Peer reviewed Journals
seron press

RESEARCH

Formal Outcome Definition and Computational Phenotyping of Anastomotic Leak in Electronic Health Records and Perioperative Decision Systems

Gabriel Sousa,[1] Matheus Oliveira,[2] and Felipe Nogueira[3]

[1]Departamento de Informática em Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre, Rua Sarmento Leite 245, Centro Histórico, Porto Alegre, Brazil
[2]Departamento de Engenharia Biomédica, Universidade Federal do ABC, Avenida dos Estados 5001, Bangú, Santo André, Brazil
[3]Departamento de Tecnologia em Sistemas Biomédicos, Universidade Estadual de Ciências da Saúde de Alagoas, Rua Doutor Jorge de Lima 113, Trapiche da Barra, Maceió, Brazil

Abstract

Outcome definition in perioperative data science is often treated as a preliminary coding task, yet for anastomotic leak it is the central methodological problem that governs label validity, error structure, transportability, and clinical interpretation. Anastomotic leak is not a single biochemical state, nor a single radiographic sign, nor a single billing event. It is a time-evolving postoperative failure process that may appear as localized contamination, systemic inflammatory deterioration, occult radiologic disruption, or operative reintervention, with each manifestation entering the record through different channels and at different delays. That heterogeneity makes binary classification deceptively simple and often unstable. This paper develops a technical framework for outcome definition and anastomotic leak phenotyping that begins from event semantics rather than algorithm choice. The discussion treats leak labels as functions of temporally indexed evidence streams, including diagnosis codes, procedures, laboratory trajectories, imaging triggers, antimicrobial escalation, and operative context. A phenotype is then defined as a probabilistic state over clinically coherent leak expressions, rather than a rigid mutually exclusive category. The paper formalizes composite outcome assembly, temporal anchoring, latent phenotype inference, uncertainty propagation, and subgroup-sensitive validation. It further argues that the mathematical object being predicted should be separated from the operational rule used to identify it, because the latter is only an imperfect measurement process. Under this view, robust AL modeling depends on constructing labels that are auditable, temporally consistent, and transportable across hospitals, while preserving clinically meaningful distinctions between contained, occult, severe, and management-defined leak states.

1. Introduction

Anastomotic leak occupies an unusual position in postoperative modeling because the event is clinically consequential, physiologically heterogeneous, and poorly represented by any single data element [1]. The inferential burden is therefore shifted backward from the prediction algorithm toward the definition of the target itself. A classifier can only be as coherent as the object it is trained to identify, and for AL that object is rarely observed directly. What becomes available to a data system is a delayed and partial trace consisting of procedure records, postoperative laboratories, radiographic decisions, free-text documentation, and, in some cases, a diagnostic code that stands in for a far more complex clinical judgment. The result is that outcome definition is not merely annotation. It is a modeling act that determines which biological failures are considered equivalent, which management pathways count as confirmatory evidence, which time windows are treated as etiologically relevant, and which forms of uncertainty are ignored or preserved.

A recent reproducible EHR-based study used a composite AL outcome requiring intestinal wound disruption or postprocedural fistula coding together with corroborating procedural or antimicrobial evidence within a 30-day window, observed an event prevalence of 7.6%, and found that early postoperative inflammatory variables dominated predictive performance, as Sadanandan (2024) [2]. That formulation makes explicit a core fact that many surgical prediction studies leave implicit: the leak state is not directly encoded in the record, and therefore must be reconstructed from heterogeneous evidence. Once this reconstruction is acknowledged, the question of AL phenotyping becomes unavoidable. A patient with a small contained pelvic collection managed by antibiotics alone is not clinically

Table 1. Outcome Definition Design Elements

Element	Description	Options	Impact
Anchor Time	Start reference	Surgery, admission	Temporal alignment
Horizon	Observation window	7, 14, 30 days	Sensitivity vs specificity
Evidence Rule	Data combination	Code, intervention	Label validity
Precedence Logic	Rule structure	Conjunctive, weighted	Interpretability

Table 2. Evidence Weighting Scheme for Composite Labels

Evidence Type	Indicator	Weight w_j	Reliability
Reoperation	R_i	High	Very specific
Drainage	D_i	Medium-high	Moderately specific
Imaging	I_i	Medium	Context dependent
Antibiotics	A_i	Low	Non-specific

identical to a patient with diffuse peritonitis requiring urgent reoperation, even if both satisfy a broad binary definition. A data system that collapses these two trajectories into a single label may improve sample size while degrading mechanistic resolution.

The technical literature on postoperative risk prediction has often privileged discrimination metrics over target formulation [3]. Such a tendency is understandable. Binary outcomes are convenient, widely supported by conventional algorithms, and easy to report. Yet in surgical complications, especially AL, this convenience can obscure an important asymmetry. The biological phenomenon is continuous, temporally unfolding, and multiscale, whereas the recorded outcome is discontinuous, delayed, and institution-dependent. If two hospitals differ in when they image, when they reoperate, how they code wound disruption, or how aggressively they escalate antibiotics, then they do not merely differ in case mix. They differ in the observation operator through which AL becomes visible. A model trained on one setting may therefore fail in another not because the biology changed, but because the mapping from latent leak state to observed record changed.

For this reason, outcome definition and phenotyping should be treated jointly. Outcome definition answers whether the data support classification of an admission as having experienced an AL event within a specified horizon. Phenotyping answers which subtype, severity regime, temporal trajectory, and evidentiary pathway characterize that event. The two tasks are mathematically linked. A noisy or overly coarse outcome definition reduces the separability of phenotype clusters [4]. Conversely, a well-structured phenotype model can inform better outcome definitions by revealing that some observed labels represent coding artifacts, postoperative inflammation without leak, or management choices that are not leak-specific. In this sense, phenotyping is not a downstream descriptive add-on. It is part of the label construction problem itself.

This paper develops a technical framework for thinking about AL labels and phenotypes in electronic health records. The argument proceeds from first principles. The central object is a latent postoperative failure process evolving after an index anastomosis. What the record contains is a set of noisy, temporally irregular, and selectively missing observations. A formal outcome definition is then a measurable map from those observations to an event indicator. A phenotype is a structured representation of within-event heterogeneity, expressed either as a discrete latent class, a continuous coordinate on a low-dimensional manifold, or a probability vector over several clinically meaningful leak states. Under this formulation, model development becomes an exercise in measurement theory, state estimation, and transportability rather than in raw pattern recognition alone.

The emphasis here is deliberately technical. The paper does not assume that AL can be perfectly observed or that a single binary endpoint is always sufficient [5]. Instead, it treats leakage as a latent construct whose manifestations depend on anatomy, timing, care processes, and observation intensity. The practical goal is not to multiply categories for their own sake. It is to define an outcome and phenotype system that supports three demands simultaneously: internal coherence for training models, external comparability across institutions, and clinical interpretability for perioperative decision making. These demands are often in tension. Narrow rules increase precision but reduce sensitivity and portability. Broad rules improve coverage but introduce semantic drift. Latent phenotypes can recover structure but may become unstable if not anchored in explicit event logic. The sections that follow formalize these trade-offs and propose a framework in which outcome definition and AL phenotyping are designed as coupled components of a single analytical system.

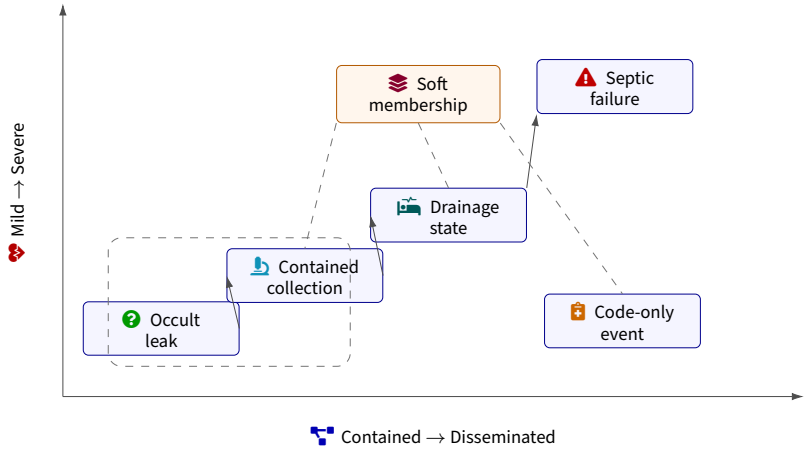


Figure 1. The phenotype space is organized by severity and containment rather than by a single event flag. Soft membership captures admissions that lie near boundaries between clinically adjacent leak states, preserving uncertainty instead of forcing unstable hard assignment.

