

Original Research

Corpus Design, Record Ecology, and Borderline Phenotype Construction for Anastomotic Leak in Longitudinal Surgical Documentation Systems

Ramesh Adhikari¹ and Bikash Koirala²

¹Faculty of Information Technology, Mid-West University, Surkhet–Jumla Road, Birendranagar 21700, Surkhet, Nepal.

²School of Engineering, Agriculture and Forestry University, Rampur Campus, Bharatpur 44200, Chitwan, Nepal.

Abstract

Background conditions in surgical informatics have changed rapidly as electronic health record repositories have become larger, denser, and more queryable, but this expansion has not resolved the core methodological difficulty of defining postoperative complications in a clinically stable way. Anastomotic leak remains one of the clearest examples of this unresolved problem. The complication is not recorded as a single transparent datum. It emerges through combinations of bedside concern, physiologic deterioration, diagnostic pursuit, procedural response, antimicrobial escalation, and retrospective documentation, all of which appear unevenly across time and across institutions. This paper develops a highly technical framework for outcome definition and phenotyping that approaches leak not as a simple binary event but as a corpus-construction problem inside a complex documentary environment. The central argument is that the quality of an AL label depends first on how the postoperative record is assembled into a case corpus, second on how evidentiary chains are distinguished from one another, and third on how borderline and adjacent states are represented instead of being silently absorbed into positive or negative classes. The paper examines record ecology, denominator discipline, sentinel documentation, trigger chains, ambiguous and mimic trajectories, annotation grammar, phenotype stabilization, and cross-institution translation. It proposes that a useful leak phenotype system must be able to describe not only whether a case satisfies an operational endpoint, but also how that case became legible inside the record and which documentary pathway produced its classification. Under this view, reliable AL phenotyping is less a matter of discovering one perfect rule than of building a reproducible corpus architecture for difficult postoperative states.

1. Introduction

Outcome definition for anastomotic leak often looks deceptively straightforward when viewed from the perspective of a finished dataset [1]. A cohort is assembled, a field is designated as the target, and model development proceeds as though the event had been directly observed. Yet that appearance of simplicity is produced only after a long sequence of clinical, documentary, and analytic transformations has already taken place. Before a leak becomes a label, it exists as a possibility in the minds of clinicians, a set of bedside concerns, a reason for ordering tests, a motive for escalation, a series of note fragments, and eventually, in some but not all cases, a coded or otherwise structured complication. The computational endpoint is therefore not merely extracted from the record. It is synthesized from heterogeneous traces whose timing, abundance, and meaning depend on how postoperative care was delivered and documented. For this reason, the most important methodological question is often not how to fit a classifier, but how to construct the corpus of cases from which a classifier would learn anything clinically coherent at all.

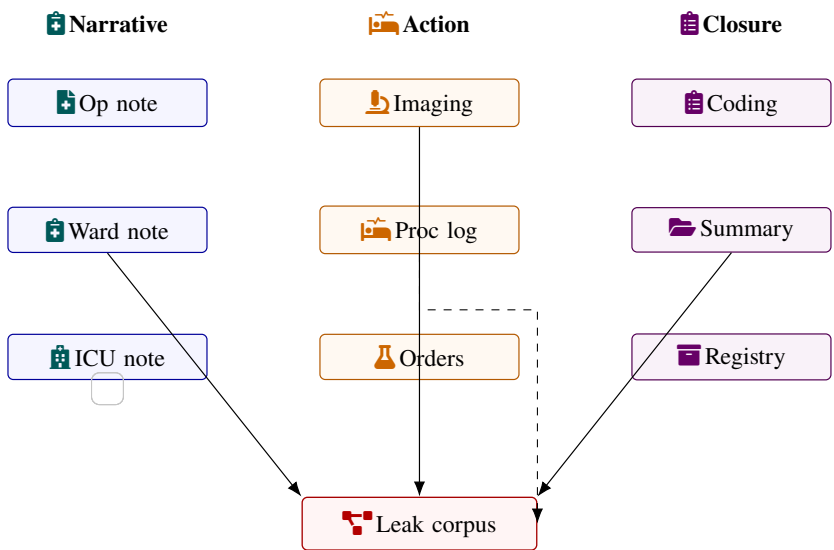


Figure 1: Record ecology for anastomotic leak corpus construction. Documentary evidence is partitioned into narrative, action, and closure lanes rather than being collapsed immediately into one flat evidence pool. The figure emphasizes that leak status is shaped not only by what is documented, but by which documentary stream first renders the postoperative course legible and which stream later consolidates the case.

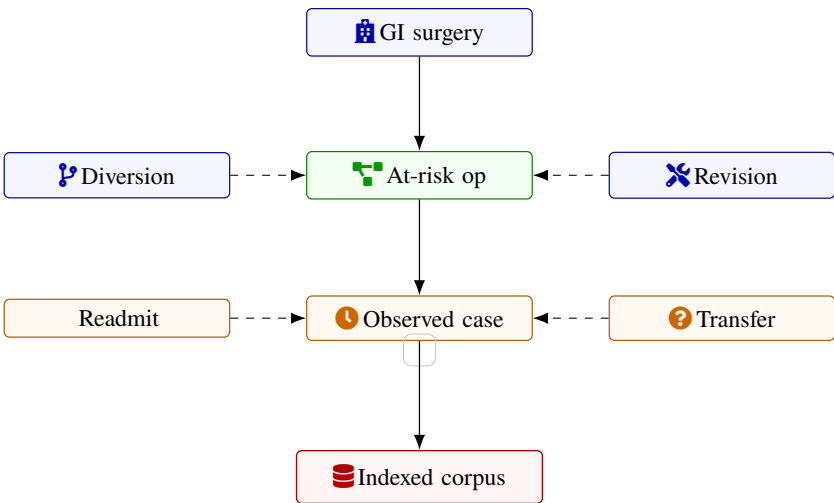


Figure 2: Layered denominator construction. Operative eligibility, anastomotic exposure, and observation adequacy are separated before leak logic is applied. Side modifiers indicate that diversion, revision context, readmission linkage, and fragmented follow-up alter the meaning of both the denominator and the negative class.

Anastomotic leak is especially sensitive to this problem because it sits at the junction of structural failure, physiologic disturbance, institutional workflow, and retrospective abstraction. The event is structural in the sense that it concerns failure of an anastomotic site and the consequences of contamination or loss of integrity. It is physiologic in the sense that the patient’s course may express that failure through inflammatory markers, hemodynamic change, organ dysfunction, and delayed recovery. It is workflow-dependent because the degree to which the event becomes visible in the record depends on who suspects it, when they suspect it, whether imaging is pursued, whether drainage or reoperation occurs, and how

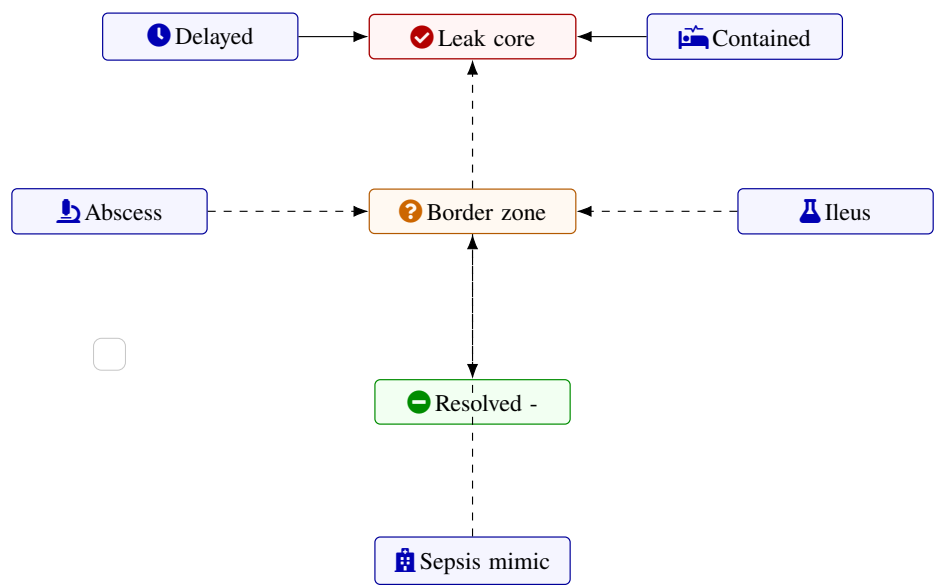


Figure 3: Borderline and leak-adjacent territory. The diagram separates a small resolved leak core from a broader uncertainty band that is fed by delayed or weakly corroborated leak-like states and by adjacent mimics such as non-anastomotic abscess, ileus-like inflammatory trajectories, and systemic sepsis pathways. This preserves the structure of disagreement instead of forcing all edge cases directly into positive or negative bins.

