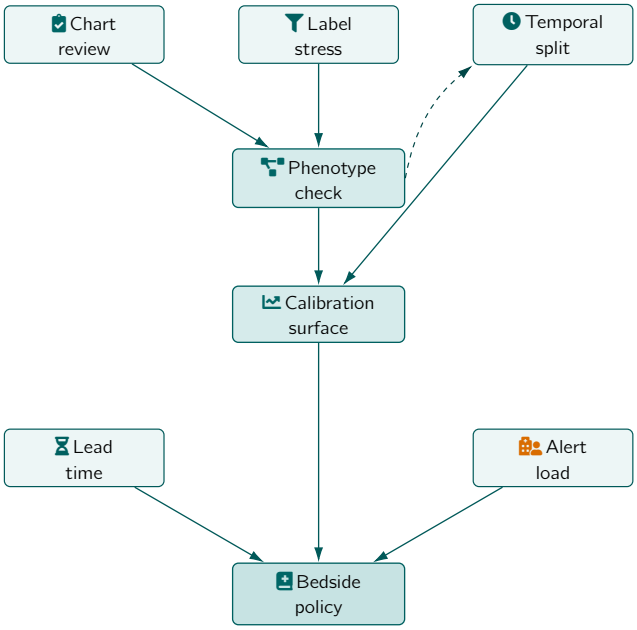

Figure 11. Validation and decision-oriented deployment stack. The framework is tested through chart review, perturbation of the label rule, temporal splits, phenotype stability, and calibration surfaces before thresholds are converted into bedside policy under explicit lead-time and alert-burden constraints.

2. Event Semantics and Estimand Design

A computational outcome definition should begin with the object of clinical interest rather than with the convenience of available variables. In the case of AL, the clinically salient object is failure of an anastomotic interface leading to pathological communication, contamination, local collection, fistulization, or systemic deterioration attributable to loss of integrity. This formulation already contains several layers. There is a structural layer concerning tissue and anastomotic continuity. There is a physiologic layer concerning inflammatory burden and systemic response. There is a management layer concerning what clinicians do in response to suspected failure. The routine record primarily captures the latter two layers and only partially captures the first. Consequently, the analyst should distinguish the latent clinical state from the observed event proxy [12].

Let the index surgery for admission i occur at time s_i, and let the latent postoperative condition be described by a state process $\xi_i(t)$ for $t \geq s_i$. The state process need not be binary. It can encode stable healing, local subclinical compromise, overt leak with containment, disseminated contamination, and post-source-control recovery. The challenge is that $\xi_i(t)$ is not directly recorded. Instead, the EHR generates an event stream

$$\begin{aligned} \mathcal{R}_i &= \Big\{(r_{i1}, \tau_{i1}), \ldots, (r_{im_i}, \tau_{im_i})\Big\}, \\ r_{ij} &\in \mathcal{A}_{\text{diag}} \cup \mathcal{A}_{\text{proc}} \cup \mathcal{A}_{\text{lab}} \cup \mathcal{A}_{\text{med}} \cup \mathcal{A}_{\text{text}} \cup \mathcal{A}_{\text{img}}, \end{aligned} \tag{1}$$

where each token r_{ij} belongs to a modality-specific alphabet and each τ_{ij} denotes occurrence time. The outcome definition problem is then to construct a map from $\mathcal{R}_i$ to an event variable whose meaning is explicit enough to support inference.

An estimand is useful only if it specifies what is being counted. In AL studies, several plausible estimands coexist. One estimand is any clinically meaningful postoperative anastomotic failure within a fixed horizon. Another is failure severe enough to generate radiographic or operative confirmation. Another is failure detectable early enough to alter management before gross decompensation. These are not equivalent. Suppose two admissions display identical structural compromise, but one is diagnosed immediately after early imaging and the other only after delayed septic deterioration. The first admission belongs to an early-detectable estimand, whereas both belong to a broad 30-day leak estimand [13]. A model trained on the latter may have limited value for the former even if both carry the same binary endpoint.

Formally, let $\mathcal{H}_i(u)$ denote the observed history up to landmark time u. A broad event estimand over horizon H may be written as

$$Y_i^{(B)}(H) = \mathbf{1}\Big\{\exists t \in (s_i, s_i + H] \text{ such that } \xi_i(t) \in \mathcal{S}_{\text{leak}}\Big\}, \tag{2}$$

where $\mathcal{S}_{\text{leak}}$ is the set of leak states. By contrast, an early-detectable estimand conditional on available information by time u is

$$Y_i^{(E)}(u, H) = \mathbf{1}\Big\{\exists t \in (u, s_i + H] \text{ such that } \xi_i(t) \in \mathcal{S}_{\text{leak}}\Big\} \quad \text{given } \mathcal{H}_i(u). \tag{3}$$

These two estimands differ because the second excludes information unavailable by the surveillance landmark. A label construction method that ignores this distinction risks contaminating predictors with post-declaration evidence.

The semantics of AL also require a distinction between intrinsic event content and extrinsic confirmation. Intrinsic content concerns the underlying structural failure and its physiological consequences. Extrinsic confirmation concerns how that failure becomes visible in the record. Operative revision, radiographic contrast extravasation, or drainage of an anastomotic collection are strong confirmation channels, but they are still not the same as the latent structural event. Their presence depends on clinician suspicion, imaging capacity, and procedural choice. A hospital with lower imaging thresholds will likely generate more confirmation events for the same latent burden. Therefore, confirmation should not be mistaken for ground truth [14]. It is better treated as an observation operator with variable sensitivity.

This motivates a measurement model. Let Y_i^* denote the latent broad-event indicator and let O_i denote an observed operational label. Then

$$\begin{aligned} \Pr(O_i = 1 \mid Y_i^* = 1, \zeta_i) &= \alpha(\zeta_i), \\ \Pr(O_i = 1 \mid Y_i^* = 0, \zeta_i) &= \beta(\zeta_i), \end{aligned} \tag{4}$$

where ζ_i encodes context such as site, service line, phenotype, and observation intensity. This pair of functions describes sensitivity and false-positive probability conditional on context. The dependence on ζ_i is not a technical embellishment. It captures the fact that mild contained leaks may be underascertained, severe leaks may be overrepresented, and local workflow may shape which cases become visible.

A clinically faithful estimand should therefore satisfy three criteria. First, it should align with a meaningful postoperative state rather than a single documentation artifact. Second, it should be temporally coherent with the intended prediction window. Third, it should disclose which parts of the ascertainment process are being incorporated into the target. If the analyst chooses a management-defined estimand because intervention burden is the relevant endpoint, that choice is legitimate [15]. If the analyst wants to characterize latent structural failure irrespective of intervention, then management data should be treated as partial evidence rather than the event itself.

Published in CURRENT ISSUE

These distinctions become especially important when the same model is expected to support multiple tasks. Consider triage, etiologic investigation, and institutional benchmarking. Triage requires an event definition with good temporal lead time and tolerable false-positive burden. Etiologic investigation requires a target that tracks biological failure rather than management convention. Benchmarking requires a definition comparable across hospitals. A single undifferentiated binary endpoint rarely serves all three purposes well. A hierarchical framework can do so more effectively because it separates a broad latent event, several operationally convenient observed labels, and a phenotype representation that captures within-event diversity.

