

Original Research

Toward Reproducible and Transportable Anastomotic Leak Intelligence: Benchmark Design, Distribution Shift, and Decision Reliability

Arif Setiawan¹, Budi Santoso² and Dimas Prakoso³

¹Department of Computer Science, Universitas Tidar, Jalan Kapten Suparman No. 39, Magelang 56116, Indonesia.

²Department of Information Systems, Universitas Trunojoyo Madura, Jalan Raya Telang, Bangkalan 69162, Indonesia.

³Department of Informatics Engineering, Universitas Siliwangi, Jalan Siliwangi No. 24, Tasikmalaya 46115, Indonesia.

Abstract

Postoperative complications remain one of the most analytically demanding targets in clinical prediction because they emerge from interacting physiologic, procedural, and organizational processes. Anastomotic leak is a particularly difficult example. It is uncommon enough to create statistical imbalance, serious enough to motivate intervention, and clinically heterogeneous enough that retrospective labels do not always map cleanly onto prospective decisions. For these reasons, model development alone does not resolve the central translational problem. What is needed is a principled account of how predictive systems should be specified, benchmarked, evaluated, and maintained when the intended use environment is a hospital rather than a static test set. This paper develops a technical framework for the transition from isolated proof-of-concept performance to reproducible and transportable surgical prediction infrastructure. The analysis treats leak prediction as a layered problem involving phenotype construction, observation process modeling, institution-sensitive generalization, rare-event calibration, utility-aware operating rules, and prospective evaluation under behavioral feedback. Rather than asking only whether a classifier separates cases from controls, the paper asks which statistical objects remain stable across centers, which measures of performance retain operational meaning under prevalence variation, and which validation designs can distinguish genuine predictive robustness from local data familiarity. The proposed framework argues for benchmark designs that preserve semantic traceability of features, explicit separation of discrimination from calibration, hierarchical treatment of site heterogeneity, and deployment governance that treats drift as expected rather than exceptional. The aim is neither to reject machine learning nor to assume its inevitability, but to specify the conditions under which surgical prediction could become a durable component of perioperative intelligence.

1. Introduction

Most clinical prediction problems are not difficult because one lacks variables [1]. They are difficult because the variables are generated by a care process whose internal logic changes across patients, units, and institutions. Anastomotic leak prediction sits squarely in that category. The event itself is pathophysiologically complex, but the deeper analytical difficulty is that the observable evidence of leak is inseparable from surveillance intensity, clinician suspicion, operative context, and documentation practice. A patient who undergoes serial laboratory testing, postoperative critical care admission, and prompt cross-sectional imaging leaves a far richer data trail than a patient managed on a routine ward pathway. A model trained on such data is learning from physiology, but also from the organizational reaction to physiology. The statistical consequences are substantial. Feature importance can become entangled with observation intensity. Calibration can depend on local ordering practices. Apparent transportability can disappear when another hospital measures similar phenomena in slightly different forms. These are not peripheral concerns. They define the problem.

Table 1: Core Components of a Reproducible Benchmark

Component	Description	Example	Risk if Omitted
Cohort (C)	Inclusion logic	Colorectal resections	Selection bias
Phenotype ($\mathcal{Y}$)	Outcome definition	Coded vs adjudicated leak	Label inconsistency
Time ($\mathcal{T}$)	Indexing scheme	24h postoperative window	Temporal leakage
Features ($\mathcal{F}$)	Variable semantics	Lab + ICU status	Misinterpretation
Missingness ($\mathcal{M}$)	Handling strategy	Median + indicators	Hidden bias
Evaluation ($\mathcal{E}$)	Validation protocol	Site-held-out split	Overfitting

Table 2: Sources of Distribution Shift in Surgical Prediction

Shift Type	Definition	Surgical Example	Impact
Covariate	$P(X)$ changes	Different case mix	Ranking drift
Label	$P(Y)$ changes	Leak prevalence variation	Calibration error
Concept	$P(Y X)$ changes	Technique differences	Model failure
Observation	Measurement differs	Imaging frequency	Feature distortion

Table 3: Performance Dimensions for Rare-Event Prediction

Metric	Captures	Strength	Limitation
AUC	Ranking	Threshold-free	Insensitive to prevalence
PR-AUC	Precision-recall	Workload relevance	Prevalence dependent
Brier Score	Calibration	Probabilistic accuracy	Scale sensitive
ECE	Calibration error	Interpretability	Bin instability

Table 4: Key Variables in Early Leak Prediction

Variable	Type	Clinical Meaning	Stability
WBC	Laboratory	Inflammation	Moderate
Lactate	Laboratory	Perfusion stress	Moderate
ICU Admission	Process	Severity/protocol	Low
Albumin	Baseline	Nutritional reserve	High
Surgery Type	Procedural	Risk variation	High

Table 5: Validation Strategies and Their Targets

Method	Split Type	Measures	Limitation
Random	Mixed samples	General fit	Inflated optimism
Temporal	Time-based	Stability over time	Same site bias
Site-held-out	External center	Transportability	Data access
Prospective	Live deployment	Real-world effect	Complex design

For that reason, a productive research agenda cannot stop at asking whether an algorithm reaches a respectable area under the receiver operating characteristic curve. The more consequential questions arise one level below and one level above that metric. One level below lies the issue of object definition. What precisely is being predicted: coded leak, adjudicated leak, clinically actionable deterioration suggestive of leak, or the downstream bundle of events by which teams discover and manage leak? One level above lies the issue of operational semantics. What does a quoted probability mean when prevalence varies across settings, when alert cost is unevenly distributed, and when the very act of issuing a

Table 6: Calibration Methods for Clinical Prediction

Method	Approach	Advantage	Limitation
Platt Scaling	Logistic map	Simple	Limited flexibility
Isotonic	Nonparametric	Flexible	Overfitting risk
Spline	Smooth fit	Balanced	Needs tuning
Temperature	Scaling factor	Stable	Less expressive

Table 7: Governance Layers in Prediction Systems

Layer	Role	Example Change	Frequency
Benchmark	Data definition	New phenotype	Rare
Model	Predictor	Retraining	Occasional
Calibration	Probability mapping	Recalibration	Frequent
Policy	Threshold/action	Alert tuning	Frequent

Table 8: Prospective Evaluation Components

Component	Focus	Example	Purpose
Silent Phase	Logging	Hidden scores	Baseline behavior
Process Metrics	Workflow	Imaging rates	Mechanism insight
Clinical Outcomes	Patient impact	Time to detection	Effectiveness
Feedback	Clinician response	Overrides	Model refinement

Table 9: Utility Trade-offs in Threshold Selection

Factor	High Threshold	Low Threshold	Trade-off
Sensitivity	Lower	Higher	Missed cases
Specificity	Higher	Lower	False alarms
Workload	Lower	Higher	Resource strain
PPV	Higher	Lower	Clinical trust

score changes subsequent observation? A technical literature that ignores those questions can still accumulate promising numbers, but it has not yet identified the conditions under which those numbers would remain meaningful outside the retrospective environment that produced them.

This paper develops a different angle on the translational problem [2]. Instead of centering the discussion on interface design or on another competition between model classes, it focuses on the mathematical and organizational properties that a benchmark must possess if the resulting predictor is to be reproducible and transportable. Reproducibility matters because a field without stable benchmarks cannot compare methods coherently. Transportability matters because a model that performs adequately in a single institutional data ecology may not survive even modest shifts in measurement patterns or patient mix. The two concerns are linked. A benchmark that does not preserve feature semantics, time-indexed availability, and label construction logic is difficult to reproduce. A predictor built on such a benchmark is equally difficult to move.