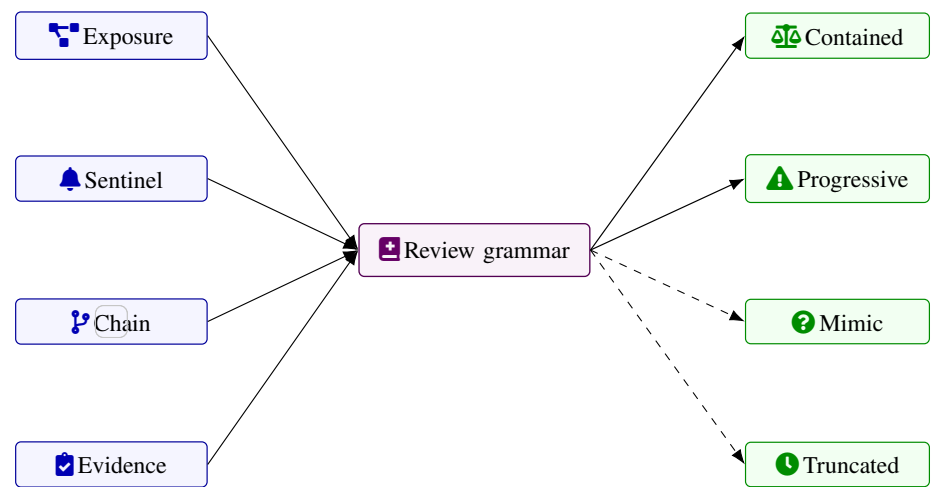


Figure 4: Annotation grammar for phenotype stabilization. Review is structured around four compact fields—operative exposure, sentinel appearance, documentary chain, and evidentiary closure—before assignment into stabilized phenotype classes. This representation supports reproducible adjudication while keeping unresolved and mimic trajectories visible as first-class outputs instead of treating them as residual error.

the subsequent clinical course is documented. It is retrospective because discharge coding or summary documentation may consolidate the event after the decisive clinical moments have already passed. These layers overlap unevenly. Some patients generate rich and early documentation of a leak-like process, while others generate sparse or delayed traces. Some non-leak complications generate remarkably similar traces. Thus, the label is not only a clinical conclusion. It is also the output of a documentary ecology.

Table 1: Documentary Sources and Their Roles in Leak Visibility

Source Type	Strength	Limitation
Progress Notes	Early clinical concern	variability
Radiology Reports	Structural detail	Interpretive uncertainty
Procedure Records	Action evidence	indication
Diagnosis Codes	Standardized labels	Retrospective delay
Medication Orders	Treatment trajectory	Non-specific

Table 2: Forms of Documentary Asymmetry

Dimension	High End	Low End
Narrative Density	Notes	Codes
Standardization	Codes	Free text
Timing	Real-time notes	Discharge summaries
Specificity	Imaging findings	Antibiotic use
Interpretability	Structured data	Context-heavy prose

Table 3: Denominator Construction Layers

Layer	Purpose	Risk if Ignored
Operative Eligibility	Identify at-risk anatomy	inclusion
Exposure Confidence	Validate anastomosis	Misclassification
Observation Adequacy	follow-up	False negatives
Context Modifiers	Capture surgical setting	Loss of nuance
Encounter Linkage	Track continuity	Fragmented cases

Table 4: Observation Adequacy Categories

Category	Description	Implication
Full Observation	Complete inpatient + follow-up	Reliable negative
Partial Observation	Limited follow-up	Uncertain status
Truncated	Early discharge/transfer	Missing outcomes
Fragmented	Multi-encounter gaps	Disjoint evidence
Delayed Capture	Late readmission	Shifted timing

A large proportion of machine learning work on postoperative complications has been designed around the assumption that the target is mostly settled and that the remaining challenge lies in prediction. That assumption is easier to sustain for outcomes that have singular operational anchors, such as mortality within a defined interval or a tightly specified reoperation. Anastomotic leak does not enjoy that convenience. Even when one adopts a widely used composite endpoint, the endpoint still reflects decisions about which evidentiary channels matter, how far into the postoperative period they remain admissible, whether coding can serve as primary evidence or only as confirmation, how to handle conservatively managed cases, and what should happen to courses that are suspicious but unresolved. None of these decisions is neutral. Each choice changes the shape of the positive class, the composition of the negative class, and the set of cases that remain in an ambiguous middle.

The usual way of coping with this heterogeneity is to sharpen the positive rule. Analysts combine diagnosis codes with procedure records, antimicrobial escalation, or other signals in order to improve specificity [2]. This is often sensible and sometimes necessary. However, such composite rules rarely articulate the broader representational problem they are solving. They tell us which cases close as positive, but they say much less about the documentary states that sit near the boundary, the kinds of negative

Table 5: Sentinel Document Types

Type	Example	Role
Clinical Note	Suspicion statement	Initiates concern
Radiology Report	Suspicious collection	Structural trigger
Order Event	Imaging request	Tacit concern
Care Escalation	ICU transfer	Severity signal
Procedure Note	Drain placement	Action threshold

Table 6: Trigger Chain Patterns

Pattern	Sequence	Interpretation
Rapid Escalation	Note → OR	Severe leak
Imaging-led	Scan → Drain	Contained leak
Gradual Concern	Notes → Antibiotics	Indeterminate
Aborted Chain	Suspicion → Recovery	Mimic/negative
Fragmented Chain	Multi-encounter	Delayed recognition

Table 7: Leak-Adjacent Mimic States

Condition	Shared Features	Distinguishing Aspect
Non-anastomotic Abscess	Imaging + drainage	Location mismatch
Severe Ileus	Distension + delay	No defect
Ischemia	Systemic decline	Different pathology
Pneumonia/Sepsis	ICU + antibiotics	Extra-abdominal source
Fluid Collection	Imaging findings	Benign evolution

Table 8: Borderline Case Characteristics

Feature	Example	Challenge
Incomplete Evidence	Suspicious scan only	Low certainty
Conflicting Signals	Notes vs imaging	Interpretation conflict
Delayed Confirmation	Late coding	Timing ambiguity
Conservative Management	Antibiotics only	Missing proof
Resolution Without Proof	Recovery	Classification difficulty

cases that are clinically adjacent, or the routes through which positive status became visible. The result is that the most interesting and difficult cases become harder to see after labeling than they were before. They either disappear into a negative residual class, are forced into a positive class on narrow grounds, or are excluded from the cohort as nuisances. A sophisticated predictive model trained on such a target may still rank patients usefully, but the target itself remains semantically underdescribed.

This paper proposes a different strategy. It treats AL outcome definition first as a corpus-design problem. A corpus, in this setting, is not simply the set of admissions in a database. It is the set of postoperative documentary objects constructed by selecting an operative denominator, organizing evidentiary streams, preserving temporal order, and defining which kinds of narratives or trigger chains count as instances of the clinical phenomenon under study. Under this view, the decisive act is not the final binary classification but the earlier architectural work of determining what a case consists of. That

Table 9: Phenotype-Relevant Trajectory Types

Trajectory	Features	Phenotype
Contained Leak	Local + drainage	Stable course
Progressive Leak	Escalation + OR	Severe course
Mimic Pathway	Partial overlap	diagnosis
Aborted Suspicion	Early concern only	Negative neighbor
Unresolved	Persistent ambiguity	Borderline class

architectural work includes deciding what belongs in the denominator, what counts as sentinel documentation, how unresolved trajectories are represented, and which kinds of negative cases must be modeled explicitly because they form the near-neighborhood of leak. Once the corpus is built in this way, binary outcomes and phenotypes can be derived more transparently.

The paper therefore does not begin from a latent variable framework, a decision-theoretic threshold, or a governance-centered lifecycle, although each of those themes is relevant. Instead, it begins from the material fact that leak becomes visible in records through a changing ecology of documents and actions. Clinical notes, operative notes, radiology reports, medication orders, ICU transitions, laboratory trajectories, diagnosis codes, and discharge summaries all participate in that ecology, but not symmetrically. Some function as early warnings, some as turning points, some as retrospective consolidations, and some as misleading echoes of other postoperative disturbances. A record-ecology perspective is especially valuable for AL because it can distinguish state from documentation, event from route of recognition, and strong evidence from abundant evidence. These distinctions are often blurred in ordinary tabular cohorts.

From this perspective, corpus design and phenotyping become intimately connected. A phenotype is not simply a cluster inside the positive class. It is a stabilized recurring pattern of documentary and clinical evolution within the corpus. To identify such patterns, one must represent not only the obvious leak cases but also the border cases, the mimic cases, and the truncated or underobserved cases. In other words, the corpus has to be built with enough negative structure and enough ambiguity retention that the phenotype system can distinguish between clinical heterogeneity and simple label noise. If everything near the boundary is either discarded or coerced into binary certainty, the phenotypes that remain may be crisp but misleading [3].