One useful way to encode this structure is through an event ontology. Let the ontology be a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes represent clinically interpretable event concepts and whose edges represent implication or precedence. Nodes may include suspected inflammatory deviation, image-supported collection, operative confirmation, drainage-supported source control, contained leak, diffuse leak, and uncertain postoperative deterioration. Observed tokens in $\mathcal{R}_i$ map to ontology nodes through a relation φ. The event definition then becomes an aggregation functional over ontology activations rather than over raw codes and procedures. This design has two advantages [16]. It reduces sensitivity to coding idiosyncrasy by grouping synonymous evidence. It also enables phenotype construction because leak states can be represented as pathways through the graph rather than as isolated labels.

The concept of estimand design thus shifts the discussion away from the false choice between broad binary outcomes and highly specific chart-adjudicated events. What matters is not only precision but semantic transparency. A broad outcome can be valid if its inferential target is explicit and its error structure is examined. A narrow outcome can be misleading if it mainly counts heavily worked-up cases. The right objective is an event representation whose layers are clear enough that analysts, surgeons, and implementation teams know whether they are modeling latent failure, documented suspicion, intervention-requiring failure, or some composite of these. Only then can AL phenotyping be built on stable conceptual ground.

3. Hierarchical Label Construction in Longitudinal Records

Once the estimand is defined, one must specify how the longitudinal record is translated into an operational label. A flat logical rule such as diagnosis code plus antibiotic exposure within 30 days is often a reasonable starting point, but it leaves too much semantic structure implicit. It treats all supporting evidence as exchangeable, overlooks the ordering of suspicion and intervention, and offers little guidance for ambiguous cases. A hierarchical construction is more appropriate because leak evidence naturally stratifies into antecedent physiologic disturbance, diagnostic or radiographic suspicion, confirmatory intervention, and delayed administrative coding [17]. These strata are not interchangeable. They correspond to different positions in the clinical recognition pathway.

Consider a hierarchy with four levels. The base level contains weak but potentially informative precursors such as persistent inflammatory deviation, unexplained tachycardia, rising lactate, or delayed recovery patterns. The second level contains suspicion events, including note-based concern, targeted imaging, or provisional diagnostic language. The third level contains stronger corroboration such as drainage procedures, operative revision, or explicit radiographic evidence. The fourth level contains administrative consolidation, including diagnosis coding and discharge abstraction. The key idea is that evidence should be aggregated upward with precedence, not merely pooled. A leak label generated from strong corroboration should not depend on whether delayed coding later appears, whereas a label generated only from weak precursors should remain low confidence.

A convenient formalization uses modality-specific scores that are fused under monotone constraints. Let $x_i^{(v)}$ denote evidence extracted from modality v, where $v \in \{0, 1, 2, 3\}$ indexes precursor, suspicion, corroboration, and

administrative layers. Construct scores

$$q_i^{(v)} = f_v\big(x_i^{(v)}\big),$$
$$q_i = \omega_0 q_i^{(0)} + \omega_1 q_i^{(1)} + \omega_2 q_i^{(2)} + \omega_3 q_i^{(3)}, \tag{5}$$

with nonnegative weights satisfying $\omega_2 \geq \omega_1 \geq \omega_0$ and $\omega_2 \geq \omega_3$ if one wishes to privilege corroborating clinical evidence over purely administrative recording [18]. The operational label can then be defined by thresholds on q_i, but the score itself carries richer information than a hard threshold because it reflects evidence composition.

Purely additive fusion, however, may still be too permissive. Certain combinations should be disallowed or down-weighted because they are clinically implausible. Broad-spectrum antibiotics after surgery may indicate many processes unrelated to AL. A radiographic collection without procedural or clinical corroboration may represent a different postoperative complication. Hierarchical construction therefore benefits from logical gating. One can write

$$O_i = \mathbf{1}\Big\{ q_i^{(2)} \geq c_2 \ \vee \ \big(q_i^{(1)} \geq c_1 \ \wedge \ q_i^{(3)} \geq c_3\big) \ \vee \ \big(q_i^{(0)} \geq c_0 \ \wedge \ q_i^{(1)} \geq c_1 \ \wedge \ q_i^{(3)} \geq c_3\big) \Big\}. \tag{6}$$

This rule reflects the idea that strong corroboration can suffice on its own, whereas weaker signals require multi-layer concurrence. The exact thresholds are context dependent, but the hierarchical principle remains stable.

Temporal anchoring is equally important. Every evidence channel must be related to the index surgery, and every triggering event should have a recorded timestamp. A single 30-day horizon may be clinically appropriate for broad capture, yet the internal structure of that window matters. Evidence in the first 24 hours often belongs to the predictor domain or to background physiologic response, not to confirmed leak declaration [19]. Evidence arising between postoperative days 3 and 7 may be most informative for early clinically manifest events. Evidence near day 30 may capture delayed collections, transfer-related documentation, or outpatient coding. Therefore, the hierarchy should encode not only evidentiary strength but also temporally appropriate roles for each signal.

Let τ_{ij} be the time of evidence token j and define time kernels $K_v(t)$ for each layer. For example, precursor evidence may be meaningful early, corroboration may be more meaningful after an initial postoperative interval, and administrative coding may remain admissible late. Then

$$q_i^{(v)} = \sum_{j=1}^{m_i} \mathbf{1}\{r_{ij} \in \mathcal{A}_v\}\, K_v(\tau_{ij} - s_i)\, \eta(r_{ij}), \tag{7}$$

where $\eta(r_{ij})$ is an evidence-specific credibility contribution. The time kernels can penalize implausibly early or late signals. This makes the label less vulnerable to arbitrary inclusion of temporally misplaced events.

A further advantage of hierarchical construction is explicit handling of indeterminate admissions. Not every postoperative inflammatory trajectory should be forced into leak or non-leak. Some admissions contain sufficient abnormality to merit surveillance but insufficient evidence for confident leak designation. For such cases, one may introduce an intermediate label U_i for uncertainty or ambiguity. In a three-way scheme,

$$Y_i^{\mathrm{tri}} = \begin{cases} 1, & q_i \geq \kappa_{\mathrm{high}}, \\ u, & \kappa_{\mathrm{low}} \leq q_i < \kappa_{\mathrm{high}}, \\ 0, & q_i < \kappa_{\mathrm{low}}, \end{cases} \tag{8}$$

where u denotes indeterminate status [20]. This triadic representation is not merely a convenience for annotation. It recognizes that the EHR often contains partial evidence and that forcing all such cases into binary classes

Published in CURRENT ISSUE

can distort both prevalence and model learning. Ambiguous cases may later be used as weakly labeled examples, excluded from certain validation analyses, or retained in phenotype models with high uncertainty.

The denominator problem also sits naturally within a hierarchical framework. The at-risk set is not just any postoperative admission. It is the set of admissions in which an anastomosis capable of leaking was created and in which follow-up is adequate for the intended horizon. If the denominator includes operations without an anastomosis or excludes complex salvage cases because documentation is messy, the label may become internally tidy at the cost of clinical meaning. Hierarchical construction should therefore include a separate eligibility layer that maps operative documentation into anatomical risk states. The event rule only operates once that eligibility layer is satisfied.