The argument proceeds from a simple observation. In surgical machine learning, the unit of failure is often not the algorithm in isolation, but the coupling between an algorithm and a local data-generating system. An operation performed in one hospital, documented in one electronic record, and followed by one style of postoperative escalation creates a specific statistical world. A model trained in that world may look more robust than it actually is because internal validation samples from the same conventions, the same coding habits, and the same rhythms of observation. External deterioration is then described as

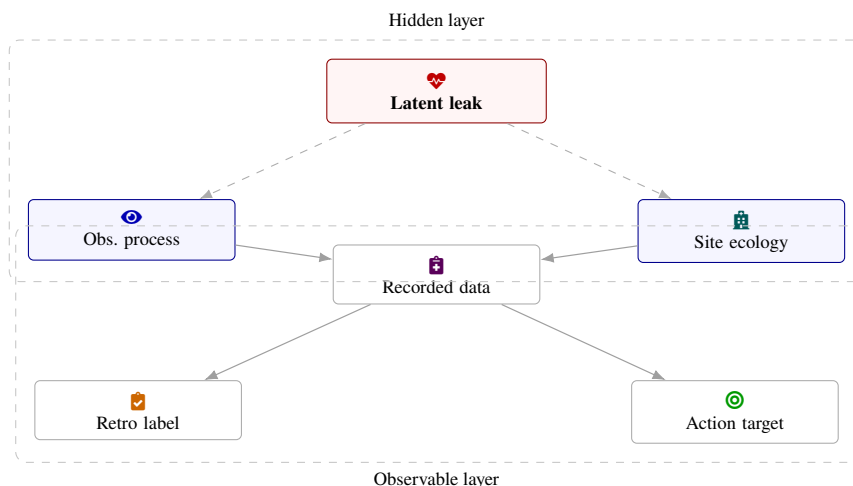


Figure 1: Leak prediction can be represented as a layered object: a latent failure process is only partially rendered into the chart through observation intensity and institution-specific care processes. The benchmarked model therefore acts on recorded signals and derived labels rather than on the biological event alone.

if it were an unfortunate but secondary detail. In reality, it reveals that the original benchmark targeted a local surrogate of the problem rather than the more general problem itself.

A rigorous response requires re-specification at several levels. The outcome definition must be sharpened so that model learning is not dominated by hospital-specific proxies for concern. Feature construction must preserve the time at which information becomes knowable, because many performance gains in retrospective work come from variables whose availability is less stable than their clinical interpretation suggests. Validation must distinguish discrimination, calibration, ranking stability, and institution-conditional utility, because these dimensions diverge under prevalence and measurement shifts. Prospective evaluation must recognize that a live predictor is an intervention on the information environment, not a passive observer of it. Governance must assume that distributional movement is normal in surgical services, where case mix, staffing, perioperative pathways, and documentation systems evolve [3].

The motivation for this perspective is not skepticism about computation. On the contrary, postoperative risk estimation is one of the domains where computational synthesis is plausibly useful because no individual clinician integrates all perioperative measurements formally and repeatedly over time. The problem is that proof-of-concept studies often compress several layers of complexity into a single question: can the model separate future leak from non-leak cases better than chance? An affirmative answer is informative, but only at the lowest level of translation. It indicates that statistical signal exists. It does not yet specify whether the signal is attached to a stable clinical object, whether the score retains meaning across sites, whether the probabilities are communicable, or whether the resulting system can remain reliable under routine maintenance.

The paper is organized around those broader requirements. The next section defines the statistical object of leak prediction in a way that separates latent pathophysiology from the observation process that renders it partially visible. The following section develops a benchmark architecture for reproducible development and uses the available proof-of-concept evidence as a concrete anchor for what such a benchmark must preserve. Subsequent sections examine transportability under institutional heterogeneity, reliability under rare-event operating constraints, prospective evaluation when prediction alters care pathways, and lifecycle governance for surgical prediction systems. Throughout, the tone is deliberately limited. The claim is not that current models are ready for general use, nor that the field should abandon algorithmic work. The claim is narrower and more technical: the next stage of progress depends less

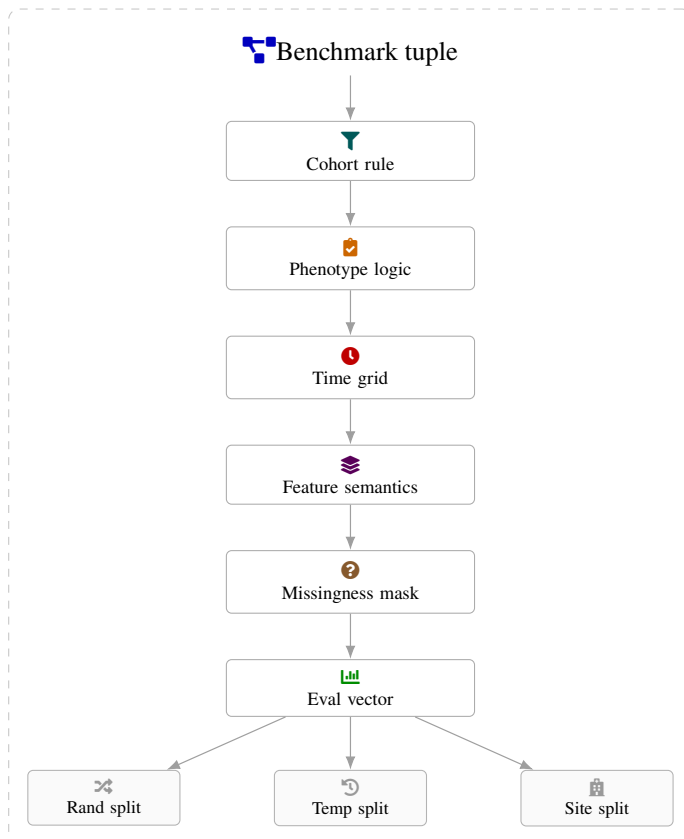


Figure 2: A reproducible benchmark is not just a feature matrix; it is a versioned tuple linking cohort construction, phenotype logic, time-indexed observability, semantic feature definitions, missingness treatment, and multi-axis evaluation. This structure makes later algorithm comparisons interpretable rather than benchmark-dependent.

on extracting another small performance gain from familiar data and more on formalizing the benchmark, validation, and maintenance properties that allow a predictive result to persist beyond the dataset in which it first appeared.

2. The Statistical Object of Postoperative Leak Prediction

A recurrent weakness in surgical prediction studies is that the target variable is treated as though it were self-evident. In practice, anastomotic leak is not directly observed as a single binary fact at the moment relevant decisions must be made. What clinicians encounter is a cascade of imperfect signals: inflammatory change, hemodynamic deviation, radiographic uncertainty, operative findings, delayed gastrointestinal recovery, drain abnormalities, and escalating concern. Some patients are diagnosed quickly because surveillance is intense. Others are recognized late because their trajectory initially resembles routine postoperative stress [4]. A retrospective label recorded at discharge or coded after intervention conceals that heterogeneity. The first requirement for a transportable benchmark is therefore to specify the prediction target as a latent event process rather than as an unexamined binary endpoint.

Let $e_i(t)$ denote the unobserved pathophysiologic state of patient i at postoperative time t , with $e_i(t) = 1$ indicating that an anastomotic failure process sufficient to qualify as leak is present at that time. Let $o_i(t)$ denote the observation process, comprising laboratory orders, care-location changes, clinician examinations, imaging studies, and other acts by which evidence becomes visible. The record

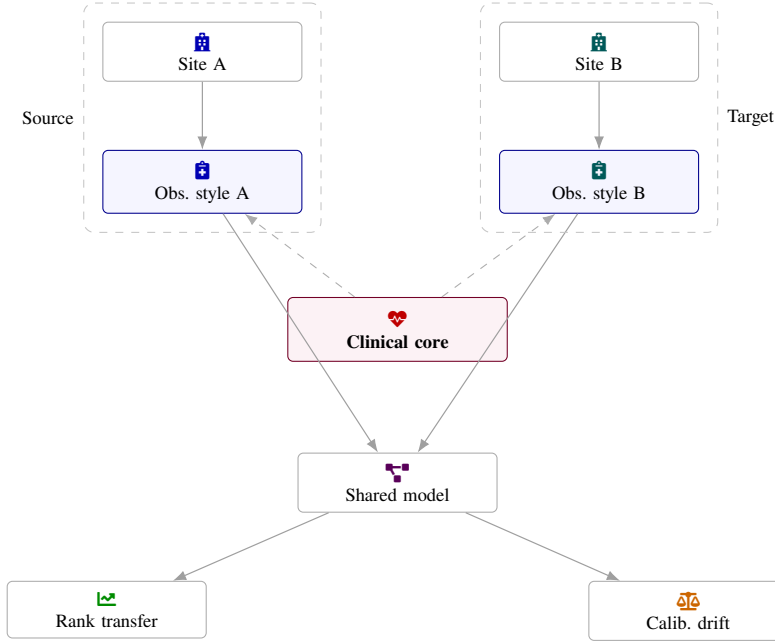


Figure 3: Transportability depends on whether the model learns a stable clinical core or a site-specific observation habit. Cross-center transfer may preserve relative ranking while still losing calibration when prevalence, measurement density, or documentation pathways shift between hospitals.

available to the model is not $e_i(t)$ itself, but $x_i(t) = \mathcal{O}(e_i(t), o_i(t), u_i(t))$, where $u_i(t)$ contains baseline and procedural context. In a stylized form,

$$\begin{aligned} x_i(t) &= \mathcal{O}(e_i(t), o_i(t), u_i(t)) \\ y_i &= \mathbb{I}(\exists t \leq T : \mathcal{A}(e_i(t), o_i(t)) = 1), \end{aligned} \quad (2.1)$$

where y_i is the retrospective label and $\mathcal{A}$ is the adjudication rule or proxy that converts the latent process into an observed endpoint by time horizon T . This representation makes clear that the label is a function not only of the complication, but of the way the health system encounters it. Two hospitals with identical pathophysiologic incidence but different imaging thresholds may therefore generate different empirical labels.