The sections that follow develop this argument in a way that is intentionally technical while avoiding formal mathematics. The discussion first examines record ecology and the asymmetry of documentary evidence. It then turns to denominator construction and corpus assembly, arguing that the meaning of the event depends critically on how the at-risk set is defined. After that, it considers sentinel documents and trigger chains, showing how leak becomes narratively legible through specific sequences of documentation and action. The paper then addresses borderline cases, mimic syndromes, and event adjacency, because the edge of the phenomenon is where most definitional instability originates. The later sections develop an annotation grammar for disciplined review, describe how phenotypes can be stabilized within a heterogeneous corpus, and examine cross-institution translation under documentary drift. The overall aim is not to provide one final leak rule. It is to describe how a research program on AL outcome definition should build its objects before modeling begins.

2. Record Ecology and Documentary Asymmetry

The postoperative record should not be imagined as a neutral container of facts. It is better understood as an ecology of documents generated by distinct actors under distinct pressures. Surgeons write operative notes and progress notes with one set of aims. Intensivists and ward teams chart physiologic response with another. Radiologists narrate imaging in a specialized descriptive register. Pharmacologic actions leave behind orders, administrations, and sometimes consult notes. Coders transform parts of

this documentary field into structured abstractions after the fact. Each documentary form carries its own timing, granularity, and relation to uncertainty. Anastomotic leak moves through this ecology unevenly. In some cases it appears first as bedside suspicion in progress notes. In others it becomes visible through radiology language before clinicians commit to a diagnosis. In still others it is fully legible only retrospectively when multiple strands have already converged. Recognizing this ecology is essential because the computational label is only ever a downstream reduction of these interacting documentary streams.

The notion of documentary asymmetry is central here. Some documents have high narrative density and low standardization. Progress notes may include tentative impressions, differential reasoning, and procedural context, but they are highly variable in style and completeness. Other sources have high standardization and low narrative nuance [4]. Diagnosis codes are compact and readily queryable, yet they often omit the path through which the diagnosis emerged and may fail to preserve uncertainty or competing explanations. Radiology reports occupy a middle position, often offering fairly detailed descriptions of collections, extravasation, free air, or inflammatory change, but phrased within the rhetorical conventions of radiologic interpretation rather than surgical decision making. Procedure records indicate that something was done, but not always why. A drain placement can reflect source control for a contained leak, management of a non-anastomotic collection, or a broader abdominal concern. The same documentary trace can therefore carry very different clinical implications depending on context.

This asymmetry matters because analytic pipelines tend to privilege what is easiest to query. Structured fields and billing abstractions are often placed at the center of the outcome definition because they are easier to scale. Yet the most conceptually decisive information may sit in less standardized documentation. A surgeon's note may say that a patient is concerning for leak but not definitively proven. A radiology report may describe a pelvic fluid collection whose interpretation depends on operative anatomy. An ICU note may link a pattern of septic physiology to the recent anastomosis without any corresponding code appearing until days later. If a corpus is built only from structured signals, these documentary moments are either flattened or lost. The resulting label may still be useful, but its relationship to the clinical process becomes thinner than the record could have supported.

Documentary asymmetry also creates different forms of delay. A process note may register concern at the moment it arises. A procedure record may appear when action is taken. A discharge code may appear only after the hospitalization has already ended. Each of these delays has a different meaning. Early concern is a clue about prospective recognizability. A later procedure may indicate that the case crossed a threshold of actionable certainty. A discharge code may be valuable for retrospective ascertainment but almost useless for early surveillance. If the corpus ignores these distinctions and simply records that all of these items happened within thirty days, it destroys one of the most informative aspects of the evidence: namely, when the system itself became aware that something leak-like was occurring.

There is a second, subtler form of asymmetry arising from institutional habit. Some services document suspicion aggressively [5]. Others prefer sparse note-writing until intervention is certain. Some institutions have highly structured radiology reports. Others rely more on progress note synthesis. Some centers code postoperative complications in great detail, while others do so minimally unless quality reporting demands it. These patterns do not simply add noise. They alter the visibility of specific classes of cases. A hospital that documents uncertainty richly may produce a large documentary border zone around leak. A hospital that documents only confident diagnoses may make the same border zone almost invisible. If a phenotype system is built at only one site and then transported without recognizing this, apparent differences in prevalence or model error may actually be differences in documentary ecology.

A record-ecology view therefore suggests that outcome construction should preserve document type as part of the case representation. Instead of immediately converting all relevant traces into one undifferentiated evidence score, the corpus should retain whether a signal arose from operative narration, bedside concern, imaging interpretation, procedure records, medication trajectories, or retrospective coding. This is not because document type is intrinsically more important than the content of the document. It is because document type helps distinguish state from route of visibility. A case recognized

primarily through radiology and drainage is not necessarily a different biological state from a case recognized primarily through narrative concern and reoperation, but it is a different documentary route, and that difference matters both for phenotyping and for transport.

One practical implication is that rare or difficult cases may be best captured by their documentary pattern rather than by any single clinical marker. A conservatively managed contained leak may not receive a definitive operative finding or a prominent diagnosis code, but it may show a recurring documentary constellation: repeated note-level concern, one or more imaging descriptions of a peri-anastomotic collection, prolonged antibiotics, and a lengthened postoperative stay. A non-leak mimic may show a similar inflammatory and management profile but differ in where the notes locate the concern or in how the radiology is framed. These are distinctions that become visible only when the corpus is allowed to preserve documentary heterogeneity rather than reducing it prematurely.

Another implication concerns the meaning of missingness. In a standard tabular dataset, absence of a variable value is a technical nuisance. In a record ecology, absence can have documentary meaning. No imaging may indicate low suspicion, rapid recovery, resource limitation, clinical confidence in another diagnosis, or truncation of observation. No explicit diagnosis code may indicate the complication was never resolved, was managed under a different label, or simply escaped abstraction. No narrative mention of leak may reflect the culture of note-writing more than the absence of concern. Thus, missingness in documentary space is not homogenous [6]. It can be informative or uninterpretable depending on context. A corpus that records only whether an item is present or absent, without regard to the ecology in which absence occurred, cannot distinguish these cases.

The strongest argument for preserving record ecology, however, is interpretive. Phenotype systems that arise from ordinary tabular data often end up describing observational styles rather than clinically intelligible postoperative states. A cluster may be distinguished not because it represents a distinct form of leak, but because those patients were worked up with more imaging or were treated on services that document more richly. When the ecology is represented explicitly, these routes of visibility can be identified and either treated as phenotype descriptors or set aside when the goal is to stabilize the underlying clinical states. Without that step, the phenotype map may look statistically coherent while remaining clinically opaque.

For AL research, then, record ecology is not an external complication. It is the environment in which the phenomenon becomes visible. A computational endpoint that ignores this ecology risks becoming a summary of documentation behavior rather than a representation of postoperative failure. A corpus built with documentary asymmetry in mind is harder to construct, but it yields objects that are more durable. It preserves the distinction between early narrative concern and late administrative closure, between strong evidence and abundant evidence, and between clinical states and the routes by which those states become visible in different care systems. Those distinctions are foundational for any subsequent work on phenotype stabilization.

3. Denominator Construction and Corpus Assembly

Outcome definition begins not with the positive rule but with the denominator. This point is so basic that it is often overlooked. Yet for anastomotic leak, the at-risk set is not merely all postoperative surgical admissions, nor even all gastrointestinal resections. It is a subset defined by the presence of an anastomosis capable of leaking, by the documentary visibility of that operative fact, and by the existence of enough postoperative observation for the event to become visible if it occurs. If the denominator is incoherent, the positive class can never be interpreted cleanly because leak prevalence, predictor distributions, and negative-class composition all become distorted. Corpus assembly therefore requires denominator discipline before any event logic is applied.

The first denominator problem is anatomical exposure. Not every relevant operation leaves behind the same risk object. A resection without an anastomosis should not inhabit the same at-risk set as a low pelvic colorectal anastomosis, even if both belong broadly to colorectal surgery. Temporary diversion

complicates the picture further. A diverted anastomosis remains an anastomosis, but its leak manifestations and documentary path may differ from an undiverted construct. Revision cases, staged procedures, and salvage operations may also alter what it means for a postoperative course to be at risk for the specific leak phenotype under study [7]. If the denominator is built through generic procedure categories rather than through explicit recognition of operative constructs, the corpus may include cases that are not truly at risk for the target event or exclude cases whose leak risk is clinically central.

The second denominator problem is documentary eligibility. In many large datasets, the presence of an anastomosis is not represented by a single clean field. It must be inferred from procedure descriptions, operative notes, codes, or combinations of these. This means that denominator assembly already involves a documentary interpretation problem. A center whose operative notes are richly structured may support fine-grained denominator construction, while another center may support only rough procedure categories. Thus, even before the positive class is defined, corpus assembly is shaped by what the record can say about operative reality. A disciplined framework should acknowledge this and preserve the confidence with which denominator membership itself was assigned. Otherwise, admissions with weakly inferred exposure will be mixed indiscriminately with admissions whose operative status is known clearly.