One can formalize eligibility through an anastomotic exposure indicator A_i and an observation adequacy indicator C_i:

$$E_i = \mathbf{1}\{A_i = 1 \ \wedge \ C_i = 1\}. \tag{9}$$

The final observed label is then defined only on admissions with $E_i = 1$. This separation matters because it allows sensitivity analyses that vary anatomical inclusion and follow-up completeness independently of the leak rule itself [21]. Too many studies merge these decisions, making it impossible to tell whether a change in prevalence stems from outcome logic or from denominator drift.

Hierarchical labels also permit uncertainty-aware training. Suppose the analyst has a subset of chart-adjudicated cases and a larger set of weak labels generated from the hierarchy. The model can be trained with confidence weights w_i derived from evidence composition, so that strongly corroborated cases contribute more to the loss than weak or ambiguous cases. A weighted empirical risk may be written as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} w_i \, \ell\big(\hat{p}_\theta(\mathcal{H}_i), Y_i^{\mathrm{obs}}\big), \tag{10}$$

where w_i increases with corroboration level or chart-review confidence. This strategy acknowledges that not all labels are equally reliable while still using the full dataset.

Finally, hierarchical construction creates a direct bridge to phenotyping. The same layers used to build the label also reveal how a leak became legible. Some admissions are corroboration-dominant, some suspicion-dominant, some administration-dominant, and some remain precursor-heavy without definitive confirmation. These patterns are clinically relevant because they often align with severity, detection delay, and local workflow. A flat endpoint discards such distinctions [22]. A hierarchical endpoint preserves them as structured metadata, making the subsequent phenotype inference problem more stable and more interpretable. In that sense, label construction is not a preprocessing nuisance but the first stage of phenotype discovery.

4. Latent Phenotype Inference on a Clinical Event Graph

A phenotype system should not be introduced merely to multiply categories. It should be introduced because the binary leak endpoint compresses heterogeneity that carries mechanistic and operational significance. The most useful computational representation is therefore not a list of manually declared subtypes but a structured latent space informed by a clinical event graph. The graph serves as an ontology linking observed events to clinically meaningful leak expressions, while the latent space allows admissions to occupy intermediate or uncertain positions rather than being forced into rigid mutually exclusive bins.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a directed event graph whose nodes represent concepts such as uncomplicated recovery, disproportionate inflammatory response, image-triggered suspicion, contained collection, diffuse contamination,

intervention-defined leak, and administratively supported leak. Each admission activates a subgraph according to its observed tokens. If a_{iv} denotes the activation intensity of node v for admission i, then the admission is represented by an activation vector $a_i \in \mathbb{R}^{|\mathcal{V}|}$. The challenge is to infer a lower-dimensional latent representation that preserves both clinical meaning and graph topology.

One approach is graph-regularized matrix factorization. Let $A \in \mathbb{R}^{n \times |\mathcal{V}|}$ be the matrix of activation vectors. We seek an embedding $H \in \mathbb{R}^{n \times d}$ and node factors $B \in \mathbb{R}^{|\mathcal{V}| \times d}$ such that

$$\begin{aligned} A &\approx HB^\top, \\ \mathcal{J}(H, B) &= \|A - HB^\top\|_F^2 + \lambda \operatorname{tr}(B^\top L_{\mathcal{G}} B), \end{aligned} \tag{11}$$

where $L_{\mathcal{G}}$ is the graph Laplacian of the ontology. The regularization term penalizes embeddings that ignore known relations among nodes. If contained leak and diffuse leak share common ancestors in the graph but differ in severity descendants, the factorization respects that geometry rather than treating nodes as unrelated columns.

The resulting latent coordinates in H can be interpreted as phenotype dimensions [23]. One axis may correspond to physiologic burden, another to local containment versus dissemination, another to observation modality, and another to management aggressiveness. Importantly, these axes need not be predefined in the data matrix. They emerge subject to the graph structure and can be audited against clinical expectations. An admission with strong activation of inflammatory precursor, targeted imaging, and drainage but no reoperation might fall in a region of the latent space corresponding to contained intervention-managed leak. Another admission with abrupt hemodynamic collapse and operative revision might occupy a more severe region. A third admission with persistent inflammatory irregularity and administrative coding but little corroboration might fall near an uncertainty boundary.

A discrete phenotype taxonomy can then be induced from the latent coordinates by mixture modeling. Let h_i denote the row of H for admission i. A Gaussian mixture representation takes the form

$$\begin{aligned} p(h_i) &= \sum_{k=1}^{K} \pi_k \mathcal{N}(h_i; \mu_k, \Sigma_k), \\ p(Z_i = k \mid h_i) &= \frac{\pi_k \mathcal{N}(h_i; \mu_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}(h_i; \mu_\ell, \Sigma_\ell)}, \end{aligned} \tag{12}$$

where Z_i is the latent phenotype index. The posterior membership probabilities allow soft assignments, which are preferable because leak phenotypes rarely have sharp boundaries. A case can simultaneously resemble a contained leak and an intervention-defined leak, or a near-leak inflammatory mimic and a low-confidence leak state [24].

Phenotypes should not be allowed to drift into purely observational clusters. For instance, one cluster may emerge simply because a hospital orders postoperative CT scans more frequently, not because its patients have a distinct biological leak state. To prevent this, the phenotype model should partially factor out observation intensity. Let m_i denote a vector of measurement-process features such as imaging count, note density, or lab frequency. A residualized latent coordinate

$$\begin{aligned} \tilde{h}_i &= h_i - P_m h_i, \\ P_m &= M(M^\top M)^{-1} M^\top \end{aligned} \tag{13}$$

can be used, where the rows of M are the m_i vectors. This projection removes linear components associated with observation density. More general nonlinear residualization is possible, but the principle is that phenotype should be less about how closely a patient was watched and more about what happened biologically and clinically.

Published in CURRENT ISSUE

Another useful device is to impose partial ordering constraints between phenotype centroids. If one phenotype is intended to represent more severe contamination than another, then its centroid should score higher on severity-linked node activations. Suppose $c^\top h_i$ is a severity functional extracted from the embedding. Then one may enforce

$$c^\top \mu_{k_1} \leq c^\top \mu_{k_2} \quad \text{whenever} \quad k_1 \preceq k_2, \tag{14}$$

where $\preceq$ is a clinically specified partial order [25]. This prevents semantically inverted solutions in which a "milder" phenotype appears more physiologically extreme than a "severe" one. Such inversions are common in unconstrained clustering when observation-process features dominate.

A graph-based phenotype framework is especially well suited to admissions that lie close to the leak boundary. Not every abnormal postoperative course reflects AL. Prolonged ileus, pneumonia, intra-abdominal abscess unrelated to the anastomosis, bleeding, and generalized sepsis from other sources can all mimic portions of the leak signal. Rather than treating all false positives as unstructured noise, the event graph can include non-leak mimic regions. Admissions then populate a broader manifold in which actual leak states lie adjacent to coherent alternative postoperative syndromes. This is valuable because many predictive models are applied exactly in this ambiguous territory. The model must distinguish not only leak from normal recovery but leak from its closest clinical competitors.