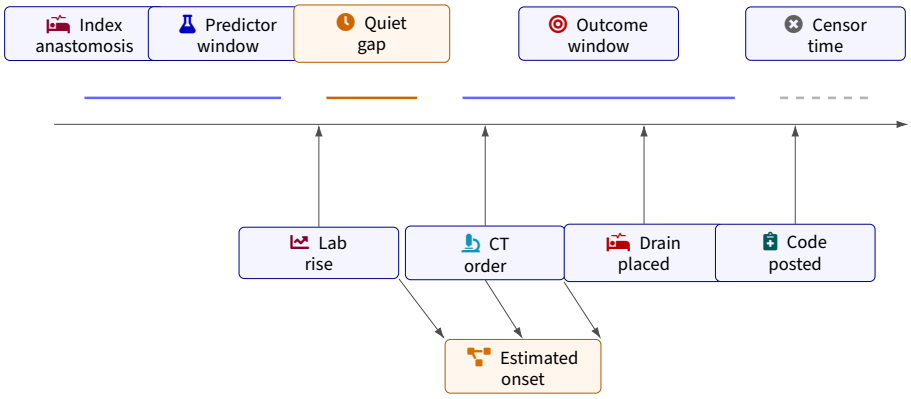


Figure 2. Temporal assignment is shown as ordered windows rather than a single retrospective flag. A predictor interval is separated from the leak horizon by a quiet gap, while observable triggers are combined into an estimated onset distinct from the unobserved physiological failure time.

2. Outcome Definition as a Statistical Object

A useful starting point is to distinguish the latent clinical state from the observed label. Let patient-admission i undergo an index anastomotic procedure at time s_i. Denote by $L_i(t)$ an unobserved stochastic process representing the true physiological integrity of the anastomosis over postoperative time $t \geq s_i$. One may interpret $L_i(t)$ as a hidden state that evolves from stable healing toward failure under influences such as ischemia, tension, contamination, nutritional impairment, immunologic dysregulation, and technical factors [6]. The object of clinical interest is not any single laboratory value or code, but whether the trajectory enters a regime of clinically meaningful dehiscence. The EHR does not provide $L_i(t)$ directly. Instead, it provides coded diagnoses, interventions, laboratory series, imaging orders, note mentions, and antimicrobial administrations, each generated by a different recording mechanism. The observed outcome is therefore a functional of these proxies.

Suppose the observed record for admission i consists of a multistream temporal object

$$\mathcal{O}_i = \Big(\mathcal{C}_i(t), \mathcal{P}_i(t), \mathcal{M}_i(t), \mathcal{V}_i(t), \mathcal{N}_i(t), \mathcal{I}_i(t)\Big), \quad t \in [s_i, s_i + \tau_i], \tag{1}$$

where $\mathcal{C}_i$ denotes diagnosis code events, $\mathcal{P}_i$ procedure and reintervention events, $\mathcal{M}_i$ medications including antimicrobial escalation, $\mathcal{V}_i$ vitals and laboratory measures, $\mathcal{N}_i$ text-derived indicators, and $\mathcal{I}_i$ imaging-related observations. An operational outcome definition is a measurable map g such that $Y_i^{\text{obs}} = g(\mathcal{O}_i) \in \{0, 1\}$. In practice, g is almost always composite. It includes a temporal horizon, a set of admissible evidentiary channels, and a logical rule specifying how those channels are combined. The important methodological point is that g is not a neutral translation of chart reality. It is a designed estimator of an unobserved target.

The simplest rule is code-based labeling, where $Y_i^{\text{obs}} = 1$ if a designated diagnosis code appears within some postoperative interval. This strategy is attractive for

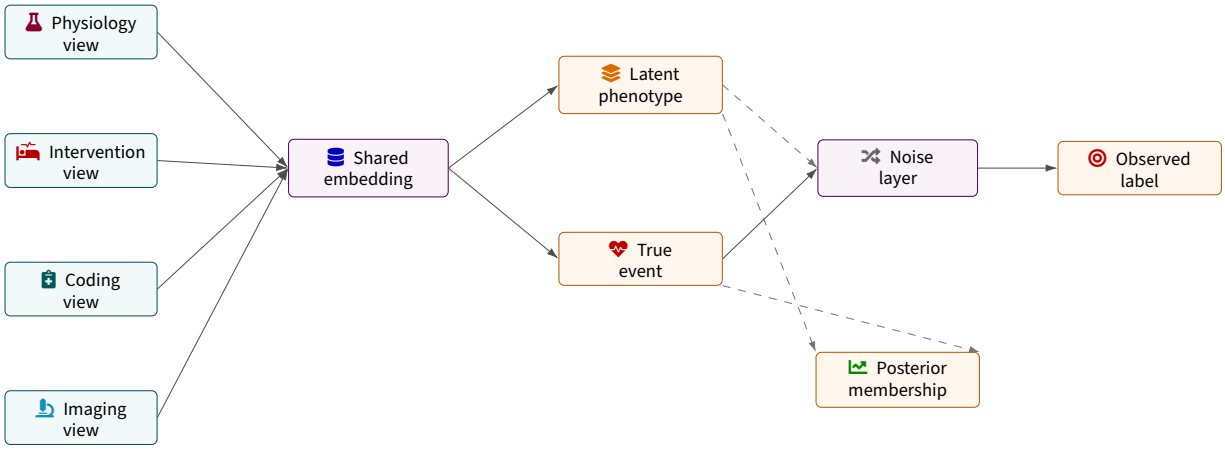


Figure 3. A latent-variable phenotyping model integrates multiple evidence views into a shared representation, infers latent subtype structure and latent event status, and then maps those hidden variables through a noise layer to the imperfect observed label used in routine datasets.

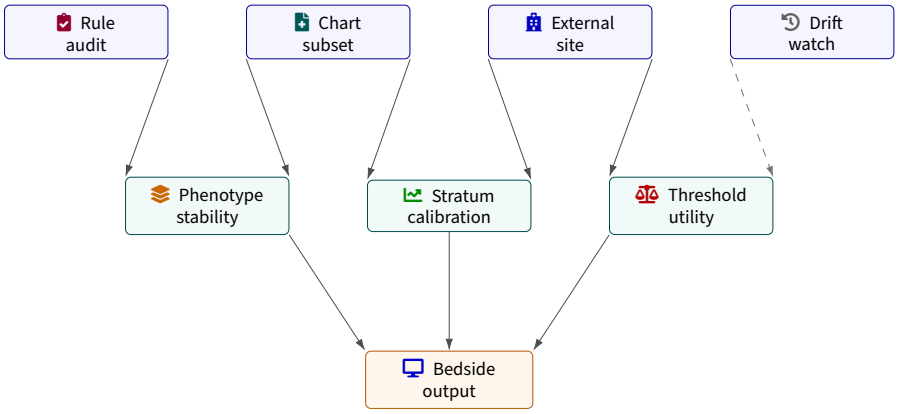


Figure 4. Validation is staged from rule auditing to chart review, external-site testing, drift monitoring, phenotype stability, calibration within strata, and threshold utility, ending in an interpretable bedside-facing output rather than an opaque score alone.

scale but fragile in meaning. Coding depends on documentation habits, billing incentives, coder interpretation, and timing conventions that vary across institutions. A more defensible formulation introduces corroboration [7]. Let C_i indicate presence of a leak-relevant diagnosis code within horizon H, and let E_i indicate supporting clinical evidence such as reoperation, drainage, or targeted antimicrobial escalation within the same horizon. Then one may define

$$\begin{aligned} Y_i^{\text{obs}} &= \mathbf{1}\{C_i = 1 \ \wedge \ E_i = 1\}, \\ E_i &= \mathbf{1}\{R_i + D_i + A_i \geq 1\}, \end{aligned} \tag{2}$$

where R_i, D_i, and A_i are binary indicators for return to the operating room, drainage intervention, and antimicrobial evidence. This rule reduces coding-only false positives, but it may exclude conservatively managed leaks or leaks recognized clinically without immediate intervention. Thus, every gain in specificity alters the target population. The label becomes increasingly management-defined.

The core technical problem is therefore not choosing a universally correct rule, because no such rule exists across all datasets. The problem is specifying a rule whose error structure is legible. Let Y_i^* denote the latent clinically true event, defined by expert judgment under ideal information. The operational label satisfies

$$\begin{aligned} \Pr(Y_i^{\text{obs}} = 1 \mid Y_i^* = 1) &= \text{Se}_i, \\ \Pr(Y_i^{\text{obs}} = 0 \mid Y_i^* = 0) &= \text{Sp}_i, \end{aligned} \tag{3}$$

where sensitivity and specificity may vary by institution, phenotype, and observation intensity [8]. If contained leaks are less likely to trigger reoperation, then the composite rule has phenotype-dependent sensitivity. If septic complications unrelated to leak trigger similar antibiotics and drainage, then specificity may fall in high-acuity settings. The observed binary label is thus generated through a misclassification operator whose parameters are not constant. This observation has major consequences. A model trained on Y_i^{obs} is

Table 3. Phenotype Axes for Leak Characterization

Axis	Symbol	Description	Clinical Meaning
Timing	τ^{onset}	Early vs late	Mechanism differences
Containment	γ^{contain}	Local vs diffuse	Anatomical spread
Physiology	η^{phys}	Inflammatory burden	Severity level
Management	μ^{manage}	Intervention intensity	Care pathway

Table 4. Discrete Leak Phenotype Categories

Phenotype	Description	Evidence Pattern	Severity
Occult	Radiologic only	Imaging + mild labs	Low
Contained	Local collection	Drainage + imaging	Moderate
Inflammatory	Systemic signs	Labs + antibiotics	Moderate
Operative	Reoperation required	Surgery evidence	High
Uncertain	Weak signals	Code only	Variable

not learning only biological propensity for leak. It is also learning the institutional process by which leak becomes observable.