This distinction changes how a benchmark should be built. If the objective is to predict coded leak, then local coding conventions are part of the target. If the objective is to identify clinically actionable deterioration related to anastomotic failure, then those conventions are nuisance structure that should be attenuated. A transportable benchmark needs to declare which of these objects it pursues. Without that declaration, model comparisons are underidentified because one method may be better at learning institutional pathways to diagnosis while another may be better at learning the underlying physiologic signature. Apparent superiority then becomes partly semantic rather than strictly predictive.

Time-indexed observability is equally important. A feature used in training has meaning only if its timestamp corresponds to the moment at which it would be available in the intended use setting. Many perioperative variables have deceptively stable names but unstable temporal semantics. A postoperative lactate measured in the recovery unit is one object; a lactate obtained after clinical deterioration on the ward is another. Similarly, intensive care admission can function as a proxy for operative complexity, early instability, or merely institutional bed-assignment policy [5]. Time-resolved feature design should therefore treat each variable as the output of a measurement process, not simply as a column in a table.

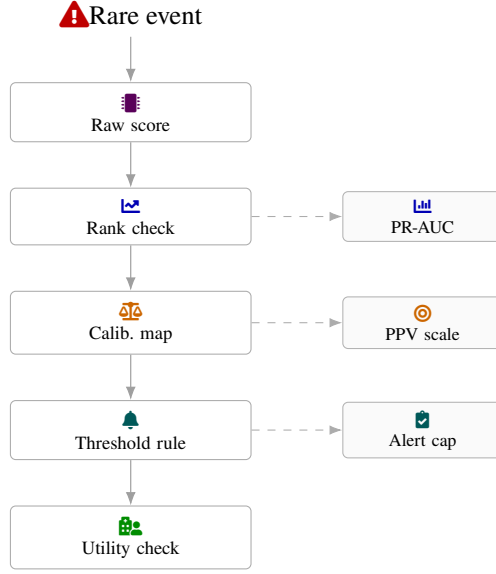


Figure 4: Under low leak prevalence, reliability is assembled in stages: ranking quality, probability calibration, operating-threshold selection, and workload-aware utility must all be checked separately. A model with acceptable discrimination can still fail operationally if its probabilities are mis-scaled or if alert burden becomes disproportionate to clinical capacity.

One can formalize the development problem by defining a sequence of sigma-fields $\mathcal{F}_t$, each representing all information legitimately available by time t . The predictor f_t must then be measurable with respect to $\mathcal{F}_t$. In other words, the score at time t should depend only on information known by that time. The empirical risk of a staged predictor family $\{f_t\}_{t=0}^T$ may be written as

$$\begin{aligned} \mathcal{R}(f) = & \sum_{t=0}^T w_t \mathbb{E}[\ell(f_t(X_i^{\leq t}), Y_i^{>t})] \\ & + \lambda \sum_{t=1}^T \mathbb{E}[d(f_t, f_{t-1})], \end{aligned} \quad (2.2)$$

where $X_i^{\leq t}$ denotes the information history up to t , $Y_i^{>t}$ denotes future occurrence beyond t , w_t weights clinically relevant times, and the penalty term discourages unstable score oscillation between adjacent stages. The second term is not an aesthetic addition. A clinically interpretable system should not reverse itself dramatically when only trivial new information has arrived. Score volatility matters because clinicians experience prediction over time, not as a single isolated event.

The rare-event structure of leak also requires that one distinguish ranking from probability assignment. Under low prevalence, it is possible to achieve respectable ranking while producing poorly scaled probabilities. In deployment, however, the scale matters because actions are tied to estimated risk, not merely to ordering. Let $s_i = f(X_i)$ be a raw score. A benchmark should evaluate both the concordance properties of s_i and the calibration map $g(s_i)$ that turns it into an estimated probability. These are different statistical objects. The first concerns whether higher scores tend to belong to leak cases. The second concerns whether a score interpreted as 0.25 behaves empirically like 25%. In the presence of site-specific prevalence or measurement shifts, the former can remain stable while the latter deteriorates substantially.

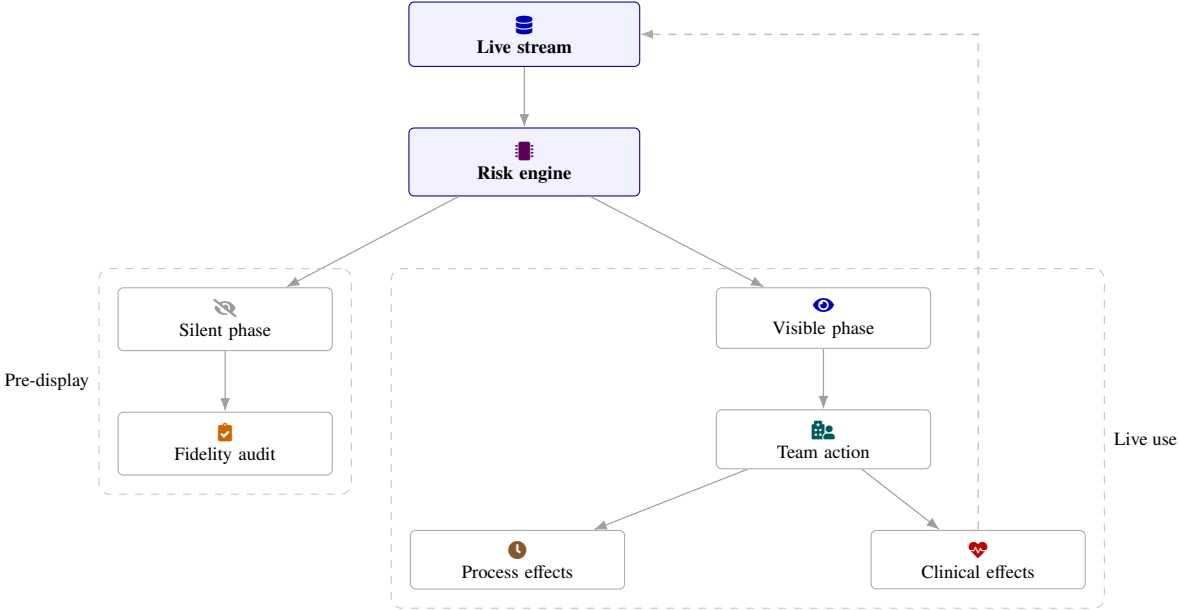


Figure 5: Prospective evaluation should begin with silent execution to verify real-time data fidelity before the score is shown clinically. Once visible, the model becomes part of the care pathway, so downstream actions, altered detection timing, and feedback into the future record all become part of the evaluation target.

A further complication is that leak prediction is not purely a patient-level problem. There is a structural hospital effect. Procedural mix, perioperative pathways, surgeon preference, laboratory panels, ward escalation norms, and discharge behavior all shape the empirical environment [6]. The data-generating mechanism should therefore be indexed by site h . A natural abstraction is

$$P_h(X, Y) = \int P_h(X | Z, \Theta_h) P_h(Y | Z, \Theta_h) dP_h(Z), \quad (2.3)$$

where Z denotes latent clinical factors and Θ_h denotes hospital-specific observation and management parameters. Internal validation estimates performance under a fixed h . Transportability asks what survives when h changes. Seen this way, external validation is not an optional prestige exercise. It is the empirical method by which one estimates sensitivity of the predictor to Θ_h .

A robust benchmark should therefore separate at least four layers: latent complication risk, observation intensity, institutional context, and action relevance. Latent complication risk is the biological object one ultimately cares about. Observation intensity determines how much of that object becomes visible in the record. Institutional context determines the mapping from clinical events to coded and documented data. Action relevance determines which errors matter more heavily at which times. Most current datasets conflate the four. That does not make them useless, but it means that any reported performance number is a compound property of physiology and process.