A third issue concerns observation adequacy. A case can belong anatomically to the at-risk set and still be poorly observed. Very early discharge, inter-institution transfer, fragmented follow-up, or incomplete capture of post-discharge care can all make an event invisible. This is not an edge case for AL. Some leaks declare after the immediate inpatient window, and some are recognized only when the postoperative course deviates enough to trigger later imaging or readmission. If the corpus treats every underobserved admission as a straightforward negative whenever no positive evidence appears, it quietly converts lack of observation into evidence of absence. A more disciplined corpus architecture distinguishes between fully observed negatives and negatives whose status rests partly on truncated follow-up. This distinction is especially important in comparative work, because institutions with different lengths of stay and follow-up capture will otherwise appear to differ in leak rates for partly documentary reasons.

Denominator discipline also affects negative-space structure. The composition of the non-event class depends on what kinds of cases were admitted into the corpus in the first place. If the denominator is too broad, the negative class will be dominated by easy, far-from-boundary admissions, making model discrimination appear stronger than it is in clinically relevant settings. If the denominator is too narrow or restricted to high-acuity populations, the negative class may become enriched for complex postoperative syndromes that closely border leak. Neither option is wrong in itself, but each supports a different inferential purpose. A surveillance-oriented corpus may deliberately emphasize difficult postoperative trajectories because those are the cases in which discrimination matters. A general benchmarking corpus may retain a broader denominator to represent overall operative risk. The key is that the choice be explicit and aligned with the intended use of the outcome [8].

Corpus assembly should therefore proceed in layers. The first layer identifies operative eligibility. The second documents exposure confidence and contextual modifiers such as diversion, urgency, and operative setting. The third establishes observation adequacy, distinguishing full, partial, and truncated postoperative visibility. Only after these layers are in place should event-related evidentiary streams be assembled. This layered approach is more cumbersome than simple cohort extraction, but it preserves the semantic foundation on which later phenotype work depends. It also makes sensitivity analysis more informative. One can see how event prevalence and phenotype distribution change when only fully observed admissions are retained, or when uncertain operative exposures are excluded, instead of confounding those issues with event-rule changes.

Corpus assembly also has a narrative dimension. The denominator should not be imagined as a static table of operations. It is a set of operative situations from which certain postoperative stories can plausibly emerge. A low pelvic anastomosis creates one narrative risk environment, an emergent right hemicolectomy another, and a revision anastomosis yet another. Preserving these contextual distinctions in the corpus helps later phenotype interpretation because the same documentary pattern may have different significance in different operative neighborhoods. For example, a small pelvic collection has one

meaning after a low anterior resection and another after a more proximal anastomosis. A corpus stripped of operative context will treat these as more similar than they should be.

One practical lesson from existing EHR work is that even relatively simple proof-of-concept studies already embody denominator decisions with substantial downstream consequences. In the MIMIC-based proof-of-concept by Sadanandan (2024) [9], the cohort was restricted to colorectal resection admissions and the target was defined within a thirty-day postoperative horizon using diagnostic and corroborating clinical evidence, but the study also acknowledged that chart validation and broader subgroup analyses remained incomplete. That acknowledgement is instructive because it points to the hidden work that denominator discipline still requires even when the predictive pipeline is otherwise transparent. The cohort can be reproducible and still rest on unresolved questions about exactly which operative exposures and postoperative visibility patterns populate it.

Denominator construction also conditions how phenotype borders behave. If underobserved admissions are allowed to accumulate in the negative class, the border between leak and unresolved postoperative concern becomes artificially sharp. If operations with unclear or irrelevant anastomotic exposure enter the corpus, the leak neighborhood becomes diluted by cases whose negative status has a different meaning. If admissions from highly selected intensive care environments dominate the denominator, mimic states may appear more central than they would in a general postoperative corpus. These effects are often attributed later to model instability, but they are already encoded in corpus assembly [10].

Another underappreciated aspect of corpus assembly is the status of readmissions and post-discharge episodes. Some leaks surface after discharge, while others generate fragmented evidence across multiple encounters. A corpus that treats only the index admission as the unit of analysis may fail to capture the true narrative arc of such cases. On the other hand, combining all postoperative encounters indiscriminately can blur the distinction between early postoperative evolution and later complications less clearly tied to the initial anastomosis. The appropriate choice depends on the target, but whatever choice is made should be reflected in how the corpus represents linkage across encounters. Otherwise, late leak recognition may appear as a deficiency of the event rule when it is actually a property of the unit of observation.

Corpus assembly, then, is not merely about selecting rows from a database. It is about constructing a meaningful field of postoperative cases in which leak, mimic, and unresolved states can be interpreted against a coherent background. The denominator defines the terrain on which the later event border will be drawn. If that terrain is poorly assembled, the border will be unstable no matter how cleverly the positive rule is designed. A research program on AL that takes corpus design seriously will therefore treat operative eligibility, observational adequacy, contextual modifiers, and encounter linkage as part of outcome definition itself. Only then can later phenotype stabilization rest on something firmer than a convenient but semantically thin cohort.

4. Sentinel Documents and Trigger Chains

Once the corpus has been assembled, the next methodological problem is to identify how leak-like states first become legible in the postoperative record. This is not the same as asking when the leak truly began. It is asking when the documentary system started to register a chain of observations and actions that pointed toward leak territory. That distinction is important because many uses of AL labeling, particularly surveillance and early recognition, depend more on documentary legibility than on unknowable biological onset. Sentinel documents and trigger chains offer a way to capture this process. A sentinel document is a piece of documentation or a structured event that marks the beginning of a leak-relevant interpretive sequence. A trigger chain is the set of subsequent documents, actions, and findings through which that sequence develops into a positive, unresolved, or mimic classification.

The idea of a sentinel document is useful because leak rarely arrives as a single definitive item in routine records. More often, it enters the chart through a first articulation of concern. That articulation may appear in a progress note stating concern for an anastomotic complication, in a radiology report describing a suspicious collection, in an unplanned return to an intensive monitoring environment, or in

a sudden management change that only makes sense if something new and serious is suspected. Not all sentinel documents are equal. Some carry explicit leak-oriented language, while others merely alter the trajectory of the record in a way that signals mounting concern. Yet each has the same structural role: it begins a chain that can later be traced and interpreted [11].

Trigger chains matter because the meaning of a sentinel event depends on what follows. A radiology report mentioning a possible peri-anastomotic collection means one thing if it is followed by drainage, persistent inflammatory concern, and prolonged antibiotics, and another if it is followed by clinical reassurance and uneventful recovery. A note-level suspicion means one thing if later confirmed by operative findings and another if displaced by a better alternative explanation. Thus, event definition should not be based solely on isolated sentinel items. It should pay attention to chains of confirmation, divergence, and abandonment. Leak becomes visible not only through positive findings but through the way the record continues to organize itself around those findings.

One advantage of trigger-chain thinking is that it restores process to what otherwise appears as a set of disconnected fields. Instead of asking whether a diagnosis code, a procedure, and a laboratory abnormality all occurred within a window, one asks whether they were parts of a coherent postoperative storyline. This is closer to clinical reasoning. Surgeons do not infer leak by noticing that several relevant items happened at some point in the same month. They infer it from how those items accumulate, reinforce, or redirect concern over time. A trigger chain captures that accumulation. It allows the corpus to distinguish a case with a short, forceful chain leading rapidly to intervention from one with a long, equivocal chain that never fully resolves.

Trigger chains are also valuable because they can differ by phenotype. Contained leaks may often begin with localized symptoms or slower inflammatory divergence, lead to imaging, and culminate in drainage or prolonged medical management. Catastrophic leaks may have a shorter chain with more abrupt escalation and less diagnostic ambiguity. Mimic states may follow chains that resemble contained or progressive leak early on but then branch away when the record supports another explanation. Unresolved cases may show repeated beginnings of chains that never settle. If these differences are preserved during outcome construction, phenotype classes become easier to stabilize later. If they are ignored, the event rule ends up mixing together cases whose only shared property is that they crossed some late threshold of documentary confirmation.

The identification of sentinel documents must also respect documentary ecology. A code entered at discharge can function as a useful retrospective anchor, but it is usually too late to serve as a sentinel for surveillance-oriented interpretation. A nursing note reporting persistent abdominal pain may be an early sentinel in one service, while a radiology recommendation embedded in a report may be the more decisive document in another. Trigger chains therefore need to be contextualized by document type and institutional style. A one-size-fits-all definition of the first meaningful signal will rarely transport well [12]. What can transport more readily is the concept of a sentinel: namely, the first documentary object that initiates a leak-relevant narrative trajectory within a given record ecology.