The latent event graph also supports transition analysis. Because nodes are temporally activatable, one can estimate likely pathways through the graph. An admission may move from mild inflammatory deviation to imaging-triggered suspicion to contained collection with drainage [26]. Another may move rapidly from early instability to operative revision. These pathways are phenotypically informative because they capture velocity and sequence, not only final state. Let $\Gamma_i = (v_{i1}, \ldots, v_{iT_i})$ be the ordered path of dominant node activations. Phenotype can then be defined jointly by latent coordinates and path structure. This is particularly useful for separating early catastrophic leaks from delayed localized collections, even if both end in the same binary outcome.

Clinical interpretability benefits from the graph formalism because each phenotype remains tied to an explicit neighborhood of concepts. Instead of reporting an abstract cluster number, the system can describe a phenotype as a region of the graph characterized by strong contained-collection activation, moderate physiologic burden, frequent drainage, and low operative revision intensity. Another phenotype can be described as early diffuse-failure activation with high hemodynamic disturbance and rapid operative confirmation. These descriptions are not decorative. They determine whether a phenotype output can be understood and trusted in a perioperative workflow.

The graph-based strategy is different from simple unsupervised clustering on tabular features because it encodes medical semantics directly into the representation. The embedding is not free to separate cases along arbitrary administrative axes if those axes conflict with the graph topology. At the same time, it remains flexible enough to discover variation within clinically meaningful regions [27]. This balance is important. Fully hand-built taxonomies are often too rigid, while unconstrained data-driven clusters are often too unstable. The event graph provides an intermediate scaffold that stabilizes inference without predetermining every subtype.

From a methodological standpoint, phenotype inference on a clinical event graph offers a way to reconcile binary benchmarking with richer clinical structure. The binary label remains available for incidence estimation and simple prediction. Yet around that binary core, the graph-based latent space captures containment, severity, timing, and evidentiary route. This dual representation is particularly attractive in AL because the event itself sits at the intersection of anatomy, physiology, and care process. A single scalar probability can summarize risk, but only a structured latent representation can explain what kind of leak state that probability refers to.

5. Dynamic Risk Surfaces and State-Space Modeling

AL surveillance is intrinsically dynamic. The probability that an admission occupies or approaches a leak state changes over time as new measurements arrive, as hemodynamics evolve, and as clinicians order imaging or interventions. A static model trained on one summary vector per admission can be useful for early screening, but it cannot represent the evolving geometry of postoperative risk. A more faithful approach treats leak progression as a partially observed state-space system in which latent clinical status changes continuously while observations arrive irregularly and selectively [28].

Let $z_i(t)$ denote a continuous latent state vector for admission i at postoperative time t. Components of $z_i(t)$ may encode local structural compromise, systemic inflammatory burden, perfusion inadequacy, and detection readiness. The state evolves according to

$$\begin{aligned} z_i(t+\Delta) &= F_\Delta z_i(t) + G_\Delta u_i(t) + \varepsilon_i(t,\Delta), \\ \varepsilon_i(t,\Delta) &\sim \mathcal{N}(0, Q_\Delta), \end{aligned} \tag{15}$$

where $u_i(t)$ includes interventions or exogenous drivers and F_Δ is a transition operator over time increment Δ. Observations $y_i(t)$ such as laboratory values, vitals, imaging indicators, or procedural events are emitted through

$$\begin{aligned} y_i(t) &= H_t z_i(t) + \nu_i(t), \\ \nu_i(t) &\sim \mathcal{N}(0, R_t) \end{aligned} \tag{16}$$

for continuous channels, with generalized observation models for binary or count channels. This representation captures a central asymmetry: the latent clinical condition evolves whether or not it is being measured, but the record reveals only irregular shadows of that evolution.

A state-space formulation is useful because it separates progression from observation. In ordinary classification, a rising leukocyte count and a CT order are simply two features. In a dynamic system, the leukocyte trajectory may be modeled as an observation of latent inflammatory burden, while the CT order may be modeled as a consequence of the posterior probability of clinically significant deterioration crossing a practical threshold. The distinction matters because care-process events are not equivalent to physiologic measurements. They are downstream responses to clinical suspicion, and thus informative about the latent state in a different way [29].

For surveillance, what matters is not merely the current latent state but the risk surface over future leak declaration. Let T_i^* be the latent time at which the state first enters a clinically significant leak region $\mathcal{L}$. Define the dynamic risk

$$\rho_i(t,h) = \Pr\Big(\exists s \in (t, t+h] \text{ such that } z_i(s) \in \mathcal{L} \mid \mathcal{H}_i(t)\Big), \tag{17}$$

where $\mathcal{H}_i(t)$ is the observed history up to time t. This quantity is more clinically meaningful than a static 30-day probability once the patient is already postoperative. It can be updated as new information arrives and can support threshold-based actions such as intensified monitoring or earlier imaging.

The risk surface can be linked to phenotype by defining several destination regions $\mathcal{L}_k$ corresponding to contained leak, diffuse leak, intervention-defined leak, and ambiguous leak-like deterioration. Then

$$\rho_{ik}(t,h) = \Pr\Big(\exists s \in (t, t+h] \text{ such that } z_i(s) \in \mathcal{L}_k \mid \mathcal{H}_i(t)\Big) \tag{18}$$

yields phenotype-specific forward risks. This is operationally important. A high forward risk of contained leak may support one monitoring strategy, whereas a high forward risk of rapidly progressive diffuse failure may support

Published in CURRENT ISSUE

another. A phenotype-aware surveillance system therefore provides more actionable information than a single undifferentiated alert probability.

Irregular sampling is a major challenge in postoperative records. Laboratory tests, bedside vital signs, and imaging are measured on different schedules, and those schedules depend on illness severity. Standard recurrent encoders may inadvertently learn that measurement frequency itself is predictive, which it is, but not necessarily in a transportable way. A state-space formulation can explicitly model observation times as informative [30]. Let $\Lambda_i(t)$ be an observation-intensity process. Then one may write a joint model in which both measurements and their arrival times depend on the latent state:

$$\begin{aligned} &\lambda_i^{\mathrm{obs}}(t) = \exp\big(\alpha_0 + \alpha^\top z_i(t)\big), \\ &N_i^{\mathrm{obs}}(t) \sim \text{Poisson process with intensity } \lambda_i^{\mathrm{obs}}(t). \end{aligned} \tag{19}$$

This structure captures the common clinical reality that sicker patients are monitored more densely. If ignored, the model may overestimate performance by using care intensity as a surrogate for leak risk.

Dynamic AL modeling also benefits from a geometric interpretation. At each time point, the posterior distribution of $z_i(t)$ defines not only a point estimate but an uncertainty ellipsoid in latent state space. Some patients occupy trajectories that move steadily toward a leak region with increasing certainty. Others wander near decision boundaries. A risk score without uncertainty can be misleading in these cases. Posterior covariance informs how robustly the model believes the patient is approaching a leak phenotype. This matters for triage. A moderate-risk estimate with low uncertainty may justify action differently from the same estimate with wide posterior spread.