This perspective also changes what interpretability should mean. In low-level modeling discussions, interpretability is often reduced to whether coefficients can be printed or whether a feature ranking can be shown. For transportable surgical prediction, interpretability should instead mean whether the model's behavior can be decomposed along clinically and operationally stable axes. A variable is not interpretable merely because it has a human-readable name [7]. It is interpretable if one can explain how its meaning is preserved across time windows and institutions. White blood cell count is interpretable in that stronger sense if its collection window, transformation, and handling of missingness are defined semantically rather than merely computationally. Intensive care admission is interpretable only if one

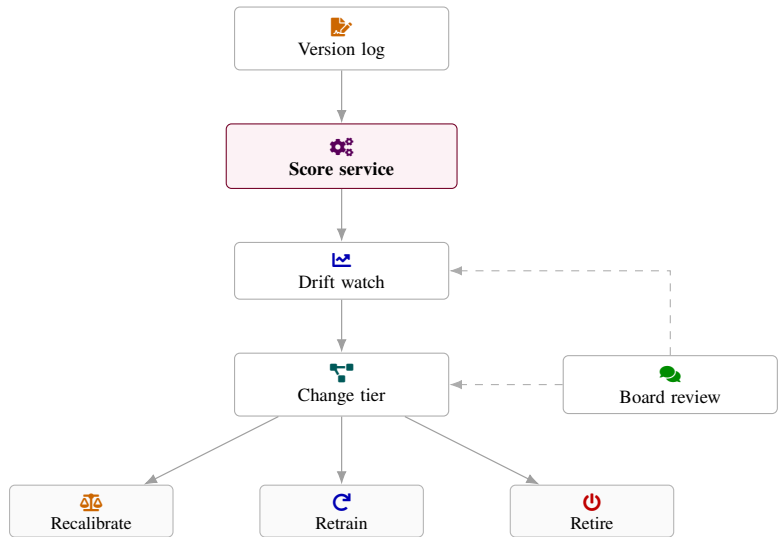


Figure 6: Lifecycle governance treats drift as a routine operating condition rather than as an exceptional failure. A maintained service therefore needs explicit versioning, continuous monitoring, tiered change responses, and a joint clinical-technical review structure that can justify recalibration, retraining, or retirement.

acknowledges that its meaning can differ between a center that routinely recovers major cases in critical care and a center that reserves ICU beds for instability.

Once prediction is framed as estimation over a latent event under observation constraints, benchmark design becomes less ad hoc. The relevant question is no longer whether a particular algorithm beats a baseline by a few hundredths of AUC. The question is whether the benchmark captures a stable enough clinical object that an algorithm trained on it could still mean the same thing when moved, recalibrated, and evaluated prospectively. That is the object to which the next section turns.

3. Benchmark Architecture and Reproducible Development

A benchmark worthy of sustained use in surgical prediction must do more than gather variables and publish a held-out metric. It must preserve the semantic and computational traceability of every major design choice. That requirement is unusually strict in perioperative data because apparently small preprocessing decisions can alter the meaning of features substantially. Winsorizing a laboratory value, recoding an operative descriptor, collapsing time windows, or imputing missing outpatient measurements each changes the object under study. If these choices are implicit, then later investigators cannot determine whether performance differences reflect model architecture or benchmark drift. A reproducible field needs benchmarks whose internal transformations are visible enough that one can distinguish methodological innovation from hidden redefinition.

The available proof-of-concept evidence gives a useful anchor for such a benchmark. One publicly reproducible study identified 3,331 colorectal resection admissions from a large critical care database, observed an anastomotic leak rate of 7.6%, and reported that structured data available within 24 hours of surgery supported moderate predictive performance, with the strongest configurations reaching AUC 0.772, PR-AUC values around 0.30 to 0.33, and influential predictors including postoperative white blood cell count, intensive care admission, postoperative lactate, surgery type, and preoperative albumin, while also emphasizing unresolved issues of external validation, subgroup analysis, and practical clinical uptake, as Sadanandan (2024) described [8]. That single result is informative not because it resolves the translational problem, but because it reveals which elements a benchmark must expose. Cohort assembly, event prevalence, temporal feature windowing, missingness handling, calibration behavior, and

comparative algorithm performance are all part of the benchmark itself rather than neutral background details [9].

A reproducible benchmark should be defined as a tuple $\mathcal{B} = (C, \mathcal{Y}, \mathcal{T}, \mathcal{F}, \mathcal{M}, \mathcal{E})$, where C is the cohort extraction rule, $\mathcal{Y}$ the phenotype logic, $\mathcal{T}$ the time-indexing scheme, $\mathcal{F}$ the feature semantics, $\mathcal{M}$ the missingness policy, and $\mathcal{E}$ the evaluation protocol. Any two studies that differ materially in one component should be understood as operating on different benchmarks even if they share a source database. In symbolically compressed form,

$$\mathcal{B} = (C, \mathcal{Y}, \mathcal{T}, \mathcal{F}, \mathcal{M}, \mathcal{E})$$

$$\Delta(\mathcal{B}_1, \mathcal{B}_2) = \sum_k \omega_k \delta_k(B_{1k}, B_{2k}), \quad (3.1)$$

where δ_k measures discrepancy in component k and ω_k weights the consequences of that discrepancy. The point of this formalism is practical: reproducibility is not binary. Two analyses can be reproducible in code while targeting meaningfully different statistical problems because their benchmark tuples diverge.

Cohort definition is the first layer at which reproducibility can fail. A colorectal resection cohort extracted from billing or procedural codes may include multiple care contexts, staged operations, admissions with repeated interventions, and heterogeneous postoperative surveillance patterns. A benchmark should therefore document not only inclusion and exclusion criteria but also encounter granularity, handling of repeated admissions, and rules for merging perioperative records. Failure to specify whether the unit of analysis is patient, admission, procedure, or anastomosis produces ambiguity in both prevalence and variance estimation. In rare-event prediction, these ambiguities are amplified because a small number of recoded cases can shift precision-recall behavior appreciably.

Phenotype logic is even more decisive. A benchmark that defines leak using diagnosis codes alone is learning one object. A benchmark that uses a composite of codes, imaging evidence, reintervention, and chart adjudication is learning another. Neither is automatically superior in all contexts, but the benchmark must disclose the trade-off. Coded labels scale easily but may embed local billing logic. Adjudicated labels are clinically sharper but expensive and potentially less reproducible across teams unless adjudication criteria are themselves standardized. One pragmatic architecture is staged phenotype construction: an initial broad label for cohort-wide model development, a nested adjudicated subset for label-noise estimation, and a prospectively maintained operational endpoint for live monitoring [10]. Such a design does not eliminate ambiguity, but it quantifies it.

Temporal design should be treated with the same discipline. Many retrospective pipelines define a window such as the first 24 postoperative hours and regard all measurements within it as equally legitimate. In reality, the statistical meaning of that window depends on care trajectory. A patient extubated promptly to a routine floor bed and a patient taken to critical care after a difficult operation both contribute observations from the same nominal time span, yet the observational density and clinical interpretation differ. A robust benchmark should preserve within-window timing or at least encode summary dynamics such as earliest value, latest value, slope, variability, and measurement count. Otherwise the model may learn from the fact of being intensively observed without making that dependence visible.

Feature semantics must also be declared in a site-agnostic language. Semantic traceability asks more than whether a variable label is understandable. It asks whether another institution could instantiate the same variable without guessing. For continuous features, that requires units, transformation rules, handling of outliers, and time anchoring. For categorical features, that requires codebook mappings and logic for rare or unknown categories. For aggregate features, such as “postoperative instability” or “surgery type,” it requires compositional definitions. A benchmark that simply releases a ready-made feature matrix is useful for method comparison but weak for scientific accumulation because one cannot inspect the semantic integrity of the features. Conversely, a benchmark that publishes its extraction grammar permits others to test not only new algorithms but also alternative semantic choices.

Missingness policy deserves separate emphasis because it silently expresses assumptions about the observation process. Median imputation, indicator augmentation, complete-case restriction, model-based imputation, and learned missingness embeddings are not interchangeable tactics. Each encodes a different view of why data are absent [11]. In postoperative care, missingness is often informative. A laboratory test may be missing because the patient was clinically stable, because the team did not suspect deterioration, or because resource constraints limited ordering. A benchmark should therefore report missingness distributions by outcome, time, and institution whenever possible. It should additionally preserve raw missingness masks so that later models can treat absence as signal where appropriate. The benchmark is richer when it exposes the observation process rather than masking it behind aggressive imputation.

Evaluation protocol is the final and often most misunderstood component. A random split at the encounter level estimates one form of generalization. A temporal split estimates another. A site-held-out split estimates a third. None subsumes the others. Reproducible development should therefore regard evaluation as a vector, not a scalar. Let

$$\mathbf{m}(f; \mathcal{B}) = (m_{\text{rand}}, m_{\text{temp}}, m_{\text{site}}, m_{\text{cal}}, m_{\text{pr}}), \quad (3.2)$$

where the entries denote performance under random, temporal, and site-held-out validation together with calibration and precision-recall summaries. Reporting only the best-performing entry is analytically insufficient because it hides which dimension is driving optimism. A model that remains stable under temporal validation but fails under site transfer suggests institution-linked feature dependence. A model with respectable site-held-out ranking but poor calibration suggests prevalence or measurement shift rather than total loss of discrimination.