This perspective also helps clarify the status of actions such as imaging orders, drainage, and antibiotic escalation. These are not merely evidence items. They are trigger-chain events in their own right because they indicate that concern has reached a level that alters care. The timing of these actions can reveal whether a chain is driven by clinical suspicion, by new structural information, or by generalized postoperative deterioration. For example, an imaging order placed before any explicit leak-related language may mark a chain initiated by tacit concern. Antibiotic escalation after a suspicious scan may mark a chain moving from possibility toward managed pathology. Reoperation without preceding imaging may signal a chain in which the record bypassed some diagnostic stages because the clinical situation became too urgent. A trigger-chain representation preserves these differences instead of collapsing them into the undifferentiated fact that action occurred.

There is also an important negative use of trigger chains. Some chains start but never mature. A patient may generate early concern, undergo a partial workup, and then clinically improve or declare another diagnosis. These aborted chains are highly informative. They show which kinds of early signals are common enough to trigger leak concern but insufficiently specific to support a final leak endpoint.

In a binary system, such cases often vanish into the negative class. In a trigger-chain corpus, they remain visible as distinct trajectories. This matters for two reasons. First, it helps identify the border between leak and mimic states. Second, it offers a more realistic substrate for model evaluation. A system that flags such a case should not be interpreted the same way as a system that flags an entirely uneventful recovery.

The language of sentinel documents and trigger chains also improves review workflow. Annotators can be asked not only for final classifications but for the first document that initiated concern and the later events that reinforced or displaced it. This yields richer data than a final verdict alone. It also makes disagreements more interpretable. Reviewers may differ on which document truly initiated the chain, or on whether a later event reinforced leak concern or pointed toward an alternative explanation [13]. Such disagreements are valuable because they indicate where documentary pathways are unstable. Over time, repeated review of trigger chains can show which chains are highly predictive of particular phenotypes and which are common but nonspecific.

Another benefit is that trigger chains can be truncated explicitly. In many real records, not every chain runs to completion within the available data. A patient may be discharged before the path resolves, transferred, or followed outside the visible system. Representing the chain as truncated is more honest than classifying the case as negative by default. It acknowledges that the documentary process, not only the clinical process, was interrupted. Truncated chains often populate the uncertainty region of a phenotype system and deserve separate treatment rather than quiet absorption into event-free status.

For AL corpus design, then, sentinel documents and trigger chains provide a practical language for describing how cases become visible. They link time, document type, and action in a way that is more faithful than static composite logic while remaining compatible with structured analysis. Most importantly, they help distinguish the obvious positive endpoint from the larger set of documentary trajectories in its neighborhood. A good phenotype system is not built only from cases that clearly ended in leak. It is built from understanding how leak-relevant chains begin, grow, abort, and branch. That understanding is one of the most important pieces of information the record can offer.

5. Borderline Cases, Mimic States, and Event Adjacency

Any serious attempt to define AL phenotypes must confront the fact that many postoperative cases live near the edge of the event. These cases are neither rare anomalies nor mere annotation nuisances. They are structurally important because they reveal where the event boundary is under strain. Borderline cases include those with significant postoperative concern but incomplete corroboration, cases with contained or delayed manifestations that never accumulate classic signals, and cases with alternative complications that imitate leak closely enough to trigger similar diagnostic and therapeutic behavior. In a binary framework these cases are typically treated as difficult positives or troublesome negatives. In a corpus-centered phenotype framework they should be treated as residents of an adjacent region whose structure is itself one of the main objects of study.

Mimic states are particularly important within this adjacent region. They are not random postoperative complications. They are the complications that occupy leak-like documentary and clinical pathways for part of their course. A patient with postoperative abscess unrelated to the anastomosis may pass through imaging, antibiotics, drainage, and prolonged inflammatory recovery that closely resemble contained leak [14]. A patient with severe ileus or ischemia may generate the same early concern and the same laboratory deviations that initiate a progressive leak trigger chain. A patient with pneumonia and systemic inflammatory response after major abdominal surgery may enter the record through ICU transfer, sepsis language, and broad treatment escalation that make the case look leak-like at a distance. These are not simply false positives once the final diagnosis differs. They are leak-adjacent states and need explicit place in the corpus.

The absence of explicit mimic representation has several consequences. It inflates the apparent heterogeneity of both positives and negatives, because some negative cases are near leak in the clinical sense while others are far from it. It makes error analysis less useful, because every disagreement with

the final binary label is counted similarly. It also encourages overconfidence in apparently good model performance, since easy negatives dominate the denominator while the difficult negative space remains hidden. When mimic states are given explicit status, evaluation becomes more discriminating. A prediction that confuses contained leak with a localized non-anastomotic abscess is not equivalent to one that confuses leak with a routine recovery. The first reveals a local ambiguity in event adjacency. The second reveals a much broader model weakness.

Borderline cases also expose the problem of evidentiary sufficiency. A patient may have imaging that is suspicious but not definitive, note-level concern, one episode of antibiotic escalation, and a prolonged stay. Another may have persistent systemic disturbance and operative exploration without a documented anastomotic defect. Another may be coded late without strong antecedent evidence. These cases are often where studies diverge from one another, not because their investigators disagree about obvious leak, but because their outcome rules have different tolerances for ambiguity. If such cases are forced into the positive class, the event expands toward clinically plausible but underconfirmed territory. If they are forced into the negative class, the event narrows toward highly resolved cases and may miss the region of greatest surveillance interest. A corpus framework should therefore preserve these borderline states explicitly rather than letting them disappear into whichever class is convenient.

Event adjacency is a useful way to conceptualize this region. Adjacent states are those that can plausibly generate similar documentary chains and similar management responses even when they do not share the same final etiologic label. In AL work, adjacency is more clinically meaningful than simple class opposition. The real question is often not whether a case is leak or non-leak in the abstract, but how far it lies from leak territory and along which dimension. A case may be adjacent because it shares local imaging features, because it shares systemic deterioration, because it shares the same diagnostic workup pathway, or because it remains unresolved despite substantial concern [15]. Representing adjacency allows the corpus to organize difficult negatives and uncertain positives relationally rather than treating them as categorical leftovers.

A particularly important form of adjacency arises when documentation, rather than clinical state, produces the ambiguity. Some cases are adjacent because the chart preserves little of the reasoning that likely occurred at the bedside. An underdocumented postoperative deterioration with late coding may sit near leak not because the evidence was weak, but because the record captured it poorly. Other cases are adjacent because the institution investigated aggressively and left behind many traces despite a non-leak final explanation. If both are simply treated as ambiguous outcomes, one loses the chance to distinguish underdocumentation from genuine clinical uncertainty. A corpus that tracks documentary routes, trigger chains, and observation adequacy can separate these forms of adjacency more effectively.

Borderline cases are also where subgroup and site differences are most likely to appear. A center with earlier imaging access may convert some unresolved or mimic-like cases into contained leak territory by generating more decisive evidence. A center with more conservative intervention thresholds may leave comparable cases in unresolved territory longer. Patient groups with different patterns of follow-up or documentation can likewise accumulate in different border regions even when the underlying clinical states are similar. This means that fairness and transport problems often first manifest at the edges of the atlas, not in its obvious center. A phenotype system that does not preserve adjacency will have difficulty diagnosing these discrepancies because it will see only shifts in binary prevalence or calibration.

One practical advantage of treating adjacency seriously is that it changes how chart review can be used. Reviewers no longer need to answer only whether a case was leak. They can place it in relation to leak. Was it close because of local contamination concern? Close because of systemic physiology? Close because of documentation delay? Or close because it remained unresolved? This produces richer labels and more clinically intelligible disagreement patterns. It also provides a better substrate for future automation, because the model is no longer asked to recover one sharp line where clinicians themselves often perceive a gradient.

A further consequence is that unresolved and adjacent states should not be used only as temporary holding categories. They should be analytically productive. One can examine which features, documentary routes, or operative contexts most often place cases near the boundary. One can study which

adjacent states are most commonly mistaken for contained leak versus progressive leak. One can ask whether certain border regions are more common in some procedure types or some institutions. These analyses deepen understanding of the event itself. They also improve the design of future outcome rules, because the border is where rule refinement is usually most needed [16].

For AL, then, the negative space around the event is part of the event's scientific description. Mimics, unresolved courses, and documentary edge cases are not detritus left over after the positive class has been purified. They are among the main reasons that outcome definition is difficult. A phenotype corpus that names and stabilizes these adjacent regions can produce more honest labels, more interpretable models, and more meaningful validation. A corpus that ignores them will rediscover them later as unexplained error and unstable transport. In that sense, adjacency is not a peripheral concern. It is one of the central structures through which leak becomes definable in records at all.