Sequential decision support can be formalized through threshold policies [31]. Let $a_i(t) \in \mathcal{U}$ denote an action such as observe, intensify labs, obtain imaging, or evaluate for source control. Given a cost functional C, one seeks a policy π minimizing

$$\mathbb{E}_\pi\Big[\int_{s_i}^{\tau_i} c\big(z_i(t), a_i(t)\big)\, dt + d\big(z_i(\tau_i)\big)\Big], \tag{20}$$

where the terminal term penalizes late recognition or missed severe leak states. The full optimal control problem may be too ambitious for initial deployment, but the formulation clarifies that leak prediction is not an end in itself. It is valuable insofar as it informs time-sensitive actions. A dynamic risk surface is therefore more aligned with clinical use than a single static probability estimated once.

Early postoperative feature windows can still be embedded into this framework. An early-only model corresponds to taking $t = s_i + 24\mathrm{h}$ and estimating $\rho_i(t, h)$ from the history available up to that landmark. This is a special case, not a different problem. By placing early models inside a state-space view, one can relate them to later updated models and ask whether the dominant predictors reflect baseline risk, early physiologic response, or emerging manifestation. The distinction is important. A variable like postoperative lactate may indicate tissue stress very early, but later measurements may reflect established pathophysiology or response to treatment.

Finally, dynamic state modeling provides a principled basis for phenotype transition analysis. Rather than assigning each positive case a single static phenotype, one can estimate a trajectory over phenotype regions [32]. A patient can move from inflammatory deviation to ambiguous leak-like state to contained leak and then toward recovery after drainage. Another can move abruptly toward diffuse severe failure without passing through a prolonged ambiguous phase. These differences are clinically meaningful and analytically useful. They define not only what happened, but how it happened and how early it could have been recognized. In AL surveillance, that temporal architecture is often more informative than the endpoint alone.

6. Identifiability, Fairness, and Distribution Shift

A sophisticated label and phenotype system does not eliminate uncertainty; it makes uncertainty more explicit. In AL modeling, three sources of instability are especially important: partial identifiability of the latent event, differential ascertainment across subgroups, and distribution shift across settings and periods. These are often discussed separately, but in practice they interact. A phenotype model that is only weakly identifiable may appear unfair because subgroup observation patterns differ. A model that fails under transport may be blamed on algorithmic weakness when the real cause is a shift in leak ascertainment or denominator semantics.

Partial identifiability arises because the latent leak state is never perfectly observed in routine data. Even if one builds a hierarchical label, several plausible latent-event models can generate similar observed records [33]. A contained leak managed empirically without definitive imaging, a postoperative abscess not clearly linked to the anastomosis, and an early inflammatory deviation that resolves without source control may all occupy nearby observational neighborhoods. Without strong anchors or chart review, the analyst cannot always distinguish them. This implies that parameter estimates in a latent phenotype model are determined jointly by data and assumptions.

Suppose Y_i^* is the latent broad leak event, Z_i is the latent phenotype, and O_i is the observed operational label. Then the observed likelihood is

$$p(\mathcal{R}_i, O_i) = \sum_{y \in \{0,1\}} \sum_{k=1}^{K} p(O_i \mid y, k)\, p(\mathcal{R}_i \mid y, k)\, p(k \mid y)\, p(y). \tag{21}$$

Without restrictions on $p(O_i \mid y, k)$ or on the phenotype-conditioned record model, multiple decompositions may fit equally well. This is the essence of nonidentifiability. The practical response is not to abandon latent modeling but to constrain it using clinically credible anchors, bounded sensitivity assumptions, adjudicated subsets, or monotonicity constraints. Operative confirmation may be treated as highly specific. Certain imaging findings may be treated as moderately specific. Antibiotics alone may be treated as weak evidence. Such assumptions narrow the set of compatible latent explanations [34].

Fairness enters because ascertainment is rarely uniform across subgroups. Let G_i denote a subgroup variable. If the probability of imaging, operative exploration, or diagnostic coding differs by G_i at the same latent severity level, then the observed label O_i has different meaning across groups. One can formalize this by allowing the measurement channel to depend on subgroup:

$$\begin{aligned} \Pr(O_i = 1 \mid Y_i^* = 1, Z_i = k, G_i = g) &= \alpha_{kg}, \\ \Pr(O_i = 1 \mid Y_i^* = 0, Z_i = k, G_i = g) &= \beta_{kg}. \end{aligned} \tag{22}$$

If α_{kg} is lower in one group for milder phenotypes because those admissions are less aggressively investigated, then training on observed labels will make the model less sensitive to those phenotypes in that group. This is a fairness problem rooted in measurement inequality. Standard parity metrics computed against O_i alone may therefore understate or misstate the problem.

A subgroup-aware phenotype system helps because it allows one to ask whether the same latent region is being labeled differently across groups. Suppose two admissions occupy similar positions in the latent event graph but receive different observed labels because one underwent imaging and the other did not. This discrepancy can be studied directly as a measurement disparity. The analyst can then report not only prediction performance by subgroup but also ascertainment sensitivity by subgroup conditional on latent phenotype. Such analyses are more informative than raw disparity in false-positive or false-negative rates, since they separate prediction error from observation bias.

Published in CURRENT ISSUE

Distribution shift across sites introduces another layer [35]. Hospitals differ in procedure mix, diversion practices, antibiotic protocols, ICU use, and coding detail. These differences affect both the prior distribution of leak phenotypes and the observation process. Let environment e index site or time period. Then the joint data-generating system may be written as

$$p_e(\mathcal{R}, O, Y^*, Z) = p_e(O \mid Y^*, Z)\, p_e(\mathcal{R} \mid Y^*, Z)\, p_e(Z \mid Y^*)\, p_e(Y^*). \tag{23}$$

Shift can occur in any factor. Case-mix shift changes $p_e(Y^*)$ or $p_e(Z \mid Y^*)$. Observation shift changes $p_e(O \mid Y^*, Z)$. Feature shift changes $p_e(\mathcal{R} \mid Y^*, Z)$. A transportable phenotype framework should try to preserve invariance in the latent clinical relations while allowing the observation layer to adapt.

Domain adaptation is easier when the model architecture already distinguishes latent event from observed label. One can recalibrate the measurement channel at a new site using a small adjudicated sample without discarding the broader phenotype representation. Let θ denote shared phenotype parameters and ψ_e site-specific observation parameters. Then

$$p_e(\mathcal{R}, O, Y^*, Z) = p_{\psi_e}(O \mid Y^*, Z)\, p_\theta(\mathcal{R} \mid Y^*, Z)\, p_\theta(Z \mid Y^*)\, p_\theta(Y^*). \tag{24}$$

This separation assumes that the latent manifestation of leak is more stable than its documentation [36]. That assumption is imperfect but often more plausible than complete end-to-end transport of a black-box predictor trained on site-specific observational artifacts.

Shift can also be temporal within the same institution. Changes in imaging availability, documentation templates, or postoperative care bundles alter the record without necessarily altering the underlying leak process. A model built on administrative signals may drift badly even if physiologic patterns remain stable. A phenotype system anchored in clinical event graphs and dynamic state trajectories should be more resilient because it places less weight on surface recording conventions. Still, resilience is not guaranteed. Drift detection is required. One can monitor the empirical distribution of latent embeddings h_i or posterior phenotype vectors over time and compare them to the reference distribution using, for example, kernel divergence or transportation distance. Sharp movement in these distributions may indicate either a real change in patient population or a change in observation mechanism.