A rigorous outcome definition should therefore state four ingredients explicitly. The first is anchor time. For AL, the natural anchor is the timestamp of index anastomosis rather than admission or discharge. The second is event horizon. A 30-day window is common because it captures clinically relevant postoperative failure while limiting unrelated later pathology, but shorter windows may be preferable for early detection studies and longer windows for surveillance registries. The third is evidence composition. This includes whether the label requires diagnostic coding, intervention, imaging confirmation, note extraction, or combinations thereof [9]. The fourth is precedence logic. One must specify whether evidence channels are treated conjunctively, disjunctively, or hierarchically. For example, operative confirmation may dominate all other evidence, whereas antibiotics alone may be insufficient without additional supporting data.

These design elements can be summarized in a formal rule with precedence weights. Let u_{ij} denote the presence of evidence type j in a specified time window and w_j a credibility weight reflecting the expected positive predictive value of that evidence source. Then a thresholded score

$$S_i = \sum_{j=1}^{m} w_j u_{ij},$$
$$Y_i^{\text{obs}} = \mathbf{1}\{S_i \geq \kappa\}, \tag{4}$$

permits graded definitions. A reoperation with intraoperative contamination may receive higher weight than isolated imaging suspicion, and isolated imaging suspicion higher weight than antibiotic escalation alone. This formulation makes the rule auditable and tunable. It also opens the possibility of uncertainty-aware labels. Instead of forcing every case into a binary state, one may retain S_i or a normalized posterior probability as a measure of label confidence.

A further complication arises from denominator definition. The at-risk set is not simply all colorectal operations [10]. It depends on whether an anastomosis was actually fashioned, whether diversion occurred, whether procedures were staged, and whether the record contains enough post-index observation time. Including non-anastomotic resections will dilute event prevalence and blur biological relevance. Excluding complex salvage surgeries may improve coherence but reduce representativeness. This is not a trivial cohorting issue. Outcome validity depends on denominator semantics. A formally defined label over an ambiguously defined at-risk population remains unstable.

Finally, outcome definition should be evaluated not only by face validity but by inferential consequences. A well-designed AL label ought to exhibit temporal plausibility, agreement with chart review in sampled subsets, expected associations with severity markers, and robustness under reasonable perturbations of the evidence rule. If small changes in antibiotic or drainage logic dramatically alter prevalence and feature importance, then the label is not stable enough to support downstream modeling without additional adjudication. In this sense, outcome definition is a statistical object whose quality can be interrogated through sensitivity analysis, latent error models, and transport tests. The question is not whether the definition is perfect, but whether its imperfections are sufficiently structured to permit meaningful prediction and phenotyping.

3. Clinical Event Semantics and the AL Phenotype Space

Once the binary outcome is recognized as a compressed measurement of a more complex process, the need for phenotyping follows naturally [11]. A phenotype should be understood as a structured summary of how leak manifests, not merely whether it occurs. The relevant

Table 5. Temporal Labeling Strategies

Strategy	Definition	Advantage	Limitation
Fixed Window	$[0, H]$	Simple	Ignores timing
Landmarking	$(u, u + h]$	Dynamic risk	Complex setup
Earliest Event	$\min(T_i)$	Clear trigger	Bias in ordering
Multi-state	State transitions	Rich structure	Data intensive

Table 6. Latent Variable Modeling Approaches

Model Type	Representation	Strength	Limitation
Mixture Model	Discrete Z_i	Interpretability	Rigid classes
Multi-view Model	Factorized views	Data integration	Complexity
HMM/Semi-Markov	Temporal states	Dynamics	Parameter heavy
Latent Trait	Continuous h_i	Smooth variation	Less discrete meaning

heterogeneity extends across timing, anatomical containment, systemic physiological burden, and management intensity. Some leaks are occult and identified radiographically after a mild inflammatory drift. Others present with abrupt hemodynamic deterioration and urgent operative source control. Some are limited to a small pelvic cavity and treated with drainage, whereas others reflect diffuse disruption of a high-risk anastomosis. These trajectories share a common failure domain but differ in clinical expression and data signature. A useful phenotype system must preserve that structure without fragmenting the cohort into unstable microclasses.

Let the latent phenotype of admission i be represented by Z_i. In a discrete formulation, $Z_i \in \{1, \ldots, K\}$ indexes mutually exclusive states such as occult contained leak, inflammatory leak without reoperation, image-confirmed pelvic leak with drainage, fulminant generalized leak requiring urgent surgery, and uncertain code-supported event. In a soft formulation, Z_i is replaced by a vector $\pi_i \in \Delta^{K-1}$ on the probability simplex, where π_{ik} gives the posterior membership probability in phenotype k. The soft representation is often preferable because phenotype boundaries are rarely sharp. A patient may exhibit features of both contained pelvic leak and evolving systemic compromise. The phenotype is then not a fixed label but a distribution over nearby clinically plausible states.

The phenotype space should be built from axes that are semantically meaningful and computationally observable [12]. One axis is temporal onset. A very early leak, emerging within the first few postoperative days, often reflects technical or perfusion-related failure, while later presentations may involve progressive tissue breakdown, occult abscess formation, or delayed recognition. Another axis is anatomical containment. A leak can remain locally confined or progress to free intraperitoneal contamination. A third axis is physiologic severity, captured by inflammatory and organ-dysfunction responses. A fourth axis is intervention dependency, reflecting whether the leak phenotype is known only because a certain management decision was taken. A fifth axis is evidence modality, separating events that are primarily code-defined, imaging-defined, operative, or text-derived. Together, these axes create a multidimensional phenotype geometry.

One may formalize this geometry by defining a feature map ϕ_i from observed data into a latent coordinate system,

$$\phi_i = \left(\tau_i^{\text{onset}}, \gamma_i^{\text{contain}}, \eta_i^{\text{phys}}, \mu_i^{\text{manage}}, \epsilon_i^{\text{evidence}}\right)^\top, \tag{5}$$

where each coordinate is either directly observed or estimated from the record. The phenotype space is then a partition or tessellation of the image of ϕ_i. Importantly, these coordinates are not merely predictors [13]. They are definitional descriptors of the event itself. For example, a rising leukocyte trajectory used to predict future leak is different from the magnitude of systemic response after leak declaration. The same variable may belong either in the predictor set or in the phenotype descriptor set depending on temporal relation to the event anchor.

A careful phenotyping scheme must also distinguish biological phenotype from ascertainment phenotype. Consider two patients with similar contained pelvic leaks. One undergoes early computed tomography because the service has a low imaging threshold; the other is observed clinically and treated empirically. In a purely data-driven clustering algorithm, these patients may separate because one has imaging evidence and the other does not. Yet the difference lies in observation policy, not in pathophysiology. To prevent such confounding, evidence modality should be modeled as one axis among others, not allowed to dominate the clustering objective. Put differently, phenotypes should remain clinically interpretable even when observation intensity varies.

This distinction suggests a layered ontology [14].

Table 7. Sources of Error in AL Phenotyping

Error Type	Source	Effect	Example
Misclassification	Coding variability	Label noise	Missing mild leaks
Missing Data	Sparse labs	Bias	Unmeasured severity
Observation Bias	Monitoring intensity	Confounding	Sicker patients measured more
Institutional Variation	Practice patterns	Transport issues	Imaging thresholds

Table 8. Validation Dimensions for Phenotype Systems

Dimension	Metric	Purpose	Insight
Discrimination	AUC	Separation	Predictive power
Calibration	Reliability curves	Accuracy	Risk validity
Stability	Bootstrap consistency	Robustness	Phenotype reliability
Transportability	External testing	Generalization	Site invariance

At the deepest layer lies structural failure of the anastomosis. Above it lies physiological expression, such as systemic inflammation or organ dysfunction. Above that lies management response, including imaging, drainage, reoperation, and antimicrobial escalation. The EHR primarily records the upper layers. A phenotype model that treats all recorded events as equivalent may inadvertently classify treatment pathways instead of disease states. The remedy is not to remove management data entirely, because management contains genuine severity information. The remedy is to model the hierarchy explicitly. An urgent reoperation is not simply another feature. It is a downstream manifestation whose relationship to the latent state passes through surgeon judgment, resource availability, and institutional practice.

A mathematically useful consequence of this layered view is that phenotype membership can be regularized toward clinical monotonicity. If phenotype k_1 is conceptually milder than phenotype k_2, then posterior membership shifts from k_1 to k_2 should align with greater physiological burden or more invasive intervention. This may be enforced through ordered latent variables or partial order constraints on phenotype centroids [15]. Such constraints are especially useful when unsupervised clustering yields unstable or semantically mixed groups. Purely geometric separation in feature space is not enough. A valid AL phenotype should remain consistent under clinical interpretation.