Reproducibility also has a computational dimension that cannot be reduced to code sharing alone. Software environments, random seeds, library versions, and data extraction dependencies all matter, especially when benchmarks are intended for longitudinal use [12]. One helpful concept is reproducibility entropy, a notional measure of how many uncontrolled degrees of freedom remain between the published analysis and a faithful rerun. If $\theta_1, \dots, \theta_p$ denote such degrees of freedom, one may write

$$H_{\text{rep}} = - \sum_{j=1}^p q_j \log q_j, \quad (3.3)$$

where q_j reflects the probability mass over plausible settings of component j left unspecified by the report. The expression is schematic, but the underlying point is concrete: methodological opacity is cumulative. A benchmark with many unspecified moving parts may be reproducible in theory and irreproducible in practice.

A mature benchmark should therefore function as both a dataset and a protocol. It should provide raw or minimally processed access where governance permits, semantic specifications for derived variables, transparent phenotype logic, multi-axis evaluation rules, and versioned changes over time. Only under those conditions can the field meaningfully compare a transparent additive model, a tree-based ensemble, a survival formulation, or a multimodal architecture without confusing algorithmic gains with hidden shifts in what is being predicted. In that sense, benchmark design is not housekeeping around the “real” science. It is the structure that determines whether the science can accumulate.

4. Transportability, Heterogeneity, and Distribution Shift

A predictive system trained at one institution exists within a locally coherent but externally fragile statistical environment. Patient demographics, procedure mix, referral intensity, preoperative optimization, ICU bed policy, postoperative testing habits, and coding conventions all influence the relationship

between recorded features and recorded outcomes. Transportability is the question of what aspects of that relationship survive when those surrounding conditions change. In surgical machine learning, this question is especially acute because the care process itself helps generate the features. A model that leans heavily on variables shaped by institutional behavior may appear strong under internal validation and still fail as soon as it encounters a new postoperative pathway.

The standard taxonomy of dataset shift remains useful here, though it requires surgical reinterpretation. Covariate shift occurs when the distribution of observed features changes across institutions while the conditional outcome relationship remains similar. Label shift occurs when prevalence changes, perhaps because one center handles more complex low pelvic cases or because surveillance intensity alters recorded incidence [13]. Concept shift arises when the mapping from features to outcome changes, which can happen when operative technique, diversion strategy, or rescue pathways differ materially across sites. Observation shift, a category especially relevant in healthcare, occurs when the same latent clinical state produces different measured data because of different monitoring practices. In leak prediction, observation shift is not a marginal nuisance. It is often the dominant mechanism of external performance loss.

A formal statement makes the issue plain. Let $P_h(X, Y)$ denote the joint distribution at institution h . Transportability of predictor f from site a to site b may be evaluated through excess risk

$$\Delta_{a \rightarrow b}(f) = \mathcal{R}_b(f) - \mathcal{R}_b(f_b^*), \quad (4.1)$$

where $\mathcal{R}_b$ is the target-site risk and f_b^* is the Bayes-optimal predictor under P_b . The challenge is that $\Delta_{a \rightarrow b}(f)$ can be large even when $\mathcal{R}_a(f)$ is small, and the size of the gap offers clues about what kind of shift is present. Severe calibration loss with modest ranking loss suggests prevalence or scale shift. Uniform ranking deterioration suggests stronger concept or observation shift. In either case, a benchmark that never leaves the source site cannot estimate transportability at all; it can only estimate internal consistency.

One promising response is hierarchical modeling of institutional heterogeneity. Instead of fitting a single global predictor and treating site effects as noise, one can allow site-indexed parameters to vary around shared population structure. Suppose $h(i)$ denotes the institution for patient i . A semipooled model may take the form

$$\begin{aligned} \eta_i &= \alpha_{h(i)} + x_i^\top \beta + x_i^\top u_{h(i)} \\ u_h &\sim \mathcal{N}(0, \Sigma_u), \end{aligned} \quad (4.2)$$

where α_h is a site intercept and u_h captures site-specific deviations from a common coefficient vector β . Such a model does not magically solve transportability, but it decomposes generalizable signal from local adjustment. Hospitals with sufficient data can express their peculiarities; hospitals with limited data shrink toward the shared structure [14]. This is conceptually preferable to training a purely local model at every site or forcing all sites into one undifferentiated parameterization.

Another route is to disentangle stable clinical representation from unstable observation artifacts. One may posit a latent variable z_i that captures the underlying postoperative state, while the observed feature vector is generated through a site-dependent measurement operator M_h . Then

$$\begin{aligned} z_i &= \Phi(u_i, \xi_i) \\ x_i &= M_{h(i)} z_i + \epsilon_i \\ P(Y_i = 1 \mid z_i) &= \sigma(\gamma^\top z_i). \end{aligned} \quad (4.3)$$

Here Φ maps baseline and procedural context u_i with latent noise ξ_i into a postoperative state representation, M_h translates that state into measured data, and γ defines leak propensity. The main implementation difficulty is that z_i is unobserved, but the formulation is still useful because it clarifies what representation learning should target. A transportable feature extractor should suppress institution-specific measurement style as much as possible while preserving information predictive of leak. In practice, this could involve temporal abstraction, normalization by site-specific reference distributions, or adversarial strategies that minimize explicit site recoverability from the learned representation.

The need for site-aware normalization is especially strong for laboratory and process variables. A postoperative white blood cell count may be measured at different times relative to surgery, in different units, and under different trigger conventions. A raw pooled model may therefore overlearn local scales. One moderate remedy is reference-space standardization, where each site's variables are mapped relative to site-specific empirical baselines before shared modeling. If x_{ij} is feature j for patient i at site h , then

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_{hj}}{\sigma_{hj} + \varepsilon}, \quad (4.4)$$

with μ_{hj} and σ_{hj} estimated on a stable reference subset. This will not eliminate observation shift, but it can reduce purely scale-driven nontransportability. More sophisticated variants could align conditional distributions within clinically matched subgroups.

Heterogeneity also enters through procedure mix. A model trained largely on one balance of right colectomy, left colectomy, low anterior resection, and urgent resections may implicitly encode risk patterns tied to that mixture. When moved to a different service profile, average performance can drift even if subgroup-wise performance remains acceptable. That is why transportability should be evaluated not only globally, but through a decomposition across clinically coherent strata. If $\mathcal{G}$ indexes such groups, overall target-site risk can be written as

$$\mathcal{R}_b(f) = \sum_{g \in \mathcal{G}} P_b(G = g) \mathcal{R}_b(f \mid G = g). \quad (4.5)$$

Observed deterioration may then be traced to altered group weights, altered within-group behavior, or both [15]. This matters operationally because the remedy differs. Weight shift may call for recalibration or threshold adjustment. Within-group degradation may indicate genuine representation failure.

A transportability program should also examine the stability of feature attributions across sites. If a model relies on physiologically sensible variables in one hospital and on process-proxy variables in another, then the predictor has not learned a stable surgical object even if AUC remains superficially adequate. One can define attribution drift between sites a and b using site-averaged explanation vectors v_a and v_b ,

$$D_{\text{attr}}(a, b) = \|v_a - v_b\|_2. \quad (4.6)$$

This diagnostic is not decisive by itself because explanations are model-dependent, but persistent large drift may signal that the model's mechanism of success is unstable. Transportability is stronger when both performance and contribution structure remain broadly consistent.

There is a temptation to treat external validation as a one-time hurdle: if a model performs acceptably at a second center, the transport problem is presumed solved. That is too weak a standard. A single external success may still reflect favorable similarity between sites. What is needed is a transport map, not a point estimate. The relevant empirical object is the distribution of performance under multiple site transfers and, ideally, under temporal evolution within sites. Only then can one begin to estimate the region of hospital-space in which the predictor is likely to remain usable.

From this perspective, multicenter collaboration is not merely a way to obtain larger sample sizes. It is the only design that reveals whether the predictor is attached to a local measurement culture or to a more durable perioperative structure. A field that continues to produce internally validated models on isolated local datasets will generate useful exploratory evidence, but it will not identify the geometry of generalization [16]. Transportability is the property that turns a promising local analysis into a candidate clinical instrument.