6. Annotation Grammar and Phenotype Stabilization

If a phenotype system is to be more than a conceptual essay, it must be supported by a disciplined annotation grammar. By annotation grammar I mean the structured language through which reviewers describe evidence, sequence, uncertainty, and phenotype placement in a way that is repeatable across cases and across annotators. Ordinary chart review often yields only a final classification, perhaps with free-text comments. That may suffice for a one-time validation exercise, but it is not enough to stabilize a complex phenotype corpus. Stabilization requires that annotators not only decide where a case belongs, but do so using a shared set of descriptive moves that can later be audited, compared, and revised.

A useful annotation grammar begins with separating extraction from interpretation. Reviewers should first record what documentary elements are present and when they occur: operative findings, note-level suspicions, radiology assertions, management escalations, procedure records, codes, and other relevant traces. Only after that should they interpret how these elements relate to one another. This separation is crucial because disagreements that appear to be about classification sometimes originate much earlier, in what evidence was noticed or counted. If extraction and interpretation are conflated, one cannot tell whether phenotypic instability is due to rule ambiguity or chart-reading inconsistency.

The next layer of the grammar concerns sequence. Reviewers should be able to indicate not only that certain documentary elements are present, but how the case evolved. Which item functioned as the first meaningful signal? Which later item converted suspicion into stronger concern? Was there a branch toward another explanation? Did the chain truncate before resolution? Did retrospective coding add certainty or merely summarize earlier material? An annotation grammar that includes these sequential roles makes it possible to compare cases at the level of narrative architecture rather than only at the level of final label. This is particularly valuable in AL because many phenotypes differ more in how they unfold than in the final inventory of evidence they contain.

A third component is attribution. Many difficult postoperative cases contain multiple plausible sources of deterioration. Reviewers therefore need a controlled way to indicate whether evidence is being attributed directly to the anastomosis, to a likely alternative, to an unresolved differential, or to a mixed situation in which several processes may coexist. Without such a grammar, attribution becomes a hidden mental act inside the final label [17]. This is risky because different reviewers may use similar evidence but assign it to different causal narratives. Making attribution explicit is especially important for mimic states and for borderline cases where the documentary chain strongly resembles leak but does not settle cleanly on an anastomotic explanation.

Confidence and uncertainty also need structured representation. Reviewers often know when they are making a strong assignment and when they are working near the edge of what the chart supports. A disciplined grammar should allow them to state this in a standardized way. Confidence can be attached to the overall case, to individual evidence items, to attribution judgments, or to the final phenotype placement. It should not be treated as an embarrassing optional extra. In a corpus designed for phenotype stabilization, uncertainty is itself information about the shape of the state space. Cases that repeatedly

attract low-confidence assignments may indicate an unstable boundary, a missing phenotype category, or inadequate observation in the record.

The grammar should further support multi-label or gradient placement when appropriate. Some cases are best described as lying between contained leak and unresolved concern, or between progressive leak and mimic. Forcing reviewers to choose one final category discards that information. A richer annotation system can allow a primary phenotype plus a secondary adjacency tag, or a primary region plus an uncertainty modifier indicating which neighboring state is most plausible. This does not make the corpus less usable. It makes the structure of the border more visible. Later analytic stages can always collapse these nuanced placements into simpler endpoints if necessary, but the reverse is impossible once the nuance has been discarded.

Phenotype stabilization depends heavily on this kind of grammar because phenotypes are not discovered fully formed. They are stabilized through repeated use, disagreement, revision, and narrative clarification. In early stages of corpus construction, reviewers may find that two nominal phenotype categories are being used interchangeably or that one category is accumulating several distinct narrative types. The annotation grammar provides the evidence needed to see this. If one class consistently requires different attributions, different trigger chains, or different confidence profiles, it may need to be split. If two classes differ mostly in late coding rather than in substantive documentary evolution, they may need to be merged. Stabilization is therefore not just a statistical exercise. It is a linguistic and interpretive one [18].

Another benefit of explicit grammar is training. New reviewers can be taught not only the phenotype names but the kinds of documentary constructions that justify each placement. They can learn, for example, what usually distinguishes a contained leak chain from a mimic chain that initially resembles it, or what makes a retrospective code evidentially weak versus merely late. Training on labels alone encourages superficial matching. Training on annotation grammar teaches reviewers how to think through a case in the same structured way. This improves both reliability and the scientific transparency of the resulting corpus.

Grammar also creates the possibility of adversarial review. Difficult cases can be deliberately re-read by a second reviewer instructed to challenge the initial placement, not by denying the evidence but by exploring alternative grammatically valid interpretations. This is especially useful for border-zone phenotypes. If a case can be placed in two different regions using equally coherent documentary grammars, the ambiguity is likely genuine and should be preserved. If the alternative interpretation fails under close reading, the boundary may be more stable than it initially seemed. Such exercises help separate true phenotype overlap from preventable reviewer inconsistency.

One of the strongest reasons to invest in annotation grammar is that it improves the interpretability of later model errors. When a model disagrees with the final binary label, the corpus can reveal whether the disagreement occurred in a region of low reviewer confidence, in a case with mixed attribution, in a mimic chain with strong early similarity to leak, or in a documentary path that lacked decisive closure. This is far more informative than simply counting the case as a false positive or false negative. It turns model evaluation into a test of how well the system navigates the same semantic terrain that challenged human reviewers.

Phenotype stabilization, then, is not achieved by declaring categories and hoping they remain coherent. It is achieved by building a review language capable of preserving what matters about difficult postoperative courses. For AL, that means evidence extraction, sequence, attribution, confidence, adjacency, and alternative explanations all need structured representation. Without that grammar, the corpus remains dependent on tacit reviewer judgment and becomes hard to revise or transport. With it, phenotypes become more than names. They become reproducible documentary objects that can be refined, compared, and projected into different operational endpoints without losing their relation to the cases from which they were built.

7. Cross-Institution Translation and Documentary Drift

A phenotype corpus that appears coherent within one institution may become unstable when moved elsewhere, even if the hospitals seem broadly similar in surgical case mix. The reason is that phenotype stabilization depends not only on the underlying clinical states but also on how those states are documented, investigated, and abstracted. Different hospitals produce different documentary ecologies [19]. Some sites emphasize radiology reports. Others rely more heavily on operative findings or bedside note language. Some maintain detailed complication coding. Others encode only what is necessary for billing or quality programs. These differences do not merely add superficial variability. They change how a given postoperative state enters the visible record. Any phenotype system that hopes to travel must therefore distinguish its core state definitions from the local documentary realizations through which those states become recognizable.

Cross-institution translation begins with acknowledging that local extraction logic is not the same as local phenotype logic. A center that identifies contained leaks mainly through CT language and drainage events may still be observing the same underlying territory as a center that identifies them through repeated narrative concern and prolonged medical management. If those sites are forced into identical raw criteria, one of them may appear deficient when in fact it is simply documenting the same state differently. Translation should instead ask whether the same phenotype region can be reconstructed from different but functionally comparable documentary pathways. This requires a corpus architecture in which route of visibility is recorded separately from core phenotype placement.

One practical way to support translation is to define each phenotype through two layers. The first is a core descriptive layer that specifies the clinical and documentary features that make the state distinct from neighboring states. The second is a local realization layer that records the most common evidence pathways through which that state appears in a specific institution. For example, a progressive leak phenotype might be defined core-wise by escalating systemic concern linked to the recent anastomosis, with or without definitive structural confirmation, while one site's realization emphasizes early ICU transitions and repeated inflammatory assessments and another's emphasizes imaging and consult escalation. This layered design permits sites to remain faithful to local records without losing the shared conceptual structure needed for multicenter work.

Documentary drift across time within a single institution poses similar problems. Changes in note templates, radiology phrasing, postoperative pathways, discharge processes, or coding support can alter the visible routes by which cases enter the corpus. A phenotype that once appeared mainly through late operative documentation may later appear earlier through standardized imaging workflows. Another phenotype may become harder to distinguish if narrative nuance is compressed by templated progress notes. If the corpus does not monitor these changes, the apparent prevalence and shape of phenotypes may shift even when the underlying clinical burden remains relatively stable. This is why phenotyping for AL cannot be treated as a one-time definitional achievement. It requires ongoing surveillance of its own documentary infrastructure.

Monitoring drift should therefore occur at the corpus level, not only at the model-performance level [20]. Analysts should examine whether certain documentary routes become more or less common over time, whether unresolved cases accumulate differently after workflow changes, and whether the border between mimic and contained leak moves when new documentation practices appear. Such review is especially important after major institutional interventions such as implementation of enhanced recovery protocols, changes in imaging availability, revised complication coding practices, or integration of new documentation templates. A performance metric alone may miss these semantic shifts if the final binary endpoint remains superficially stable. The corpus, by contrast, can show that the internal composition of the event has changed.