Soft phenotyping also provides a natural way to represent indeterminate cases. Outcome definition inevitably produces borderline admissions: coded leak without clear corroboration, drainage without specific diagnostic coding, or severe sepsis after colorectal surgery without direct radiologic confirmation of dehiscence. Forcing these admissions into a hard leak or no-leak category injects label noise. A more coherent strategy assigns them high entropy phenotype vectors and propagates that uncertainty downstream. In training, such cases can contribute fractional membership to multiple states or be down-weighted by confidence. In validation, they can be examined as ambiguity strata rather than as errors in the ordinary sense.

Phenotyping should further respect anatomical and procedural context. A low pelvic colorectal anastomosis, a stapled intracorporeal ileocolic anastomosis, and an upper gastrointestinal anastomosis do not share identical baseline geometry, contamination risk, or leak presentation. A universal phenotype system that ignores operative topology may conflate distinct failure mechanisms [16]. Yet a procedure-specific phenotype taxonomy can become so narrow that it loses generalizability. The practical compromise is to define a common latent backbone, such as timing, containment, physiology, and intervention, while allowing procedure-specific modifiers to enter the mapping from raw features to phenotype coordinates.

The value of such a phenotype space is methodological rather than ornamental. Binary outcome models implicitly assume exchangeability among all positive cases. Phenotyping relaxes that assumption by acknowledging that positive cases occupy multiple local regimes in feature space. This improves not only descriptive analysis but also predictive modeling, because one can estimate phenotype-specific risks, detect differential calibration, and ask whether certain interventions alter transition probabilities between mild and severe states. In this way, phenotyping becomes an intermediate representational layer linking outcome definition to decision support. The target is no longer a single opaque event flag. It is a structured object whose components correspond to clinically intelligible modes of postoperative failure.

4. Temporal Windows, Label Assignment, and Censoring

Time is not an ancillary attribute of AL. It is constitutive of the event. The same laboratory profile may indicate baseline postoperative inflammation on day 1,

Table 9. Deployment Considerations in Clinical Settings

Aspect	Requirement	Risk	Mitigation
Interpretability	Clear outputs	Misuse	Structured summaries
Lead Time	Early detection	Late alerts	Temporal validation
Workflow Fit	Actionable triggers	Alert fatigue	Threshold tuning
Auditability	Traceable logic	Hidden bias	Version control

occult leak evolution on day 4, or recovery by day 8 [17]. Similarly, the evidentiary meaning of reoperation depends on when it occurs relative to the index procedure and to prior imaging, antimicrobial change, or physiological deterioration. Outcome definition and phenotyping therefore require explicit temporal operators. Without them, the label becomes a static overlay on a fundamentally dynamic process.

Let s_i denote the time of index anastomosis and let T_i^* denote the unobserved time at which clinically meaningful leak first becomes present. The EHR rarely contains T_i^*. Instead, it contains a set of observed times such as the first leak-relevant diagnosis code T_i^C, first leak-related intervention T_i^P, first decisive imaging order or report T_i^I, and first leak-suggestive antibiotic escalation T_i^A. These times satisfy no guaranteed ordering. A patient may undergo empiric antibiotics before imaging, or receive a diagnosis code only after reoperation. The observed event time is thus itself a constructed quantity. Common options include the earliest corroborating time, the first intervention time, or the time of code assignment. Each has consequences for causal interpretation and predictor eligibility.

If the modeling goal is postoperative surveillance, a landmarking strategy is appropriate [18]. Choose a time u after surgery and define a risk prediction problem conditional on being event-free up to u. Predict whether a leak will occur in $(u, u+h]$ for some horizon h. The label then becomes

$$Y_i(u,h) = \mathbf{1}\{\, T_i^* \in (u, u+h] \mid T_i^* > u\}. \tag{6}$$

This dynamic framing is superior to a single static 30-day label when one wants time-updated risk. It avoids leakage of future information into predictors and aligns naturally with bedside monitoring. However, it requires repeated at-risk sets and careful treatment of informative observation schedules. Patients with more frequent labs or imaging are not merely better monitored; they are often clinically sicker. Observation intensity thus becomes a time-dependent confounder.

For retrospective EHR labeling, many studies instead define a fixed horizon after surgery, say $H = 30$ days, and set

$$Y_i^{\mathrm{obs}}(H) = \mathbf{1}\{\min(T_i^C, T_i^P, T_i^I, T_i^A) \le s_i + H\}. \tag{7}$$

This is easy to implement but conflates early and late leak processes. It also obscures the fact that the opportunity to observe a late event depends on discharge status, transfer, death, and follow-up completeness. Censoring therefore matters [19]. Let C_i denote censoring time from loss of follow-up, death from unrelated causes, or end of observable record. Then the observed time is $\tilde{T}_i = \min(T_i^*, C_i)$ with event indicator $\Delta_i = \mathbf{1}\{T_i^* \le C_i\}$. If discharge truncates record access and post-discharge leakage is incompletely captured, then naive fixed-window labels can undercount events nonrandomly. Patients with shorter stays may appear event-free not because they were biologically protected, but because the observation mechanism ended.

The appropriate mathematical framework is therefore closer to multistate event history analysis than to ordinary static classification. One may define states such as stable postoperative course, occult inflammatory suspicion, image-triggered concern, confirmed contained leak, severe disseminated leak, and death or discharge. Transition intensities between these states can be written as

$$\lambda_{ab}(t \mid X_i(t)) = \lim_{\delta \downarrow 0} \frac{\Pr\left(S_i(t+\delta) = b \mid S_i(t) = a, X_i(t)\right)}{\delta}, \tag{8}$$

where $S_i(t)$ is the latent or observed state and $X_i(t)$ is the time-dependent covariate process. Even if one does not estimate the full multistate model, this representation clarifies that leak is not a single transition from zero to one. It is a path through intermediate evidentiary and management states, only some of which are recorded explicitly.

Temporal window selection also determines whether predictors are genuinely antecedent to the event. Using all data from the first 24 postoperative hours is sensible only if the goal is early surveillance and if most leak-defining evidence occurs later [20]. If a subset of cases declares itself clinically within that window, then features drawn late in the window may partly capture the event rather than its precursor. This is not simply data leakage in the colloquial sense. It is a failure to respect the distinction between predictor time and label time. A robust pipeline should therefore define a predictor window $[s_i, s_i + u]$, an exclusion buffer to avoid capturing incipient overt events, and an outcome window $(s_i + u, s_i + H]$. That separation may reduce discrimination but improves interpretability.

Competing events complicate matters further. Early postoperative death, urgent reoperation for hemorrhage,

or diversion unrelated to leak can preclude or alter the observation of leak. Standard binary labeling treats these admissions as non-events or excludes them, but either choice can distort estimates. A competing-risks formulation with cause-specific hazards or subdistribution hazards is more appropriate when the goal is population-level risk. For phenotype modeling, competing events may define separate absorbing states that interact with leak probabilities. A frail patient who dies before imaging may belong to a high-risk latent phenotype even though no leak is observed.

Temporal heterogeneity also has semantic importance [21]. A phenotype system should distinguish early catastrophic leak from delayed localized failure even if both satisfy the same 30-day rule. One may therefore define onset-sensitive phenotypes by combining event time with severity. If $T_i^{\dagger}$ is an estimated onset time and σ_i a severity score, then a phenotype map can depend on the pair $(T_i^{\dagger}, \sigma_i)$. Early severe events and late mild events will then separate naturally. This matters for both biology and intervention, since the actionable window differs substantially across temporal regimes.

From an implementation standpoint, temporal labeling should be versioned and auditable. Every admission should have a traceable record of which timestamp triggered the label, which evidence channels were active before and after that time, and what censoring assumptions were invoked. Without such provenance, disagreements between institutions cannot be decomposed into differences in biology, documentation, or observation horizon. The overarching point is that time is not merely a field in the dataset. It is part of the ontology of AL itself. Any outcome definition or phenotype system that ignores temporal structure risks converting a dynamic postoperative failure process into an unstable static artifact.

5. Latent Variable Models for AL Phenotyping

A principled phenotype model begins with the acknowledgment that the clinically meaningful state is latent. The observed record is a noisy emission from that state [22]. This immediately suggests latent variable methods. Let Z_i denote the latent phenotype for admission i, and let X_i denote the collection of observed features assembled over a defined postoperative interval. In the simplest mixture formulation,

$$p(X_i) = \sum_{k=1}^{K} \pi_k \, p(X_i \mid Z_i = k),$$
$$p(Z_i = k \mid X_i) = \frac{\pi_k p(X_i \mid Z_i = k)}{\sum_{\ell=1}^{K} \pi_\ell p(X_i \mid Z_i = \ell)}, \qquad (9)$$

where π_k is the prevalence of phenotype k. The mixture model gives a posterior phenotype vector for each admission and naturally supports soft assignment. However, the choice of likelihood $p(X_i \mid Z_i = k)$ is not straightforward because AL-relevant data are mixed-type, irregularly sampled, and temporally structured. Continuous laboratory trajectories, binary interventions, count features from notes, and time-to-event variables must coexist in the same generative framework.