5. Reliability, Calibration, and Utility Under Rare-Event Burden

In rare-event prediction, the difference between a scientifically interesting model and an operationally reliable model often lies in how uncertainty is communicated and managed rather than in how cleverly the score was generated. A predictor can rank patients reasonably well and still be unsuitable for use because its probabilities are unstable, its threshold behavior is brittle, or its false-alarm burden overwhelms the clinical environment in which it operates. Anastomotic leak prediction has all of these difficulties. The event rate is low, the consequences of delayed recognition are substantial, and the costs of indiscriminate escalation are nontrivial. Reliability should therefore be treated as a multidimensional property composed of ranking, calibration, uncertainty width, and utility-sensitive operating behavior.

Ranking is usually summarized by AUC, but AUC alone is only a partial description of rare-event performance. Because the negative class is large, substantial changes in clinically relevant behavior may alter AUC only modestly. Precision-recall summaries are more informative about workload and positive predictive value, yet they remain prevalence-sensitive and threshold-free. Calibration adds a further layer because the same precision-recall profile can correspond to very different probability scales. A deployment-oriented evaluation should therefore represent performance as a vector rather than a single number. For a predictor f and calibration map g , one may define

$$\mathbf{Q}(f, g) = (\text{AUC}, \text{PRAUC}, \text{Brier}, \text{ECE}, \text{PPV}_t, \text{TPR}_t), \quad (5.1)$$

where ECE is expected calibration error and the thresholded terms are evaluated at clinically meaningful operating points t . The vector reminds us that models can exchange strengths and weaknesses across dimensions. A small gain in AUC may be irrelevant if it comes with loss of calibration or unstable threshold behavior.

The calibration problem deserves particular emphasis. A score that is monotone in risk is useful for ranking, but clinical communication requires stronger structure. If a predictor issues a probability p , clinicians implicitly treat that number as carrying absolute meaning. Calibration is the requirement that [17] $\Pr(Y = 1 \mid \hat{P} = p) \approx p$. In practice, one estimates a calibration function g such that $\hat{P} = g(s)$, where s is a raw model score. A convenient specification is a temperature-shifted logistic map,

$$\hat{P}_i = \sigma\left(\frac{s_i - \alpha}{\tau}\right), \quad (5.2)$$

with intercept α and temperature τ . More flexible isotonic or spline-based mappings can be used when sample size permits, but the central point remains: calibration is an object that should be estimated, monitored, and updated separately from the core ranking function.

Low event rates complicate probability communication because even moderately high scores can still correspond to modest absolute risks. This is not just a pedagogic detail. Threshold selection becomes highly sensitive to local capacity and the relative cost of false negatives versus false positives. Consider a threshold c that maps predicted probabilities into action. The expected utility of using c can be written as

$$U(c) = \pi \text{TPR}(c) u_{\text{TP}} + (1 - \pi) \text{TNR}(c) u_{\text{TN}} \\ - \pi \text{FNR}(c) u_{\text{FN}} - (1 - \pi) \text{FPR}(c) u_{\text{FP}}, \quad (5.3)$$

where π is prevalence and the u -terms encode local utilities or disutilities. The formula is familiar in decision theory, but its practical implication is often neglected: there is no universally correct threshold for leak prediction. A center with rapid imaging access and high tolerance for additional workup may rationally choose a different operating point from a center already strained by alert burden. Reliability is therefore inseparable from local operational context.

Because both calibration and operating burden can drift, reliability should include a measure of uncertainty around the score itself. Bootstrap, conformal, or Bayesian methods can generate intervals for the predicted probability or for the rank position. An interval-valued predictor $[\underline{p}_i, \overline{p}_i]$ is often more honest than a single point estimate when data are sparse or covariate profiles are underrepresented. One may then define a conservative decision rule

$$a_i(c) = \begin{cases} 1, & \underline{p}_i \geq c, \\ 0, & \overline{p}_i < c, \\ \emptyset, & \text{otherwise,} \end{cases} \quad (5.4)$$

where $\emptyset$ denotes an indeterminate zone. Such set-valued prediction may appear unattractive because it withholds decisiveness, yet in low-incidence, high-consequence problems it can be preferable to false precision [18]. An indeterminate region flags cases where the model's own uncertainty is too large for confident categorization, preserving room for clinical judgment or additional data acquisition.

Reliability also has a subgroup dimension. A score can be well behaved overall and still exhibit poor calibration or unstable precision in clinically important strata. Since postoperative services differ by urgency, procedure type, and patient reserve, subgroup reliability should be part of routine assessment rather than an optional fairness appendix. Let G denote subgroup membership. One may define calibration disparity as

$$\Delta_{\text{cal}} = \sup_{g \in G} |\mathbb{E}[Y \mid \hat{P}, G = g] - \hat{P}|, \quad (5.5)$$

interpreted as the maximal conditional miscalibration across groups. Empirically this quantity must be estimated with smoothing or shrinkage because event counts are limited, but the conceptual point stands. Reliable surgical prediction is not merely globally reliable; it should avoid concentrated unreliability in subpopulations that differ in case complexity or surveillance density.

A related concern is threshold instability under prevalence change. Precision at a fixed threshold is strongly prevalence-dependent. A score calibrated in a tertiary center with more complex cases may appear overcautious in a community setting even if its likelihood ratios remain useful. This is another reason to separate ranking from calibration and both from policy thresholds. A transported model may require recalibration and threshold retuning while preserving its internal ordering. Treating any performance change as evidence that the entire model must be retrained confuses these layers and can lead to unnecessary redevelopment.

Finally, reliability under rare-event burden should be judged longitudinally. A model that looks acceptable in an annual aggregate may still oscillate month to month because the number of observed leak events is small. Rolling monitoring should therefore use stabilized estimators and prespecified alert-rate guardrails [19]. One can model the observed number of threshold crossings A_t over time and compare it to an expected range under the benchmark distribution. Likewise, rolling calibration slope

and intercept can be tracked with Bayesian updating to avoid overreacting to transient noise. The broader principle is that rare-event reliability is not established once; it is maintained through repeated reconciliation between the score's mathematical behavior and the environment in which the score is being used.

6. Prospective Evaluation as an Intervention Problem

Retrospective validation answers a limited question: given a fixed historical record, how well can a model anticipate a later-coded outcome? The moment a predictor is made visible in clinical practice, the question changes. The predictor becomes part of the care process. Clinicians may order imaging earlier, repeat laboratories more frequently, admit borderline patients to higher levels of care, or document concern with greater specificity. These actions can alter both the time of detection and the empirical incidence of recorded leak. Prospective evaluation must therefore be framed as the study of an intervention on the information and response system, not as a mere extension of retrospective scoring.

This point is often missed because prediction and intervention are conceptually separated in methodological writing. In the hospital, however, the separation is porous. A visible score changes what clinicians know, what they attend to, and what they do next. Let A_i denote exposure to the live predictor and D_i the downstream decisions that become available or more likely because of that exposure. Outcomes Y_i then depend not only on baseline features X_i but also on decision-mediated pathways. A simplified structural system is

$$\begin{aligned} D_i &= \psi(X_i, A_i, \nu_i) \\ Y_i &= \varphi(X_i, D_i, \epsilon_i), \end{aligned} \tag{6.1}$$

where ν_i and ϵ_i are unobserved factors. Under this representation, evaluating the live predictor is closer to policy evaluation than to standard predictive validation. The relevant quantity is not simply $P(Y | X)$, but the effect of allowing $A_i = 1$ within a specified response protocol.

Prospective work should therefore begin with silent deployment [20]. In a silent phase, the predictor runs on current data, logs its outputs, but is not shown to clinicians. This phase estimates real-time data fidelity, latency, threshold crossing frequency, subgroup coverage, and concordance with observed outcomes without yet perturbing care behavior. Silent deployment often reveals discrepancies between retrospective assumptions and production reality: delayed data availability, source-table instability, missing timestamps, or greater than expected frequency of edge cases. Importantly, it also provides a baseline against which the later behavioral effect of visibility can be measured.

Once the predictor becomes visible, evaluation should distinguish process effects from clinical effects. Process outcomes include changes in review timing, diagnostic ordering, escalation patterns, and care-location transitions. Clinical outcomes include time to confirmed detection, severity at detection, rescue success, unplanned reoperation, prolonged stay, or mortality. These layers should not be collapsed. A predictor may improve process regularity without yet changing hard outcomes, especially when the intervention acts on recognition rather than on the biological formation of leak. Conversely, an apparent outcome change without a corresponding process pathway may reflect confounding or coding drift rather than a genuine predictive benefit.