Cross-institution translation is also where fairness and subgroup analysis often become inseparable from documentation practice. Patient populations are not only clinically different. They can also be differentially documented. Some subgroups may have shorter postoperative stays, less complete follow-up, or different pathways of care escalation. These differences can change where cases fall in the corpus

and how often they occupy unresolved or mimic-adjacent territory. If the phenotype system is not built to detect such patterns, subgroup differences will later appear as puzzling calibration shifts or performance disparities. A corpus-centered approach can locate them earlier by asking whether certain patient groups accumulate in specific documentary border zones or whether the routes that make leak visible vary systematically across populations.

Translation also benefits from preserving uncertain and partial equivalence rather than forcing all sites to full harmony. In multicenter work, some institutions may support fine-grained contained and progressive distinctions, while others can only reliably recover a broader leak-adjacent territory because of thinner records. A brittle phenotype system would exclude the latter. A more flexible system can represent partial mappings, allowing some sites to contribute coarser but still meaningful phenotype information. This is preferable to pretending that exact equivalence has been achieved. It also better reflects the real state of routine surgical data, where information richness is uneven. The cost is complexity, but the reward is a more honest multicenter phenotype space.

Another practical issue is how transport failure should be diagnosed. If a model trained on one site performs poorly elsewhere, it is not enough to say that the model failed to generalize. One must ask whether the phenotype definitions remained coherent, whether documentary routes changed, whether the denominator shifted, and whether the target itself was being constructed from different evidence ecologies. A corpus that preserves this information makes such diagnosis possible. One can see, for example, that a “contained leak” phenotype at the external site contains many cases that would have been unresolved under the source-site corpus, or that a mimic region is being overpopulated because alternative diagnoses are documented differently. This is far more informative than a simple decline in discrimination.

Cross-institution translation also requires humility about what can and cannot be harmonized [21]. Some documentary differences may be reducible to shared families of evidence. Others may reflect truly different clinical pathways. A center that routinely reoperates for early concern is not merely documenting the same contained leak differently from a center that observes and drains. The clinical management pathway is part of the state’s lived reality in that setting. Translation, therefore, should not always erase route differences. Sometimes route itself is a meaningful descriptor of local phenotype expression. The key is to know whether route is being used to define the state or to describe one realization of the state. Confusing those roles is one of the main reasons phenotype systems drift under transport.

For AL corpus design, then, cross-institution translation is not a post hoc validation step. It is a constitutive test of whether the phenotypes have been defined at the right level of abstraction. If the phenotype concepts are too tied to one site’s documentation style, they will fracture when moved. If they are too abstract and disconnected from local evidence pathways, they will become empty generalities. The corpus approach seeks a middle position: stable enough to survive translation, specific enough to remain clinically meaningful, and explicit enough about documentary route that drift can be recognized rather than denied. In a field where records are heterogeneous by nature, that is a more realistic goal than expecting simple rule portability.

8. Operational Consequences for Surveillance, Benchmarking, and Scientific Inference

The way an AL corpus is constructed has consequences that extend well beyond label aesthetics. It shapes what kinds of questions can be asked, what kinds of errors are visible, and which uses of the resulting outcome are credible. A phenotype system built from record ecology, denominator discipline, trigger chains, adjacency structure, and annotation grammar does not merely support a different descriptive vocabulary. It changes how surveillance tools, benchmarking exercises, and etiologic analyses can be conducted. These are not secondary benefits. They are among the main reasons to invest in richer outcome construction in the first place.

For surveillance, the most important gain is temporal honesty. A corpus built around sentinel documents and trigger chains can distinguish between cases that became legible early and cases that became clear only after later retrospective consolidation. This means that the same broad event can be projected

differently depending on whether the use case is retrospective analysis or near-real-time monitoring. A surveillance system built on such a corpus can focus on the documentary pathways that actually emerge during actionable windows [22]. It can also report when a patient occupies a leak-adjacent or unresolved state rather than pretending that only fully closed positive events deserve attention. This is more aligned with clinical work, where teams often act under uncertainty and where waiting for perfect documentary closure may be too late to matter.

For benchmarking and quality comparison, the corpus offers a different advantage. It makes explicit which regions of the postoperative state space are being counted and which are being excluded. A benchmarking endpoint can therefore be intentionally narrower, focusing on well-supported clinically significant leak states, while still being understood as one projection from a richer corpus rather than as the whole truth of leak. This is preferable to narrow operational rules that present themselves as if they defined the phenomenon completely. It also permits fairer comparison across sites, because institutions can examine whether differences in rates derive from genuinely different burdens of severe states or from different occupancy of unresolved and mimic-adjacent regions.

Scientific inference also changes under this framework. Many studies use leak as a binary endpoint in association or prediction analyses and then interpret coefficients or feature importance as if they referred to one uniform clinical entity. A corpus-aware phenotype system cautions against that move. The same feature may be associated with one documentary route into contained leak but not with catastrophic operative presentations. Another feature may mainly identify prolonged inflammatory border states. Without phenotype structure, these distinctions are hidden. With it, inference becomes more local and therefore more clinically interpretable. Instead of asking which variables predict leak in general, one can ask which variables distinguish progressive from contained states, or unresolved from clearly non-leak states, or which documentary routes are associated with delayed recognition.

The corpus also changes how implementation failures are understood. If a predictive system performs worse than expected after deployment, the problem may not lie only in calibration or threshold choice. It may be that the deployed environment populates different border regions, documents suspicion differently, or produces different sentinel documents than the development corpus assumed. A binary endpoint offers little help in diagnosing this. A phenotype corpus can reveal whether errors are concentrated in unresolved trajectories, in mimic states, or in cases whose documentary route was rare in training. This makes operational debugging more concrete and reduces the temptation to attribute every performance problem to abstract “distribution shift” without specifying what actually shifted.

Another operational consequence concerns clinician trust. Systems that output only a leak probability often appear opaque, particularly when the clinicians using them know that leak is a heterogeneous phenomenon. A corpus-aware approach permits outputs that refer not only to aggregate risk but to the kind of postoperative documentary trajectory currently unfolding. A case can be described as moving through a contained-leak-like trigger chain, a progressive systemic concern chain, an unresolved inflammatory chain, or a mimic-prone route [23]. These distinctions need not be displayed with cumbersome detail in every interface, but the system’s internal ability to preserve them makes clinician-facing explanations more plausible. Trust is easier to build when the system appears to understand the same difficult middle ground that clinicians struggle with, rather than reducing every abnormal course to one opaque scalar.

There are also consequences for study design. A corpus that preserves border zones and documentary routes allows investigators to choose outcome projections tailored to specific hypotheses while retaining comparability to other projections. One study may examine only severe resolved leak states. Another may study the broader neighborhood of leak-adjacent postoperative deterioration. A third may focus on unresolved cases as a category of their own. If all of these derive from the same underlying corpus architecture, their relationship is much easier to understand than if each study invents a separate leak rule. This supports cumulative science because findings can be located within a shared representational field rather than compared across opaque endpoints.

Importantly, the corpus perspective also clarifies the limits of routine data. Some regions of the postoperative state space may remain weakly observed no matter how carefully the corpus is built. Some

distinctions may rely on bedside nuance or external follow-up that the record captures poorly. A stronger corpus does not eliminate these limits. What it does is make them visible. This matters operationally because it prevents institutions from overinterpreting what their leak labels mean. A center may decide that certain unresolved states should not be used for benchmarking because the local record cannot support confident distinction. Another may decide that those same states are valuable for surveillance precisely because they mark where earlier recognition is most needed. Both decisions are legitimate if the corpus has made the representational limits explicit.

The corpus can also support learning over time. When new workflows or documentation improvements are introduced, institutions can examine not only whether model performance changes, but whether the underlying documentary representation of leak has become richer or more stable. If a new radiology reporting template causes contained leak trajectories to become easier to distinguish from mimics, the corpus should reflect that. If a change in discharge documentation inflates late coding without improving early legibility, the corpus should show that as well. This turns the outcome system into a tool for institutional learning, not merely for model supervision.

For AL research and practice, then, the choice is not between simple binary labels and unmanageable descriptive complexity. The real choice is between a label that hides its own construction and a corpus that makes its construction inspectable [24]. The latter permits more careful surveillance, more interpretable benchmarking, more cumulative scientific inference, and more realistic implementation. It does demand more upfront work. But much of the field's current instability around leak definitions stems precisely from trying to save that effort by moving too quickly to a final endpoint. A corpus-centered approach spends more time on the object so that later modeling is not forced to compensate for a poorly articulated target.

9. Conclusion

Anastomotic leak is difficult to model not simply because it is clinically serious or statistically infrequent, but because it becomes visible through a heterogeneous documentary ecology. The postoperative record does not present leak as a single stable fact. It presents an evolving set of notes, tests, procedures, management shifts, and retrospective abstractions from which leak-related states must be inferred. A binary endpoint can be useful within this environment, but only if one first understands what kind of corpus of postoperative cases made that endpoint possible. Without such understanding, event definitions risk becoming compressed summaries of local documentation habits rather than transparent representations of postoperative failure.