The interaction between identifiability, fairness, and shift suggests a general methodological lesson. AL modeling should not treat the observed label as an unquestioned target and then append fairness or transport analyses afterward. The observation model is itself part of the fairness and transport problem. A system can only be judged fairly if it is clear what clinical state the labels approximate and how that approximation varies by group and environment [37]. Likewise, a model can only be transported responsibly if it is clear which parts of the pipeline are expected to generalize and which require site-specific recalibration.

A practical implication is that external validation should include measurement audits, not merely performance summaries. Analysts should examine whether the same latent phenotype regions correspond to similar operative, radiographic, and coding patterns across sites and subgroups. If not, the model should be recalibrated at the observation layer or the outcome rule should be revised. Such work is analytically demanding, but it is preferable to presenting a site-wide AUC decline as if all degradation were algorithmic. In AL prediction, much of what appears as model failure is in fact label-system mismatch. Making that mismatch explicit is a prerequisite for equitable and portable deployment.

7. Validation and Decision-Oriented Use

A rigorous AL phenotype framework should be judged by whether it remains coherent under perturbation, interpretable under uncertainty, and useful within a real surveillance workflow. Validation must therefore be layered. The first layer concerns the operational label: does the hierarchical outcome rule correspond plausibly to clinically meaningful leak events, and how often does it disagree with adjudication or with strong corroborating evidence? The second layer concerns the phenotype representation: are the latent states stable, clinically ordered, and reproducible across resampling and across institutions? The third layer concerns dynamic risk output: are the probabilities calibrated in time, sufficiently early to support action, and sufficiently specific to avoid overwhelming the care environment? A single headline metric cannot answer all three questions.

For label validation, sampled chart review remains indispensable. Even if the full cohort cannot be adjudicated, small targeted review strata can be chosen from regions of high confidence, low confidence, and disagreement between evidence channels [38]. The goal is not only to estimate sensitivity and specificity but to understand which kinds of cases fail. Do false positives cluster in inflammatory mimics? Do false negatives cluster in conservatively managed contained leaks? Does delayed administrative coding rescue otherwise unrecognized true events? These qualitative patterns matter because they determine whether the operational label is narrowing the target in clinically acceptable ways or systematically excluding important phenotypes.

Phenotype validation should emphasize stability and semantic consistency. If the graph-based latent space is bootstrapped or reconstructed on temporal splits, do the same broad phenotype regions reappear? Are severity-ordered phenotypes associated with progressively greater physiologic burden or intervention intensity? Do ambiguous cases retain high posterior entropy rather than being arbitrarily absorbed into one cluster? A useful criterion is whether local neighborhoods in latent space remain clinically interpretable. If one region alternates between "contained leak" and "non-leak mimic" depending on random initialization, the phenotype system is too fragile for deployment or scientific inference.

Because leak phenotypes are not fully observed, internal validation should be supplemented by perturbation analysis. The outcome rule can be tightened or relaxed by modifying evidence thresholds, time windows, or denominator requirements. A robust phenotype system should not collapse under modest perturbation. Prevalence may change, but the broad geometry of severe, contained, ambiguous, and mimic regions should remain recognizable. This kind of semantic robustness is often more informative than a small change in conventional discrimination statistics. It indicates that the phenotype representation is tracking stable clinical structure rather than overfitting one particular operational definition.

Dynamic validation requires temporally aware metrics [39]. The surveillance task is not to detect events after they are obvious but to provide useful lead time. Therefore one should evaluate time-dependent discrimination and calibration at clinically relevant landmarks. If the system emits a high-risk signal only after imaging has already been ordered or after a reoperation note appears, then its statistical performance may be respectable while its operational value is negligible. Lead-time analysis should quantify the interval between a risk threshold crossing and the earliest strong leak confirmation. This interval, rather than static AUC, determines whether the model can plausibly alter management.

Calibration deserves special emphasis because phenotype-aware outputs are intended for decision support. A reported probability should correspond to a clinically interpretable frequency, and this should hold not only overall but within phenotype strata and subgroups. A model that is well calibrated for severe diffuse leak but overstates risk for contained leak may generate too many invasive evaluations in one region of the phenotype space while remaining apparently acceptable in aggregate. Calibration surfaces over time and phenotype are therefore more revealing than

Published in CURRENT ISSUE

ordinary reliability plots. If the system outputs both a broad leak probability and a vector of phenotype-specific risks, then both layers require calibration auditing.

Decision-oriented evaluation should also account for intervention burden. In practice, a surveillance alert can trigger repeat labs, CT scanning, bedside evaluation, prolonged hospitalization, or early source-control consideration [40]. These actions carry costs and risks. The suitable threshold is not determined solely by statistical optimization. It depends on local capacity, acceptable false-positive burden, and the relative cost of delayed recognition. Decision-curve analysis and resource-constrained simulation are useful here. They allow analysts to estimate net benefit under different thresholds and to compare phenotype-specific strategies. For instance, a lower threshold may be acceptable for the phenotype region associated with rapid severe deterioration, while a higher threshold may be preferable for ambiguous contained-leak-like states.

Interpretability is part of validation because a model that cannot be interrogated will not be used reliably. Clinicians need to know not only that a patient is high risk but also whether the model is signaling rising inflammatory burden, a trajectory toward a contained collection, or evidence consistent with a rapidly progressive diffuse leak state. The event-graph formulation makes this possible because outputs can be rendered as contributions from clinically named nodes and as position within a latent phenotype neighborhood. Such explanations do not guarantee correctness, but they permit plausibility checks and foster calibrated trust.

Prospective validation should be staged rather than immediate and fully interventional. A silent deployment phase can compare real-time phenotype trajectories and alerts against clinical events without changing care [41]. This phase reveals alert burden, calibration drift, workflow timing, and subgroup-specific failure modes. Only after the system demonstrates stable behavior should it be linked to action recommendations. Even then, the recommendations should probably be advisory and phenotype-aware rather than fully automated. In AL surveillance, the clinical context is too heterogeneous for a single rigid response rule. The role of the model is to sharpen situational awareness and support earlier, more structured evaluation.

The broader point is that validation of AL outcome definition and phenotyping should mirror the layered nature of the model. Labels need one kind of validation, latent phenotypes another, dynamic probabilities a third, and workflow integration a fourth. Compressing all of this into a single discrimination measure obscures what the system actually knows and where it is likely to fail. A decision-oriented perspective instead asks whether the model tracks the right clinical object, distinguishes the right latent states, expresses uncertainty honestly, and becomes informative early enough to matter. Those are the conditions under which a technically sophisticated phenotype framework can contribute meaningfully to perioperative care.

8. Conclusion

Outcome definition and AL phenotyping are best understood as a coupled inverse problem in perioperative data science. The latent clinical event is not directly recorded, and the observed postoperative record is a selective projection shaped by physiology, care processes, timing, and documentation [42]. For that reason, the binary leak label should not be treated as a self-evident target. It should be treated as an operational construction whose evidentiary hierarchy, temporal scope, and denominator logic are explicit and auditable. Once this is acknowledged, phenotyping is no longer optional decoration. It becomes the mechanism by which clinically important heterogeneity is preserved instead of being collapsed into a single event flag.