A useful design for staggered implementation is a panel or stepped-introduction framework. Let Y_{it} denote an outcome for service unit or patient i at time t , and let A_{it} denote exposure to the visible predictor. A simple model is

$$Y_{it} = \mu_i + \lambda_t + \theta A_{it} + \beta^\top X_{it} + \varepsilon_{it}, \tag{6.2}$$

with unit fixed effects μ_i and time effects λ_t . The coefficient θ captures the average effect of exposure under the assumptions of the design. Depending on the outcome, this can be embedded in logistic, count, or survival models. What matters is not the exact regression family but the recognition that implementation unfolds over time and across service structures, making repeated-measures designs more realistic than a single before-after comparison.

Counterfactual policy value provides another lens [21]. Suppose the predictor supports a policy π that maps observed patient history to one of several surveillance states. The value of the policy can be written as

$$V(\pi) = \mathbb{E}^{\pi} \left[\sum_{t=0}^T r_t(S_t, A_t) \right], \quad (6.3)$$

where S_t is the patient state, A_t the chosen surveillance or diagnostic action, and r_t the reward or cost at time t . Inverse probability weighting or doubly robust estimation can be used to estimate $V(\pi)$ from observational data if a suitable action model is available. While such methods are demanding, they clarify a crucial point: the value of a predictor is mediated by the policy it induces. A highly discriminative score with an ill-chosen response policy may have lower clinical value than a modest predictor embedded in a coherent escalation protocol.

Prospective evaluation must also account for feedback in outcome ascertainment. If a visible score causes more imaging, some occult leaks will be detected earlier or more frequently than before. An increase in recorded leak diagnoses after deployment may therefore reflect improved ascertainment rather than worse care. This is not a statistical inconvenience to be ignored; it is part of the intervention effect. The evaluation plan should predefine how detection timing, severity at diagnosis, and downstream rescue markers will be interpreted so that changes in incidence are not misread. Composite outcomes that include severe deterioration, urgent reintervention, or sepsis progression may sometimes be more stable than leak coding alone.

Randomized evaluation remains valuable but should be designed around realistic units of intervention. Patient-level randomization can be difficult when clinicians work on teams and share awareness of the tool. Unit-level or time-block randomization may better reflect how exposure actually occurs. Hybrid designs are also possible, with silent deployment everywhere and visible activation randomized across teams. The appropriate design depends on the maturity of the predictor and the degree of operational standardization already achieved. The broader point is that outcome evaluation should be aligned with the intervention mechanism, not imported from abstract trial design without regard to how surgical services function [22].

Finally, prospective evaluation should include clinician behavior as data rather than noise. Override rates, reasons for nonadherence, qualitative disagreement patterns, and service-specific usage trajectories are analytically valuable. They reveal whether the predictor's failures are due to poor representation, poor communication, mismatched action recommendations, or reasonable clinical context not captured in the model. In many predictive deployments, structured disagreement is the most informative post-launch signal because it identifies the boundary between algorithmic abstraction and bedside reality. A system that cannot learn from its own contested cases is unlikely to mature.

The decisive implication is that the passage from proof-of-concept to clinical use cannot be understood as merely "turning on" a model. It is the initiation of a dynamic experiment in which information, behavior, and outcome interact. A prospective design that respects this interaction is essential if one wants to know whether a predictor has become clinically meaningful rather than merely visible.

7. Lifecycle Governance for Surgical Prediction Infrastructures

A live surgical prediction system is not a finished artifact. It is a maintained infrastructure whose statistical meaning can erode, improve, or mutate as the hospital changes around it. New perioperative

pathways are introduced, staffing patterns shift, laboratory panels change, documentation interfaces are redesigned, and surgical populations drift over time. Governance is the discipline that keeps the model legible and revisable under those ordinary forms of movement. Without governance, performance monitoring becomes episodic, threshold changes become informal, and responsibility diffuses across technical and clinical teams. With governance, the institution can decide which kinds of change require recalibration, which require retraining, and which require retirement of the predictor.

The first principle of governance is versioned transparency. Every deployed predictor should have a documented identity consisting of benchmark version, model version, calibration version, threshold policy, and monitoring window. These layers should not be conflated because they change for different reasons [23]. A recalibration update does not imply a new representation model. A threshold change to reduce workload does not mean that the model's ranking has improved. A benchmark revision reflecting phenotype refinement means that past and present scores are no longer directly comparable without qualification. A compact notation is

$$\mathcal{S}_t = (\mathcal{B}^{(k)}, f^{(\ell)}, g^{(m)}, \pi^{(r)}), \quad (7.1)$$

where $\mathcal{B}^{(k)}$ is benchmark version k , $f^{(\ell)}$ the predictive model, $g^{(m)}$ the calibration map, and $\pi^{(r)}$ the action policy. Logging $\mathcal{S}_t$ with each score ensures that later audits can reconstruct exactly what the system meant at the moment of use.

The second principle is prespecified drift response. Distributional movement should be expected, not treated as evidence of failure in itself. Governance should therefore distinguish surveillance metrics from intervention triggers. A rise in missingness for preoperative albumin, for example, may warrant investigation without immediate model suspension. Sustained deterioration in calibration slope, unexplained growth in threshold-crossing frequency, or concentrated subgroup miscalibration may justify recalibration or policy revision. More severe signals, such as loss of site-to-site ranking stability or systematic conflicts with adjudicated outcomes, may justify retraining or withdrawal. The point is not to automate these decisions entirely, but to ensure that they are made through recognized criteria rather than ad hoc reaction.

The third principle is joint clinical and technical stewardship. Prediction systems in surgery fail when either side governs alone. Technical teams may focus on model metrics without appreciating care-pathway implications. Clinical teams may respond to salient anecdotes without recognizing statistical uncertainty or base-rate effects. An effective governance body needs both kinds of competence and a structured review cadence. Cases of false reassurance, high-consequence false alarm, and recurrent override should be examined not to assign blame, but to identify whether the issue lies in label construction, feature semantics, calibration drift, threshold policy, or context unavailable to the model. This is why comprehensive logging is indispensable [24]. Without it, retrospective judgment collapses into memory and impression.

Equity should also be operationalized within governance rather than left as a one-time fairness statement. In surgical prediction, inequity can arise through multiple channels: different baseline risk distributions, different rates of missing data, different surveillance intensity, different precision at the same threshold, and different access to downstream action once flagged. Governance should therefore review not only subgroup discrimination and calibration but also alert exposure, override patterns, and realized care consequences. If one subgroup receives more indeterminate outputs because its data are sparser, that is part of the model's practical behavior even if aggregate AUC remains stable. Institutions should decide in advance what level of subgroup disparity triggers remedial analysis or policy adjustment.

Human interpretability belongs here as well. A system can be mathematically stable and still lose legitimacy if clinicians no longer understand its role. Governance should thus monitor not only quantitative performance but also the conceptual alignment between the score and clinical reasoning. Are the main contributors still recognizable? Do changes in local postoperative care make some variables less

meaningful than before? Has the action recommendation drifted away from what teams regard as proportionate? These questions are not peripheral to the model's function. A prediction system used in a skeptical environment must continually justify its own intelligibility.

Maintenance strategy should be tiered. Minor changes include recalibration against new prevalence and scale, or modest threshold adjustment to maintain workload targets. Intermediate changes include feature remapping after source-system alteration or subgroup-specific recalibration where justified. Major changes include retraining on expanded multicenter data, revised phenotype logic, or redesigned temporal representation. Each tier should have its own validation requirements. One should not require a full prospective relaunch for every threshold change, but neither should a major phenotype revision proceed under the guise of routine maintenance [25]. Governance supplies this taxonomy so that adaptation remains orderly.

An underappreciated governance issue is retirement. Predictive systems accumulate institutional inertia. Once embedded in dashboards and review routines, they may persist despite declining usefulness because no one owns the decision to remove them. Retirement criteria should therefore be specified at deployment. Persistent inability to maintain calibration, failure to retain acceptable utility after reasonable adjustment, or replacement by a demonstrably superior benchmark-model stack are all legitimate grounds for retirement. Removing a predictor that no longer has defensible meaning is evidence of mature governance, not of project failure.

Lifecycle governance becomes especially important once multicenter benchmarking is achieved. A model may need local calibration at each hospital, but governance should also preserve comparability across sites so that learning remains cumulative. Shared audit templates, common phenotype updates, and federated monitoring summaries can help. The broader aim is to create prediction infrastructures that are locally governable while remaining scientifically connected. A field advances when hospitals do not merely deploy their own black boxes, but participate in a common discipline of versioning, validation, and revision.