This paper has argued for a different starting point. Instead of treating outcome definition as a thresholding exercise over a finished dataset, it treats outcome definition as the architectural work of building a corpus. That work includes denominator construction, preservation of record ecology, identification of sentinel documents and trigger chains, representation of mimic and unresolved border zones, development of annotation grammar, and stabilization of phenotype regions within a heterogeneous documentary field. Under this view, the final binary event is one projection from a richer representational object rather than the primary scientific object in its own right.

A corpus-centered approach offers several advantages. It makes explicit the distinction between clinical state and documentary route of recognition. It retains the negative and ambiguous space around leak rather than collapsing it into residual classes. It improves interpretation of model errors by locating them in specific border regions and documentary pathways. It supports translation across institutions by allowing phenotype concepts to remain stable even when local realizations differ. It also clarifies where routine data remain insufficient, which is itself a valuable scientific result. The implication is that progress in AL informatics depends at least as much on the quality of corpus design as on the sophistication of prediction algorithms. Better models trained on semantically thin labels may produce attractive metrics while learning unstable targets. Better corpora, by contrast, make later modeling more honest, more interpretable, and more transferable. For a complication whose difficulty lives largely in its documentary middle ground, that is not a secondary concern. It is the methodological center of the problem.

A stable science of leak phenotyping will therefore emerge not from one universally accepted rule, but from more disciplined ways of constructing, describing, and revising the case corpora from which those rules are drawn [25].

References

- [1] M. Sugara, V. Peruvaje, S. Krishnamurthy, K. V. V. Kumar, and R. Chowdappa, "Effect of preoperative biliary stenting on outcomes of pancreaticoduodenectomy.," *Indian journal of surgical oncology*, vol. 12, pp. 554–560, 7 2021.
- [2] C. Papadimitriou, W. Martin, A. E. Dellis, C. Deliveliotis, and I. Mitsogiannis, "Twenty-four-month efficacy of ureteroureterostomy combined with unilateral nephrostomy following radical cystectomy.," *Cureus*, vol. 14, pp. e30478–e30478, 10 2022.
- [3] B. Ayhan, M. Erikoglu, S. S. Tavli, and H. Toy, "A comparison of the application of fibrin glue and adhesive film for repair of anastomotic leaks in the rat.," *Clinical and investigative medicine. Medecine clinique et experimentale*, vol. 35, pp. 216–222, 8 2012.
- [4] E. Ugliciono, F. Rebecchi, C. Vicentini, A. Salzano, and M. Morino, "Cost-effectiveness analysis of revisional roux-en-y gastric bypass: laparoscopic vs. robot assisted.," *Updates in surgery*, vol. 75, pp. 189–196, 11 2022.
- [5] E. Pai and T. Kumar, "Comment on "oncological outcomes after anastomotic leakage after surgery for colon or rectal cancer: Increased risk of local recurrence".," *Annals of surgery open : perspectives of surgical history, education, and clinical approaches*, vol. 3, pp. e171–e171, 6 2022.
- [6] R. Souche, M. Nayeri, R. Chati, E. Huet, I. Donici, J.-J. Tuech, F. Borie, M. Prudhomme, S. Jaber, and J.-M. Fabre, "Thoracoscopy in prone position with two-lung ventilation compared to conventional thoracotomy during ivor lewis procedure: a multicenter case-control study," *Surgical endoscopy*, vol. 34, pp. 142–152, 3 2019.
- [7] A. Mathew, G. Kaushal, D. Ramachandra, N. R. Rakesh, and P. Dhar, "Staple line dehiscence in gastric conduit following esophagectomy: A complication conspicuously missing a mention it deserves.," 1 2022.
- [8] L. Claassen, F. van Workum, and C. Rosman, "Learning curve and postoperative outcomes of minimally invasive esophagectomy.," *Journal of Thoracic Disease*, vol. 11, 12 2018.
- [9] 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.
- [10] J. C. Alverdy, "Rationale behind phosphate therapy to modulate the gut microbiome and protect against surgery-related infection.," *Microbiota and host*, vol. 1, 9 2023.
- [11] T. Kamal, A. Pai, V. R. Velchuru, M. Zawadzki, J. J. Park, S. J. Marecik, H. Abcarian, and L. M. Prasad, "Should anastomotic assessment with flexible sigmoidoscopy be routine following laparoscopic restorative left colorectal resection," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 17, pp. 160–164, 1 2015.
- [12] U. Aday, E. Gündeş, H. Çiyiltepe, D. A. Çetin, S. Gülmez, A. S. Senger, K. C. Değer, and M. Duman, "Examination of anastomotic leak with aqueous contrast swallow after total gastrectomy: Should it be carried out routinely?," *Contemporary oncology (Poznan, Poland)*, vol. 21, pp. 224–227, 9 2017.
- [13] E. L. Watkins, N. Schellack, V. Abraham, and B. Bebington, "Men and those with a history of smoking are associated with the development of postoperative ileus following elective colorectal cancer resection at a private academic hospital in johannesburg, south africa: A retrospective cohort study.," *Frontiers in surgery*, vol. 8, pp. 667124–, 6 2021.
- [14] 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.
- [15] N. El-Sourani, H. Bruns, A. Troja, H.-R. Raab, and D. Antolovic, "Routine use of contrast swallow after total gastrectomy and esophagectomy: Is it justified?," *Polish journal of radiology*, vol. 82, pp. 170–173, 3 2017.
- [16] K. I. Deen, "Annual academic sessions august 2012," *Sri Lanka Journal of Surgery*, vol. 30, pp. 33–, 1 2013.
- [17] F. Dip, L. Boni, M. Bouvet, T. Carus, M. Diana, J. Falco, G. C. Gurtner, T. Ishizawa, N. Kokudo, E. L. Menzo, P. S. Low, J. Masia, D. D. Muehrcke, F. A. Papay, C. Pulitano, S. Schneider-Korath, D. A. Sherwinter, G. Spinoglio, L. P. S. Stassen, Y. Urano, A. L. Vahrmeijer, E. Vibert, J. M. Warram, S. D. Wexner, K. P. White, and R. J. Rosenthal, "Consensus conference statement on the general use of near-infrared fluorescence imaging and indocyanine green guided surgery: Results of a modified delphi study.," *Annals of surgery*, vol. 275, pp. 685–691, 11 2020.

- [18] N. Angamuthu, F. Froghi, S. Shah, R. Mirnezami, and J. Knowles, “188 right hemicolectomy with end-to-side circular stapled ileo-colic anastomosis: Evaluation of outcomes in a series of 55 consecutive patients,” *British Journal of Surgery*, vol. 108, 9 2021.
- [19] S. Bischofberger, F. Adshead, K. Moore, M. Kocaman, G. Casali, C. Tong, S. Roy, M. Collins, and W. Brunner, “Assessing the environmental impact of an anastomotic leak care pathway,” *Surgery open science*, vol. 14, pp. 81–86, 7 2023.
- [20] A. Brownlee, E. Bromberg, and M. Roslin, “Outcomes in patients with helicobacter pylori undergoing laparoscopic sleeve gastrectomy,” *Obesity surgery*, vol. 25, pp. 2276–2279, 5 2015.
- [21] Çağatay Çetinkaya, Z. Bilgi, S. Aslan, and H. F. Batreil, “Evolution of a minimally invasive oesophagectomy program - effective complication management is key,” *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques*, vol. 18, pp. 481–486, 8 2023.
- [22] A. Rieth, B. Balogh, and T. Kovács, “Jejuno-ileal atresia: evaluation of the efficacy of laparoscopic approach,” *Journal of Pediatric Endoscopic Surgery*, vol. 3, pp. 33–37, 9 2020.
- [23] M. K. Kim, D. D. Won, J.-K. Lee, W.-K. Kang, J.-G. Kim, and S. T. Oh, “Comparative study between transanal tube and loop ileostomy in low anterior resection for mid rectal cancer: a retrospective single center trial,” *Annals of surgical treatment and research*, vol. 88, pp. 260–268, 4 2015.
- [24] M. Zawadzki, M. Rząca, R. Czarnecki, Z. Obuszko, K. Jacyna, L. Stewart, and W. Witkiewicz, “Beginning robotic assisted colorectal surgery - it's harder than it looks!,” *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques*, vol. 9, pp. 562–568, 9 2014.
- [25] S. Lauricella, D. Peyser, F. M. Carrano, and P. Sylla, “Intraluminal anastomotic assessment using indocyanine green near-infrared imaging for left-sided colonic and rectal resections: a systematic review,” *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, vol. 27, pp. 615–625, 1 2023.