One solution is a multi-view latent model. Partition the data into views, such as physiology $X_i^{(1)}$, management events $X_i^{(2)}$, coded diagnoses $X_i^{(3)}$, and text or imaging abstractions $X_i^{(4)}$. Conditional on phenotype,

$$p(X_i \mid Z_i = k) = \prod_{v=1}^{4} p\big(X_i^{(v)} \mid Z_i = k, \theta_k^{(v)}\big), \qquad (10)$$

with view-specific parameters $\theta_k^{(v)}$. This preserves interpretability because each view can be inspected separately. It also limits domination by any single high-dimensional view. For example, diagnostic codes may be sparse but highly specific, while laboratory summaries may be dense but nonspecific [23]. Multi-view factorization allows both to inform phenotype assignment without forcing them into a single homogeneous feature family.

Temporal structure can be introduced through hidden Markov or semi-Markov models. Let $Z_i(t)$ represent phenotype state over time and let observations be emitted conditionally on the current state. Then

$$\Pr\big(Z_i(t+\delta) = b \mid Z_i(t) = a\big) = q_{ab}(t)\delta + o(\delta),$$
$$p\big(X_i(t) \mid Z_i(t) = a\big) = f_a\big(X_i(t)\big), \qquad (11)$$

where $q_{ab}(t)$ are transition intensities and f_a emission laws. Such a model is attractive because it captures phenotype evolution. A patient may move from postoperative stability to occult inflammatory suspicion and then to overt contained leak or severe disseminated failure. Semi-Markov variants are particularly useful because sojourn times in each state carry clinical meaning. A brief excursion in inflammatory markers is not equivalent to sustained progressive deterioration.

Yet pure latent class discovery has limitations. The EHR contains strong sources of nuisance variation, including service practice, testing frequency, and coder behavior. Unconstrained clustering may therefore recover institutional processes rather than disease subtypes. To mitigate this, one can couple the latent phenotype model to anchor variables with high clinical credibility. Let A_i denote a set of anchors such as operative confirmation, radiographic extravasation, or drain evidence, and let the model enforce partial alignment between phenotype states and anchor prevalence [24]. A structured prior of the form

$$p(\Theta) \propto \exp\Big(-\lambda \sum_{k=1}^{K} |0\mu_k - \mu_k^{\text{anchor}}|0^2\Big) \qquad (12)$$

shrinks phenotype centroids μ_k toward clinically interpretable regions suggested by the anchor variables. This does not force the phenotypes to replicate predefined categories, but it discourages solutions that are statistically convenient and clinically meaningless.

Latent trait models provide a complementary formulation when phenotypes are better viewed as coordinates than classes. Let $h_i \in \mathbb{R}^d$ be a continuous latent representation of leak expression. Then mixed observed features can be modeled as generalized linear emissions from h_i:

$$x_{ij} \mid h_i \sim \mathcal{F}_j\big(\alpha_j + \beta_j^\top h_i\big), \tag{13}$$

where $\mathcal{F}_j$ is a feature-specific family. For a continuous laboratory summary, $\mathcal{F}_j$ may be Gaussian; for intervention occurrence, Bernoulli; for note counts, Poisson or negative binomial. Continuous latent representations are particularly useful when severity and containment vary smoothly rather than discretely. A downstream clinician-facing phenotype can still be obtained by partitioning the latent space into named regions, but the model itself is not forced into hard discreteness.

The relationship between latent phenotype and observed binary AL outcome should also be explicit. Rather than modeling phenotype only among positive cases, one may define a joint model over event and subtype. Let Y_i^* denote latent event status and Z_i the phenotype conditional on event. Then

$$p(X_i, Y_i^*, Z_i) = p(Y_i^*)\, p(Z_i \mid Y_i^*)\, p(X_i \mid Y_i^*, Z_i). \tag{14}$$

When $Y_i^* = 0$, Z_i may be set to a null state representing uncomplicated recovery or non-leak postoperative inflammation [25]. This formulation enables distinction between true leak phenotypes and near-miss or mimic states, which is important because many false positives in binary models are not random noise but clinically coherent mimics such as ileus with inflammatory elevation, postoperative abscess without anastomotic disruption, or generalized sepsis from another source.

A further extension is to treat the observed operational label Y_i^{obs} as an imperfect measurement of Y_i^*. Then

$$p(X_i, Y_i^{\text{obs}}, Y_i^*, Z_i) = p(Y_i^*)\, p(Z_i \mid Y_i^*)\, p(X_i \mid Y_i^*, Z_i)\, p(Y_i^{\text{obs}} \mid Y_i^*, Z_i). \tag{15}$$

The final factor allows phenotype-dependent label noise. Mild contained leaks may be undercoded, severe operative leaks overcaptured, and uncertain management-defined events inconsistently observed. Estimating such a model typically requires anchors, partial adjudication, or strong priors, but even an approximate version clarifies the inferential structure: the target label is a noisy sensor, not a ground truth.

Model selection for phenotyping should not rely solely on likelihood or discrimination improvement. A latent model can fit well while producing unstable or uninterpretable states. Useful criteria include phenotype stability under bootstrap resampling, alignment with chart-reviewed subsets, monotonic relation with severity proxies, and transportability across hospitals. Phenotypes that collapse or permute under slight dataset shifts are not suitable for deployment or scientific interpretation. In practice, this means latent variable modeling must be paired with semantic constraints, audit trails, and external validation. The goal is not to discover arbitrary clusters, but to estimate a hidden structure that remains clinically intelligible when confronted with the messy reality of perioperative data.

6. Representation Learning and the Geometric Structure of Leak States

Phenotyping can also be understood geometrically. The observed postoperative record defines a high-dimensional point cloud in which each admission is represented by a vector of temporal summaries, interventions, codes, and physiological measurements [26]. The challenge is that Euclidean proximity in this raw space often corresponds poorly to clinical similarity. Two admissions may be close because they share measurement frequency rather than biological state, or distant because one hospital records more granular procedure codes. Representation learning seeks a lower-dimensional space in which neighborhoods better reflect latent postoperative trajectories. In AL phenotyping, this learned geometry is valuable because it provides a substrate for both subtype discovery and uncertainty estimation.

Let $X \in \mathbb{R}^{n \times p}$ be a matrix of admission-level features, where columns may include standardized summaries of postoperative WBC, lactate, hemodynamic variability, transfusion burden, imaging triggers, drainage events, and code indicators. A classical approach is low-rank factorization

$$X \approx UV^\top, \qquad U \in \mathbb{R}^{n \times d}, \quad V \in \mathbb{R}^{p \times d}, \quad d \ll p, \tag{16}$$

where the rows of U define a latent embedding of admissions. If the reconstruction loss is chosen appropriately for mixed data types, the columns of U can be interpreted as latent factors such as inflammatory burden, intervention intensity, delayed recognition, or organ dysfunction. However, unconstrained factorization is sensitive to nuisance variance. Sparse coding, group penalties, and temporal weighting may be needed to prevent the embedding from being dominated by highly frequent but weakly specific features.

A graph-based formulation can be more robust. Construct a similarity graph on admissions with weight matrix W, where W_{ij} is large if admissions i and j have similar trajectories under a clinically informed metric. The objective is then to find embeddings U that pre-

serve local neighborhood structure. A Laplacian regularized criterion takes the form [27]

$$\mathcal{L}(U, V) = |0X - UV^\top|0_F^2 + \lambda \operatorname{tr}(U^\top L U), \quad L = D - W, \tag{17}$$

where L is the graph Laplacian. The trace term penalizes embeddings that separate nearby admissions. If the graph metric is designed to discount purely administrative similarity and emphasize physiological and intervention trajectories, the resulting manifold can reflect clinically meaningful leak structure. Admissions lying on the boundary between uncomplicated recovery and occult leak will occupy intermediate positions, which is useful for ambiguity-aware labeling.

Temporal representations can be made richer by encoding sequences rather than fixed summaries. Suppose each admission is represented by an event sequence

$$\mathcal{S}_i = \{(e_{i1}, t_{i1}), \ldots, (e_{im_i}, t_{im_i})\}, \tag{18}$$

where events include laboratory measurements crossing thresholds, medication changes, imaging, procedures, and note-derived mentions. A sequence encoder maps $\mathcal{S}_i$ into an embedding h_i. The specifics of the encoder need not be tied to any one architecture. The key design principle is that the map should be order-aware, time-gap aware, and censoring-aware. A patient with antibiotic escalation before imaging differs meaningfully from one with imaging before antibiotics, even if the unordered set of events is identical. The representation should preserve that ordering.

For phenotype discovery, the geometry should also reflect uncertainty [28]. One useful device is to represent each admission not as a single point but as a localized distribution in latent space. Let

$$h_i \sim \mathcal{N}(m_i, \Sigma_i), \tag{19}$$

where m_i is the mean embedding and Σ_i encodes uncertainty due to missingness, conflicting evidence, or sparse observation. Large covariance indicates that the admission could plausibly belong to multiple nearby phenotypic regions. This is particularly appropriate for borderline cases, such as isolated code-supported leak without intervention or postoperative inflammatory deterioration lacking direct confirmation. Decision systems can then use not only phenotype location but also phenotype confidence.