A hierarchical event framework offers a practical way to organize this complexity. It allows weak precursors, suspicion signals, corroborating interventions, and administrative abstractions to contribute differently to label formation. A graph-based latent representation then provides a structured phenotype space in which severe diffuse

failure, contained leak, intervention-defined leak, and ambiguous leak-like states can be represented with uncertainty rather than forced into brittle categories. Dynamic state-space modeling extends the same logic over time, supporting landmarked surveillance and phenotype-specific forward risk rather than static retrospective labeling alone.

Such a system does not eliminate error. It instead exposes where error enters: in latent-state identifiability, in subgroup-dependent ascertainment, and in cross-site observation shift. Those sources of instability must be incorporated into validation and deployment rather than treated as afterthoughts. The practical value of AL prediction depends on whether it identifies the right postoperative state, distinguishes meaningful event expressions, and produces interpretable risk estimates early enough to support action. A phenotype-aware framework is well suited to that task because it treats leak not as a single recorded fact but as a family of postoperative failure trajectories observed through imperfect clinical systems [43].

References

[1] Y. Dai, S. Chopra, S. Kneif, and M. Hünerbein, "Management of esophageal anastomotic leaks, perforations, and fistulae with self-expanding plastic stents," *The Journal of thoracic and cardiovascular surgery*, vol. 141, no. 5, pp. 1213–1217, 12 2010.

[2] J. Roy, K. D. Sims, P. Rider, L. Grimm, J. Hunter, and W. G. Richards, "Endoscopic technique for closure of enterocutaneous fistulas," *Surgical endoscopy*, vol. 33, no. 10, pp. 3464–3468, 1 2019.

[3] J. Tschmelitsch, H. Wykypiel, R. Prommegger, and E. Bodner, "Colostomy vs tube cecostomy for protection of a low anastomosis in rectal cancer." *Archives of surgery (Chicago, Ill. : 1960)*, vol. 134, no. 12, pp. 1385–1388, 12 1999.

[4] P. Ortega-Deballon, F. Radais, O. Facy, P. d'Athis, D. Masson, P. E. Charles, N. Cheynel, J.-P. Favre, and P. Rat, "C-reactive protein is an early predictor of septic complications after elective colorectal surgery," *World journal of surgery*, vol. 34, no. 4, pp. 808–814, 1 2010.

[5] M. Y. Liu, H.-C. Tang, S.-H. Hu, H.-L. Yang, and S. J. Chang, "Influence of preoperative peripheral parenteral nutrition with micronutrients after colorectal cancer patients," *BioMed research international*, vol. 2015, pp. 535 431–535 431, 4 2015.

[6] P. Scognamiglio, M. Reeh, N. Melling, M. Kantowski, A.-K. Eichelmann, S.-H. Chon, N. El-Sourani, G. Schön, A. Höller, J. R. Izbicki, and M. Tachezy, "Management of intra-thoracic anastomotic leakages after esophagectomy: updated systematic review and meta-analysis of endoscopic vacuum therapy versus stenting." *BMC surgery*, vol. 22, no. 1, pp. 309–309, 8 2022.

[7] D. B. Evans, P. W. Pisters, J. E. Lee, R. J. Bold, C. Charnsangavej, N. A. Janjan, R. A. Wolff, and J. L. Abbruzzese, "Preoperative chemoradiation strategies for localized adenocarcinoma of the pancreas," *Journal of hepato-biliary-pancreatic surgery*, vol. 5, no. 3, pp. 242–250, 11 1998.

[8] 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.

[9] C. K, A. T. Jacob, M. Muralee, W. M. Sudam, M. L, and S. Balakrishnan, "Low anterior resection syndrome and quality of life of patients post sphincter preservation surgery: A prospective study," 7 2023.

Published in CURRENT ISSUE

[10] J. Ma, H. Hu, J. Xiao, Y. Zhao, X. Chen, S. Wei, and H. Zhang, "Esophageal covered stent treatment for gastroesophageal cervical anastomotic fistula," *Chinese journal of radiology*, vol. 53, no. 5, pp. 385–388, 5 2019.

[11] I. Kamaleddine, A. Hendricks, M. Popova, and C. Schafmayer, "Adequate management of postoperative complications after esophagectomy: A cornerstone for a positive outcome." *Cancers*, vol. 14, no. 22, pp. 5556–5556, 11 2022.

[12] R. M. Dahl, J. Wetterslev, L. N. Jorgensen, L. S. Rasmussen, A. M. Møller, and C. S. Meyhoff, "The association of perioperative dexamethasone, smoking and alcohol abuse with wound complications after laparotomy," *Acta anaesthesiologica Scandinavica*, vol. 58, no. 3, pp. 352–361, 1 2014.

[13] R. Mirnezami, A. Rohatgi, R. P. Sutcliffe, A. Hamouda, and R. Mason, "Transhiatal oesophagectomy: treatment of choice for high-grade dysplasia," *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery*, vol. 36, no. 2, pp. 364–367, 5 2009.

[14] F. S. Mari, M. Gasparrini, G. Nigri, G. Berardi, G. G. Laracca, B. Flora, A. Pancaldi, and A. Brescia, "Can a curved stapler made for open surgery be useful in laparoscopic lower rectal resections? technique and experience of a single centre," *The surgeon : journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*, vol. 11, pp. S23–6, 11 2012.

[15] G. A. Prevost, M. Odermatt, M. Furrer, and P. M. Villiger, "Postoperative morbidity of complete mesocolic excision and central vascular ligation in right colectomy: a retrospective comparative cohort study," *World journal of surgical oncology*, vol. 16, no. 1, pp. 214–214, 10 2018.

[16] I. Pasternak, M. Dietrich, R. J. Woodman, U. Metzger, D. Wattchow, and U. Zingg, "Use of severity classification systems in the surgical decision-making process in emergency laparotomy for perforated diverticulitis." *International journal of colorectal disease*, vol. 25, no. 4, pp. 463–470, 11 2009.

[17] Y. Mazni, A. Syafiuddin, and A. S. Putranto, "Intraoperative pancreatic assessment in pancreaticoduodenectomy the correlation with pancreatic fistula formation," *The New Ropanasuri : Journal of Surgery*, vol. 5, no. 1, pp. 12–15, 6 2020.

[18] C. G. Segal, D. K. Waller, B. C. Tilley, L. B. Piller, and K. Y. Bilimoria, "An evaluation of differences in risk factors for individual types of surgical site infections after colon surgery," *Surgery*, vol. 156, no. 5, pp. 1253–1260, 8 2014.

[19] P. Rai, S. S. Johnston, R. Chaudhuri, E. Naoumtchik, and E. Pollack, "Association of complications with healthcare utilization and hospital-borne costs among patients undergoing open low anterior resection using curved cutter staplers." *Medical devices (Auckland, N.Z.)*, vol. 14, pp. 87–95, 3 2021.

[20] M. K. Ismael and A. H. M. Al-Azzawi, "Comparison study stapled versus hand sewn method for large bowel anastomosis surgery," *International Journal of Surgery Science*, vol. 4, no. 4, pp. 164–168, 10 2020.