The geometry of the learned space should ideally respect known clinical monotonicities. Severity ought to align with a smooth direction in latent space, and contained-to-disseminated progression should correspond to another. These monotonicities can be encouraged by partial supervision. Let s_i be a severity proxy, perhaps derived from reoperation urgency, ICU-level support, or organ dysfunction burden. One may require that a linear functional $a^\top h_i$ correlate with s_i, adding a term

$$\mathcal{R}_{\text{ord}} = \rho \sum_{i<j} \mathbf{1}\{s_i < s_j\} \max\left(0, \delta - a^\top (h_j - h_i)\right) \tag{20}$$

to the objective. This does not reduce phenotyping to severity scoring [29]. Rather, it ensures that the learned geometry remains compatible with an important clinical axis while still allowing other axes, such as timing or evidence modality, to vary independently.

Representation learning also helps address cross-institution transportability. Raw features differ across hospitals because of coding granularity, laboratory naming conventions, and practice style. A phenotype space based directly on raw features may therefore fail to align externally. Domain-invariant learning seeks a representation h_i that preserves leak-relevant signal while suppressing site-specific artifacts. One formal objective is

$$\min_{\Phi} \; \mathcal{L}_{\text{event}}(\Phi(X), Y) + \lambda \mathcal{L}_{\text{site}}(\Phi(X), S), \tag{21}$$

where Φ is the representation map, Y the outcome or phenotype target, and $\mathcal{L}_{\text{site}}$ penalizes recoverability of site label S from the learned representation. In practice, one should not remove all site information, because some hospitals treat different populations. The objective is instead to minimize dependence on irrelevant observational idiosyncrasies while preserving clinically real heterogeneity.

From a scientific perspective, the value of geometric representation lies in its ability to reveal whether AL is best viewed as a handful of discrete phenotypes or as a continuum with dense intermediate states. If the embedding forms well-separated basins corresponding to distinct clinical regimes, then hard phenotypes may be appropriate. If the space is elongated and smoothly varying, then a continuous severity-containment manifold may be more faithful. This is not a minor technical preference [30]. It determines how labels are constructed, how decision thresholds are set, and how clinicians interpret the model output. A geometry-aware approach thus complements explicit latent class models by providing a flexible and auditable substrate on which leak states can be organized, compared, and transported.

7. Error Propagation, Bias, and Identifiability

Any AL phenotype system built from routine EHR data inherits error from multiple layers: coding inaccuracy, selective intervention, missing laboratory measurements, variable imaging thresholds, and imperfect denominator definition. These errors do not remain local to the label construction step. They propagate into feature importance, subgroup performance, calibration, and external transport. The technical difficulty is that many of these errors are differential. They

depend on the latent phenotype, the hospital, and the patient subgroup. Standard assumptions of independent and identically distributed label noise are therefore usually inappropriate.

Let Y^* denote the latent true event and Y^{obs} the operational label. Under asymmetric misclassification,

$$\begin{aligned} \Pr(Y^{\text{obs}} = 1 \mid Y^* = 1, G = g) &= \alpha_g, \\ \Pr(Y^{\text{obs}} = 1 \mid Y^* = 0, G = g) &= \beta_g, \end{aligned} \tag{22}$$

where G indexes a subgroup such as hospital, procedure type, sex, or race. If α_g and β_g vary across groups, then the observed outcome does not have uniform meaning [31]. A model may appear better calibrated in one subgroup simply because labels are more reliably captured there. Similarly, an algorithm may seem unfair when in fact the label quality differs systematically across populations. This does not eliminate fairness concerns; it complicates them by showing that fairness cannot be assessed independently of outcome ascertainment.

Phenotype-dependent misclassification is even more consequential. Suppose severe disseminated leaks almost always trigger reoperation and coding, while mild contained leaks are frequently managed conservatively without explicit diagnostic documentation. Then observed event status overrepresents one region of phenotype space. In training, the model will infer that leak is strongly associated with severe systemic disturbance, because milder phenotypes are undercounted. The resulting predictor will perform poorly on contained leaks and may encourage a clinically narrow understanding of AL. The remedy is not simply more complex algorithms. It is explicit modeling of sensitivity variation across phenotype strata or calibration against adjudicated subsets.

Measurement error in covariates also matters. Laboratory values may be absent not at random, especially preoperative nutritional markers or outpatient anthropometrics [32]. Imaging-derived features are observed only in patients who are imaged, and note-derived signals depend on clinician documentation behavior. If missingness is informative, complete-case analysis can produce biased phenotype structures by excluding precisely those admissions with unusual or emergent pathways. More generally, the observation operator is itself time-dependent. Sicker patients generate denser records, so richer feature vectors may reflect severity rather than better information. This can cause a model to exploit observation intensity as a surrogate for outcome, which is particularly dangerous in transport settings where monitoring patterns differ.

A formal way to capture informative missingness is to augment each feature x_{ij} with an observation indicator m_{ij}. Then the joint model includes

$$p(X_i, M_i \mid Z_i) = p(M_i \mid Z_i)\, p(X_i \mid M_i, Z_i), \tag{23}$$

where M_i encodes which features were observed. The term $p(M_i \mid Z_i)$ allows missingness itself to depend on latent phenotype. This is realistic. Severe suspected leaks trigger more measurement, not less. Ignoring this term can lead to optimistic internal performance, because the model effectively learns care-process intensity. A transportable phenotype model should therefore either adjust for observation process explicitly or design representations that are less sensitive to measurement density [33].

Identifiability is another central concern. If one has only routine EHR data and no chart-reviewed gold standard, the joint distribution of latent event, latent phenotype, and observation process may not be uniquely recoverable. Multiple parameter settings can generate similar observed data. This is especially true when using flexible latent variable models. To achieve partial identifiability, one must introduce anchor assumptions, bounded error rates, or small adjudicated samples. For example, operative confirmation of leak may be treated as nearly specific but not fully sensitive, whereas isolated antibiotic escalation may be treated as low-specificity evidence. Such assumptions can be encoded as priors or inequality constraints. The important point is that identification comes from assumptions, not from optimization alone.

One can express uncertainty propagation using posterior distributions over latent event and phenotype. Let θ denote model parameters and $\mathcal{D}$ the observed dataset. Then

$$p(Y_i^*, Z_i \mid \mathcal{D}) = \int p(Y_i^*, Z_i \mid \theta, \mathcal{D})\, p(\theta \mid \mathcal{D})\, d\theta. \tag{24}$$

A downstream risk model that ignores this uncertainty and treats the maximum a posteriori phenotype as fixed will understate variance and overstate confidence. This is especially problematic for subgroup analyses, where sample sizes shrink and labeling error may differ. A more coherent pipeline propagates posterior uncertainty into risk estimates, calibration intervals, and decision thresholds [34].

Bias analysis should also separate structural from allocative concerns. Structural bias arises when the data generation process itself differs across groups, such as lower imaging frequency in resource-constrained settings or differential coding practices across service lines. Allocative bias arises when the model's decisions or risk scores lead to different downstream care actions across groups. In AL decision support, the two are linked. If the label overcaptures severe leaks in one population and undercaptures contained leaks in another, then triage thresholds trained on the observed data may allocate surveillance unequally. Fairness metrics computed on noisy labels can therefore mislead unless they are contextualized by ascertainment quality.

From a methodological perspective, sensitivity analyses are indispensable. One should perturb the out-

come definition, vary the evidentiary threshold, relax or tighten the time horizon, and examine whether phenotype prevalence, severity ordering, and subgroup performance remain stable. If the phenotype map changes drastically under modest definitional shifts, then the system has poor semantic robustness. Such fragility is common when the model is implicitly learning management conventions rather than disease expressions. Stability under definitional perturbation is thus a stronger sign of validity than a marginal gain in AUC.

In sum, error in AL phenotyping is not a residual nuisance [35]. It is a primary object of analysis. The label is measured through fallible institutional processes, the features are partially observed through informative monitoring, and the phenotype structure is only partially identifiable without constraints. A serious framework must therefore include explicit error models, subgroup-aware ascertainment analysis, and uncertainty propagation. Without these components, apparently precise leak predictions may rest on unstable and unexamined observation mechanisms.

8. Validation, Transportability, and Deployment

A technically elegant phenotype system has limited value unless it survives contact with new hospitals, new care pathways, and the operational demands of perioperative decision making. Validation must therefore extend beyond internal discrimination. For AL outcome definition and phenotyping, the validation target is multidimensional. One needs to know whether the binary event label remains semantically stable across sites, whether phenotype assignments preserve meaning under altered observation policies, whether risk estimates are calibrated within phenotype strata, and whether the system supports actionable clinical workflows without collapsing into alert fatigue or semantic drift.

Internal validation is necessary but narrow. Random splits or patient-level partitions test whether the model can reproduce structure already present in the source institution. They do not test whether that structure is portable. In AL, portability is especially fragile because observation processes vary [36]. One center may image early and drain aggressively; another may reoperate sooner; another may code wound disruption inconsistently. A binary leak label that appears stable internally may therefore correspond to different latent mixtures across hospitals. Phenotype-aware external validation is preferable because it allows one to ask not merely whether overall performance declines, but which regions of phenotype space move, collapse, or disappear.

Transportability can be formalized by considering environments $e \in \mathcal{E}$, each representing a hospital or time period with its own data-generating process. Let $P_e(X, Y, Z)$ denote the joint distribution of features, observed event label, and latent phenotype in environment e. The ordinary external validation question is whether a model trained under P_{e_1} performs well under P_{e_2}. A more informative question is whether the conditional relations of interest are invariant:

$$P_{e_1}(Y^* \mid Z, X_{\text{causal}}) \approx P_{e_2}(Y^* \mid Z, X_{\text{causal}}), \qquad (25)$$

while nuisance features and observation-driven variables may shift. This suggests that transportable phenotyping should privilege features with stable mechanistic interpretation, such as physiological response patterns and operative context, over purely administrative features whose meaning is site-specific. It also suggests that site adaptation should occur at the observation layer rather than by redefining the latent phenotypes wholesale.

Validation metrics should be chosen to reflect this layered structure. Standard AUC is insufficient. Calibration is essential because perioperative decisions are thresholded on risk [37]. Yet even calibration should be evaluated conditionally. A model may be globally calibrated but systematically underpredict contained leaks and overpredict fulminant leaks. Phenotype-conditioned calibration curves, severity-ordered calibration, and cumulative incidence plots stratified by latent phenotype provide more informative diagnostics. Agreement between phenotype assignments and chart-reviewed cases, where available, should be reported not only as overall accuracy but as entropy and uncertainty concentration. A system that confidently misclassifies a subgroup of mild leaks is more problematic than one that expresses uncertainty near phenotype boundaries.

Temporal validation is likewise critical. Because hospitals change documentation systems, antibiotic protocols, and imaging thresholds, the observation operator drifts over time. A model validated only on random contemporaneous splits may therefore look more robust than it is. Evaluating on future periods tests sensitivity to documentation drift and changes in care pathways. For phenotype systems, one should examine whether centroid locations, transition probabilities, and prevalence patterns remain stable over time. Drift detection methods can monitor whether incoming admissions fall outside the support of the historical latent space or cluster in new regions suggestive of emerging practice patterns.

Clinical deployment introduces a further distinction between predictive validity and interventional validity [38]. A risk estimate is clinically useful only if it changes action at a time when action still matters. For AL, the intervention triggered by a high-risk phenotype might be intensified monitoring, earlier imaging, nutritional optimization, delayed drain removal, or heightened threshold for source-control evaluation. The value of the phenotype system therefore depends on lead time relative to overt declaration. A repre-

sentation built largely from post-declaration interventions may validate statistically yet fail operationally, because it becomes informative only after the clinician already knows the likely diagnosis. Deployment-oriented validation must quantify this lead time explicitly.

Human factors matter as much as metrics. Surgeons and perioperative teams do not act on latent vectors; they act on interpretable summaries. A deployed phenotype system should therefore render its inference in clinically intelligible terms: probability of leak, dominant phenotype regime, key evidence channels, and uncertainty level. Explanations need not expose the entire mathematics, but they should reveal enough structure that clinicians can relate model output to the patient's trajectory. For example, an output stating that the admission occupies an intermediate region between occult inflammatory leak and image-triggered contained leak, driven by rising WBC, persistent lactate elevation, and delayed bowel recovery, is far more actionable than a solitary probability number. Interpretability is not only a matter of trust. It is also a safeguard against silent transport failure, because clinicians can detect when the model is reasoning from implausible signals [39].

Evaluation should also include consequence-aware analyses. A phenotype-specific decision curve can quantify whether using the model improves net benefit at plausible surveillance thresholds. Resource-sensitive simulation can estimate how many additional imaging studies, drainage evaluations, or ICU reviews are generated per leak detected. These analyses are especially important because different leak phenotypes impose different costs of delay. Missing a contained mild leak has a different consequence profile than missing a rapidly progressive disseminated leak. A single threshold for all phenotypes may therefore be suboptimal. Deployment may require differential thresholds or two-stage systems in which a broad screen is followed by phenotype-specific refinement.

Finally, reproducibility remains foundational. Outcome definition rules, feature extraction logic, phenotype mapping, and uncertainty calculations should be version-controlled and auditable. External sites should be able to run the same pipeline and identify exactly where differences arise. Without this transparency, failure of transport cannot be decomposed. One would not know whether the disagreement originated in cohort selection, evidentiary logic, phenotype representation, or downstream calibration [40]. In surgical informatics, where data standards are uneven and local practices shape documentation strongly, this auditability is not a luxury. It is a condition for scientific accumulation.

The central implication is that validation of AL outcome definition and phenotyping must be hierarchical. First validate the event rule against adjudicated subsets and temporal logic. Then validate the phenotype structure for stability, monotonicity, and semantic coherence. Then validate risk models conditionally on phenotype and across sites and periods. Finally validate the workflow consequences in prospective settings. Only through such staged validation can one determine whether the system captures leak biology, institutional observation, or some unstable mixture of both.

9. Conclusion

Outcome definition and phenotyping for anastomotic leak should be treated as coupled inferential problems rather than as separate preprocessing and modeling steps. The binary leak label is not a direct observation of a singular clinical fact. It is an operational estimate assembled from heterogeneous postoperative evidence streams that enter the record with varying delay, precision, and dependence on local care processes. When this is ignored, prediction models absorb undocumented choices about time windows, corroborating evidence, denominator construction, and intervention thresholds [41]. Apparent improvements in discrimination may then reflect institutional observation habits as much as clinically meaningful signal.

A more coherent framework begins by separating the latent leak process from the observed event rule. The operational label should be explicit, auditable, and sensitivity-tested under plausible perturbations. Phenotyping should then represent within-event heterogeneity in terms of timing, containment, physiological burden, intervention pathway, and evidentiary modality. Depending on the data and use case, this representation may take the form of discrete latent classes, continuous manifold coordinates, or uncertainty-aware probability vectors over clinically interpretable states. In all cases, phenotype structure should be constrained enough to preserve meaning and flexible enough to accommodate borderline cases and incomplete observation. Time-aware labeling, latent variable models, geometry-preserving representations, and explicit misclassification modeling become central rather than optional. Validation must move beyond single-site AUC and examine semantic stability, subgroup-sensitive ascertainment, temporal drift, and deployment consequences. Under this view, AL phenotyping is not merely a richer way to describe positive cases. It is a mechanism for improving target validity, clarifying uncertainty, and enabling decision systems that distinguish between biologically and operationally distinct postoperative failure trajectories. The overall goal is not to multiply categories, but to ensure that what is predicted corresponds as closely as possible to the underlying clinical process of interest [42].

References

[1] A. Ulrich, C. M. Seiler, K. Z'graggen, T. Löffler, J. Weitz, and M. Büchler, "Early results from a randomized clinical trial of colon j pouch versus transverse coloplasty pouch after low anterior resection for rectal cancer.," The British journal of surgery, vol. 95, pp. 1257–1263, 8 2008.

[2] 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.

[3] K. Garude, C. Tandel, S. Rao, and N. J. Shah, "Single layered intestinal anastomosis: a safe and economic technique.," The Indian journal of surgery, vol. 75, pp. 290–293, 4 2012.

[4] C. Wong, D. Chan, I. Wong, B. Law, F. Chan, and S. Law, "479. transthoracic versus transabdominal surgical approach for siewert type ii esophagogastric junctional adenocarcinomas—a propensity score-matched analysis," Diseases of the Esophagus, vol. 36, 8 2023.

[5] K. Siaw-Acheampong, S. K. Kamarajah, R. R. Gujjuri, J. R. Bundred, P. Singh, and E. A. Griffiths, "Minimally invasive techniques for transthoracic oesophagectomy for oesophageal cancer: systematic review and network meta-analysis," BJS open, vol. 4, pp. 787–803, 9 2020.

[6] M. Hussain, "Tp7.2.10 the un-abating debate of using drains following colorectal anastomosis: a systematic review," British Journal of Surgery, vol. 108, 10 2021.

[7] A. Truong, T. Wood, P. R. Fleshner, and K. N. Zaghiyan, "A single-center experience of transanal proctectomy with ipaa for ibd.," Diseases of the colon and rectum, vol. 65, pp. 1121–1128, 6 2021.

[8] O. Yassaie, M. Thompson-Fawcett, and J. Rossaak, "Management of sigmoid volvulus: is early surgery justifiable?," ANZ journal of surgery, vol. 83, pp. 74–78, 8 2012.

[9] V. Bencurik, M. Skrovina, L. Martínek, J. Bartoš, M. Machackova, M. Dosoudil, E. Štěpánová, L. Přibylová, R. Bris, and K. Vomackova, "Intraoperative fluorescence angiography and risk factors of anastomotic leakage in mini-invasive low rectal resections.," Surgical endoscopy, vol. 35, pp. 5015–5023, 9 2020.

[10] L. A. Moukarzel, M. Byrne, S. Leiva, M. Wu, Q. C. Zhou, A. Iasonos, N. R. Abu-Rustum, Y. Sonoda, G. J. Gardner, M. M. Leitao, V. Broach, D. S., K. L. Roche, and O. Zivanovic, "The impact of near-infrared angiography and proctoscopy after rectosigmoid resection and anastomosis performed during surgeries for gynecologic malignancies," Gynecologic oncology, vol. 158, pp. 397–401, 5 2020.