[21] S. Vural, O. Civil, M. Kement, Y. E. Altuntas, N. Okkabaz, C. Gezen, M. C. Haksal, E. Gundogan, and M. Oncel, "Risk factors for early postoperative morbidity and mortality in patients underwent radical surgery for gastric carcinoma: a single center experience." *International journal of surgery (London, England)*, vol. 11, no. 10, pp. 1103–1109, 9 2013.

[22] A. Saleh, U. Ihedioha, B. Babu, J. Evans, and P. Kang, "Is estimated intra-operative blood loss a reliable predictor of surgical outcomes in laparoscopic colorectal cancer surgery?" *Scottish medical journal*, vol. 61, no. 3, pp. 167–170, 7 2015.

[23] S. Satoskar, S. Kashyap, F. Benavides, R. Jones, R. Angelico, and V. Singhal, "Success of endoscopic vacuum therapy for persistent anastomotic leak after esophagectomy - a case report." *International journal of surgery case reports*, vol. 80, no. C, pp. 105 342–, 11 2020.

[24] 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, no. 7, pp. 501–503, 6 2008.

[25] T. A. Bowles, K. M. Sanders, M. E. Colson, and D. A. K. Watters, "Simplified risk stratification in elective colorectal surgery." *ANZ journal of surgery*, vol. 78, no. 1-2, pp. 24–27, 1 2008.

[26] D. S. Nirhale, A. A. Ghalsasi, and V. Nisarga, "Assessment of nutritional status, pre-operative nutrition supplementation and its' impact on the outcome of surgery in gastrointestinal malignancies: a prospective study," *International Surgery Journal*, vol. 7, no. 1, pp. 178–, 12 2019.

[27] P. Albers, S. Schäfers, H. Löhmer, and P. de Geeter, "Seminal vesicle-sparing perineal radical prostatectomy improves early functional results in patients with low-risk prostate cancer." *BJU international*, vol. 100, no. 5, pp. 1050–1054, 8 2007.

[28] A. Spinelli, M. Carvello, P. G. Kotze, A. Maroli, I. Montroni, M. Montorsi, N. Buchs, and F. Ris, "Ileal pouch-anal anastomosis with fluorescence angiography: a case-matched study." *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 21, no. 7, pp. 827–832, 4 2019.

[29] D. W. Good, J. M. O'Riordan, D. Moran, F. B. V. Keane, E. Eguare, D. S. O'Riordain, and P. Neary, "Laparoscopic surgery for rectal cancer: a single-centre experience of 120 cases," *International journal of colorectal disease*, vol. 26, no. 10, pp. 1309–1315, 6 2011.

[30] A. D. Politano, T. Hranjec, L. H. Rosenberger, R. G. Sawyer, and C. A. T. Leon, "Differences in morbidity and mortality with percutaneous versus open surgical drainage of postoperative intra-abdominal infections: A review of 686 cases," *The American surgeon*, vol. 77, no. 7, pp. 862–867, 7 2011.

[31] M. Sohn and A. Agha, "Preservation of the superior rectal artery," *coloproctology*, vol. 40, no. 1, pp. 42–46, 12 2017.

[32] J. Owono-Mbouengou, D. Ngabou, D. Folly, M. Essomo-Megnier-Mbo, H. Nyamatiengui, and R. Nguema-Mve, "Distal ileal necrosis: Right ileo-colic intussuscepted anastomosis as an alternative to ileostomy," *Journal of visceral surgery*, vol. 151, no. 5, pp. 341–346, 6 2014.

[33] S. Kusamura, "Peer review report for: Hyperthermic intraperitoneal chemoperfusion with high dose oxaliplatin: Influence of perfusion temperature on postoperative outcome and survival [version 2; peer review: 1 approved, 2 approved with reservations]," 11 2015.

[34] N. Hyman, T. M. Osler, P. A. Cataldo, E. H. Burns, and S. R. Shackford, "Anastomotic leaks after bowel resection: what does peer review teach us about the relationship to postoperative mortality?" *Journal of the American College of Surgeons*, vol. 208, no. 1, pp. 48–52, 11 2008.

[35] M. Lanuti, P. E. de Delva, C. D. Wright, H. A. Gaissert, J. C. Wain, D. M. Donahue, J. S. Allan, and D. J. Mathisen, "Post-esophagectomy gastric outlet obstruction: role of pyloromyotomy and management with endoscopic pyloric dilatation," *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery*, vol. 31, no. 2, pp. 149–153, 12 2006.

[36] C. lue Mei, M. Huang, and X. you Zhang, "The risk factors for anastomotic leaks due to colorectal surgery," *International Medicine and Health Guidance News*, vol. 18, no. 21, pp. 3111–3113, 11 2012.

Published in CURRENT ISSUE

[37] G. K. Weston-Petrides, R. E. Lovegrove, H. S. Tilney, A. G. Heriot, R. J. Nicholls, N. Mortensen, V. W. Fazio, and P. P. Tekkis, "Comparison of outcomes after restorative proctocolectomy with or without defunctioning ileostomy." *Archives of surgery (Chicago, Ill. : 1960)*, vol. 143, no. 4, pp. 406–412, 4 2008.

[38] T. D. Lyon, S. A. Boorjian, P. Shah, R. F. Tarrell, J. C. Cheville, I. Frank, R. J. Karnes, R. H. Thompson, and M. K. Tollefson, "Comprehensive characterization of perioperative reoperation following radical cystectomy." *Urologic oncology*, vol. 37, no. 4, pp. 292.e11–292.e17, 1 2019.

[39] R. Peltrini, M. Podda, S. Castiglioni, M. M. D. Nuzzo, M. D'Ambra, R. Lionetti, M. Sodo, G. Luglio, F. Mucilli, S. D. Saverio, U. Bracale, and F. Corcione, "Intraoperative use of indocyanine green fluorescence imaging in rectal cancer surgery: The state of the art," *World journal of gastroenterology*, vol. 27, no. 38, pp. 6374–6386, 10 2021.

[40] D. J. Aaron, A. Anandhi, G. S. Sreenath, S. Sureshkumar, O. Shaikh, V. Balasubramaniyan, and V. Kate, "Serial estimation of serum c-reactive protein and procalcitonin for early detection of anastomotic leak after elective intestinal surgeries: a prospective cohort study." *Turkish journal of surgery*, vol. 37, no. 1, pp. 22–27, 3 2021.

[41] N. Akhtar, "Iddf2018-abs-0242 sleeve omentopexy over pancreatico jejunostomy – a new technique," *Clinical Gastroenterology*, vol. 67, pp. A78.2–A78, 6 2018.

[42] K. Epari and R. J. Cade, "Oesophagectomy for tumours and dysplasia of the oesophagus and gastro-oesophageal junction," *ANZ journal of surgery*, vol. 79, no. 4, pp. 251–257, 4 2009.

[43] S. Awad, R. Aguilo, S. Agrawal, and J. Ahmed, "Outcomes of linear-stapled versus hand-sewn gastrojejunal anastomosis in laparoscopic roux en-y gastric bypass," *Surgical endoscopy*, vol. 29, no. 8, pp. 2278–2283, 11 2014.