[11] K. Lynes, D. A. Lawes, and C. Bailey, "Ileo-anal pouch surgery in a district general hospital results in good outcomes and high patient satisfaction.," International journal of surgery (London, England), vol. 28, pp. 34–38, 2 2016.

[12] 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.

[13] M. El-Ahmar, F. Peters, M. Green, M. Dietrich, M. Ristig, L. Moikow, and J.-P. Ritz, "Robotic colorectal resection in combination with a multimodal enhanced recovery program - results of the first 100 cases.," International journal of colorectal disease, vol. 38, pp. 95–95, 4 2023.

[14] N. R. de Barros, A. Gangrade, A. Elsebahy, R. Chen, F. Zehtabi, M. Ermis, N. Falcone, R. Haghniaz, S. Khosravi, A. Gomez, S. Huang, M. Mecwan, D. Khorsandi, J. Lee, Y. Zhu, B. Li, H. Kim, F. G. Thankam, and A. Khademhosseini, "Injectable nanoengineered adhesive hydrogel for treating enterocutaneous fistulas.," Acta biomaterialia, vol. 173, pp. 231–246, 10 2023.

[15] B. Mahendran, B. Rossi, M. Coleman, and S. Smolarek, "The use of endo-sponge® in rectal anastomotic leaks: a systematic review.," Techniques in coloproctology, vol. 24, pp. 685–694, 5 2020.

[16] M. Grieco, D. Cassini, D. Spoletini, E. Soligo, E. Grattarola, G. Baldazzi, S. Testa, and M. Carlini, "Laparoscopic resection of splenic flexure colon cancers: a retrospective multicenter study with 117 cases," Updates in surgery, vol. 71, pp. 349–357, 11 2018.

[17] N. A. Hirst, "Development of biosensors for early detection of anastomotic leak and sepsis," 8 2014.

[18] P. Gaur and S. J. Swanson, "Should we continue to drain the pylorus in patients undergoing an esophagectomy," Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus, vol. 27, pp. 568–573, 2 2013.

[19] P. H. Pucher, S. Rahman, R. C. Walker, B. Grace, A. Bateman, T. Iveson, A. Jackson, C. Rees, J. P. Byrne, J. Kelly, F. Noble, and T. J. Underwood, "Outcomes and survival following neoadjuvant chemotherapy versus neoadjuvant chemoradiotherapy for cancer of the esophagus: Inverse propensity score weighted analysis.," European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, vol. 46, pp. 2248–2256, 7 2020.

[20] A. Maher, P. D. Coppi, S. Blackburn, S. Loukogeorgakis, S. Eaton, K. Cross, S. Giuliani, J. Curry, and D. Mullassery, "Short and medium term outcomes of open and laparoscopic assisted oesophageal replacement procedures.," Journal of pediatric surgery, vol. 59, pp. 192–196, 10 2023.

[21] R. Fazylov, E. Soto, and S. Merola, "Laparoscopic roux-en-y gastric bypass in morbidly obese patients 55 years old," Obesity surgery, vol. 18, pp. 656–659, 4 2008.

[22] A. N. Townsend, A. Denton, N. Gohel, J. Lozano, P. R. de la Vega, G. Castro, and R. Seetharamaiah, "An association between comorbidities and postsurgical complications in adults who underwent esophagectomy.," Cureus, vol. 15, pp. e36395–e36395, 3 2023.

[23] S. Mills, S. Kelly, and N. McKigney, "Pth-335 permanent stoma rates following elective resection of rectal cancer," Colon and Anorectum – Cancer Including Diagnosis, Prevention and Screen, vol. 64, pp. A556.2–A557, 6 2015.

[24] P. Gioia, S. Gloor, R. Troller, and M. Adamina, "Rate of local recurrence in a cohort of 125 patients treated by transanal total mesorectal excision due to rectal cancer," British Journal of Surgery, vol. 108, 5 2021.

[25] L. Rawlins, M. P. Rawlins, C. C. Brown, and D. L. Schumacher, "Effect of helicobacter pylori on marginal ulcer and stomal stenosis after roux-en-y gastric bypass," Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery, vol. 9, pp. 760–764, 7 2012.

[26] 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.

[27] K. A. Garrett, B. J. Champagne, B. T. Valerian, D. Peterson, and E. C. Lee, "A single training center's experience with 200 consecutive cases of diverticulitis: Can all patients be approached laparoscopically?," Surgical endoscopy, vol. 22, pp. 2503–2508, 3 2008.

[28] A. Ammar, M. Paul, A. El-Sharkawy, P. Leeder, A. Bohra, S. Awad, and S. Iftikhar, "O090 the sweet approach to oesophagectomy is still a valid surgical approach seldom used in the uk," British Journal of Surgery, vol. 109, 7 2022.

[29] J. K. Pye, M. K. H. Crumplin, M. E. Foster, A. Biffin, and J. Charles, "Carcinoma of the oesophagus and stomach in wales: a 1-year survey," Journal of British Surgery, vol. 86, pp. 422–422, 3 1999.

[30] D. A. Nguyen, T.-A. Mai-Phan, T. T. Thai, and H. V. Nguyen, "Laparoscopic hartmann reversal: Experiences from a developing country.," Annals of coloproctology, vol. 38, pp. 297–300, 6 2021.

[31] I. Kyrochristou, E. Spartalis, G. Anagnostopoulos, G. Tsourouflis, D. Dimitroulis, and N. I. Nikiteas, "Crp in drain fluid as a predictive marker of anastomotic leak in colorectal surgery: A systematic review of the literature.," In vivo (Athens, Greece), vol. 37, pp. 1450–1454, 6 2023.

[32] M. M. B. Olyaee, "Anastomotic leakage after thoracic esophagectomy for esophageal cancer: Systematic review and meta-analysis," Zenodo (CERN European Organization for Nuclear Research), 9 2023.

[33] V. Jayaraman, C. Hammerle, S. K. Lo, L. H. Jamil, and K. Gupta, "Clinical application and outcomes of over the scope clip device: Initial us experience in humans," Diagnostic and therapeutic endoscopy, vol. 2013, pp. 381873–381873, 7 2013.

[34] J. Orford, D. T. Cass, and M. J. Glasson, "Advances in the treatment of oesophageal atresia over three decades: the 1970s and the 1990s," Pediatric surgery international, vol. 20, pp. 402–407, 5 2004.

[35] S. R. Markar, C. Gronnier, A. Duhamel, J.-M. Bigourdan, B. Badic, M. C. du Rieu, J. H. Lefevre, K. Turner, G. Luc, and C. Mariette, "Pattern of postoperative mortality after esophageal cancer resection according to center volume: Results from a large european multicenter study.," Annals of surgical oncology, vol. 22, pp. 2615–2623, 1 2015.

[36] N. Sato, K. Yamaguchi, S. Shimizu, T. Morisaki, K. Yokohata, K. Chijiiwa, and M. Tanaka, "Coil embolization of bleeding visceral pseudoaneurysms following pancreatectomy: the importance of early angiography.," Archives of surgery (Chicago, Ill. : 1960), vol. 133, pp. 1099–1102, 10 1998.

[37] S. N. Lukmonov, Y. V. Belenkaya, M. S. Lebedko, S. S. Gordeev, and Z. Z. Mammadli, "The effect of neoadjuvant treatment on postoperative morbidity in upper rectal cancer," Surgery and Oncology, vol. 13, pp. 46–53, 6 2023.

[38] H. Zhao, A. M. Misariu, J. L. Ramirez-GarciaLuna, T. B. Nobel, C. L. Mueller, J. Cools-Lartigue, J. Spicer, D. Molena, M. S. Bains, S. G. Swisher, W. L. Hofstetter, and L. E. Ferri, "Synchronous esophageal and lung cancers - is combined anatomic resection appropriate?," The Annals of thoracic surgery, vol. 113, pp. 1354–1360, 4 2021.

[39] J. V. Reynolds, J. A. Elliott, N. E. Donlon, C. L. Donohoe, N. Ravi, M. K. Kuppusamy, and D. E. Low, "548 comparison of esophagectomy outcomes between a national center, a national audit collaborative, and an international database using eccg standardized definitions," Diseases of the Esophagus, vol. 34, 9 2021.

[40] B. Vidarsson, M. Sundbom, and D. Edholm, "Incidence and treatment of small bowel leak after roux-en-y gastric bypass: a cohort study from the scandinavian obesity surgery registry.," Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery, vol. 16, pp. 1005–1010, 4 2020.

[41] C. Papadimitriou, C. Deliveliotis, A. Dellis, W. Martin, and I. Mitsogiannis, "Uretero-ureterostomy combined with unilateral nephrostomy as a method of urinary diversion following radical cystectomy.," Cureus, vol. 14, pp. e27501–e27501, 7 2022.

[42] I. Bartolini, M. Risaliti, M. N. Ringressi, F. Melli, G. Nannini, A. Amedei, P. Muiesan, and A. Taddei, "Role of gut microbiota-immunity axis in patients undergoing surgery for colorectal cancer: Focus on short and long-term outcomes.," World journal of gastroenterology, vol. 26, pp. 2498–2513, 5 2020.