What emerges from this governance perspective is a shift in how one should think about translation. The central challenge is not only whether a model can be developed, but whether a predictive service can remain interpretable and controllable over time. Hospitals already manage medications, devices, and quality metrics through lifecycle processes. Surgical machine learning deserves the same seriousness. Without it, proof-of-concept systems will continue to generate scattered enthusiasm without becoming dependable parts of perioperative intelligence.

8. Conclusion

The route from proof-of-concept surgical prediction to durable clinical intelligence is narrower and more technical than retrospective performance tables sometimes suggest [26]. The core challenge is not merely to extract signal from perioperative records. It is to specify a benchmark whose semantic structure survives scrutiny, to understand how institutional measurement practices shape apparent performance, to separate ranking from probability assignment under rare-event conditions, and to evaluate the live system as an intervention that changes the very data environment in which it operates. These requirements move the discussion beyond isolated model comparison and toward a fuller account of reproducibility, transportability, and lifecycle control.

A neutral reading of the present state of the field supports that shift. Moderate predictive signal appears to exist in routinely collected perioperative data, and this is enough to justify continued work. It is not enough to justify casual translation. The crucial unresolved questions concern stability across sites, clarity of the target phenotype, calibration under changing prevalence and observation intensity, and the governance structures needed to maintain statistical meaning after deployment. In other words, the field does not need less modeling, but it does need a different ratio between model invention and benchmark discipline.

The framework developed here treats leak prediction as a layered object. At one layer lies latent post-operative failure. At another lies the observation system that renders that failure partially visible. Above

those lie institutional heterogeneity, utility-sensitive action rules, and prospective behavioral feedback. A predictor becomes clinically credible only when its development and evaluation acknowledge all of these layers. That is why reproducible benchmarks, multicenter transport studies, calibrated probability reporting, policy-aware evaluation, and explicit governance belong to the core scientific agenda rather than to later implementation work.

References

- [1] A. R. Khawaja, M. A. Mir, A. Beg, S. M. Baba, M. S. Wani, and A. Hamid, "Outcome of extra mucosal single interrupted layer (emsil) ileo-ileal anastomosis in urinary diversions during open radical cystectomy," *African Journal of Urology*, vol. 29, 5 2023.
- [2] G. M. Jung, S. H. Lee, D. S. Myung, W. S. Lee, Y. E. Joo, M. R. Jung, S. Y. Ryu, Y.-K. Park, and S. B. Cho, "Novel endoscopic stent for anastomotic leaks after total gastrectomy using an anchoring thread and fully covering thick membrane: Prevention of embedding and migration," *Journal of gastric cancer*, vol. 18, pp. 37–47, 2 2018.
- [3] K. E. Rollins, H. Javanmard-Emamghissi, M. J. Scott, and D. N. Lobo, "The impact of peri-operative intravenous lidocaine on postoperative outcome after elective colorectal surgery: A meta-analysis of randomised controlled trials.," *European journal of anaesthesiology*, vol. 37, pp. 659–670, 2 2020.
- [4] H. Impellizzeri, A. Pulvirenti, M. Inama, M. Bacchion, E. Marrano, M. Creciun, A. Casaril, and G. Moretto, "Near-infrared fluorescence angiography for colorectal surgery is associated with a reduction of anastomotic leak rate.," *Updates in surgery*, vol. 72, pp. 991–998, 4 2020.
- [5] M. B. Witte, J. Saupe, J. Reiner, K. Bannert, C. Schafmayer, G. Lamprecht, and P. Berlin, "Ileocolonic healing after small ileocecal resection in mice: Nod2 deficiency impairs anastomotic healing by local mechanisms.," *Journal of clinical medicine*, vol. 12, pp. 3601–3601, 5 2023.
- [6] T. Hinoi, M. Okajima, M. Shimomura, H. Egi, H. Ohdan, F. Konishi, K. Sugihara, and M. Watanabe, "Effect of left colonic artery preservation on anastomotic leakage in laparoscopic anterior resection for middle and low rectal cancer," *World journal of surgery*, vol. 37, pp. 2935–2943, 9 2013.
- [7] M. A. Klose, J. Walldorf, M. Damm, S. Krug, J. Klose, U. Ronellenfitch, J. Kleeff, P. Michl, and J. Rosendahl, "Treatment of esophageal leakages with the microtech®-vac-stent: a monocentric early experience of three cases.," *Therapeutic advances in gastrointestinal endoscopy*, vol. 16, pp. 26317745231200312–26317745231200312, 9 2023.
- [8] B. Sadanandan, "Data-driven risk stratification for anastomotic leak: A proof-of-concept study using electronic health records," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 6, pp. 1–27, 2024.
- [9] J. Ma, H. Hu, J. Xiao, Y. Zhao, X. Chen, S. Wei, and H. Zhang, "Esophageal covered stent treatment for gastroesophageal cervical anastomotic fistula," *Chinese journal of radiology*, vol. 53, pp. 385–388, 5 2019.
- [10] T. Adanir, O. Nazli, C. Kara, M. Aksun, A. Sözütek, A. Şencan, and G. Oktay, "The relationship between vasopressor dose and anastomotic leak in colon surgery: An experimental trial," *International journal of surgery (London, England)*, vol. 8, pp. 221–224, 1 2010.
- [11] A. Ghuman, A. A. Karimuddin, C. J. Brown, M. J. Raval, and P. T. Phang, "Colorectal surgery surgical site infection prevention practices in british columbia.," *Canadian journal of surgery. Journal canadien de chirurgie*, vol. 64, pp. E516–E520, 10 2021.
- [12] R. M. Engel, K. Oliva, S. Centauri, W. Wang, P. J. McMurrick, and R. Yap, "Impact of anastomotic leak on long-term oncological outcomes after restorative surgery for rectal cancer: A retrospective cohort study.," *Diseases of the colon and rectum*, vol. 66, pp. 923–933, 11 2022.
- [13] D. Borowski, S. Kelly, A. Ratcliffe, D. Bradburn, R. Wilson, A. Gunn, S. Pillinger, A. Sharma, A. Taylor, T. White, and J. Monson, "Six of the best, colorectal 29–34," *British Journal of Surgery*, vol. 91, pp. 73–75, 6 2004.
- [14] 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.
- [15] Y. W. Min, T. Kim, H. Lee, B.-H. Min, H. K. Kim, Y. S. Choi, J. H. Lee, P.-L. Rhee, J. J. Kim, J. I. Zo, and Y. M. Shim, "Endoscopic vacuum therapy for postoperative esophageal leak.," *BMC surgery*, vol. 19, pp. 37–37, 4 2019.

- [16] V. A. Fleetwood, K. Gross, G. Alex, C. S. Cortina, J. Smolevitz, S. Sarvepalli, S. Bakhsh, J. Poirier, J. Myers, M. Singer, and B. A. Orkin, "Common side closure type, but not stapler brand or oversewing, influences side-to-side anastomotic leak rates.," *American journal of surgery*, vol. 213, pp. 590–595, 1 2017.
- [17] M. A. Mf, L. Lf, K. Ja, and U. R, "Dual pathology: Synchronous existence of gastrointestinal stromal tumour (gist) of stomach and adenocarcinoma of colon," *Scholars Journal of Medical Case Reports*, vol. 8, pp. 945–947, 11 2020.
- [18] Y. Kozlov, S. Poloyan, A. Marchuk, A. Rozhanski, A. Byrgazov, K. Kovalkov, C. Ochirov, V. Kapuller, and A. Narkevich, "Systematic bibliographical review and meta-analysis of thoracoscopic vs. open-surgical treatment of esophageal atresia," *Pediatrics. Journal named after G.N. Speransky*, vol. 102, pp. 112–124, 8 2023.
- [19] M. A. Ziegler, J. A. Catto, T. W. Riggs, E. R. Gates, M. B. Grodsky, and H. Wasvary, "Risk factors for anastomotic leak and mortality in diabetic patients undergoing colectomy: analysis from a statewide surgical quality collaborative.," *Archives of surgery (Chicago, Ill. : 1960)*, vol. 147, pp. 600–605, 7 2012.
- [20] L. M. Zhang, D. Schuitevoerder, M. G. White, S. Feldt, P. Krishnan, N. Hyman, and B. D. Shogan, "Combined mechanical and oral antibiotic bowel preparation is associated with prolonged recurrence-free survival following surgery for colorectal cancer.," *Journal of surgical oncology*, vol. 124, pp. 1106–1114, 7 2021.
- [21] A. Schiff, B. L. Brady, S. Ghosh, S. Roy, C. Ruetsch, and E. J. Fegelman, "Intra-operative anastomotic leak rates and testing methodology in colorectal resection surgery," *Journal of Surgery and Surgical Research*, vol. 2, pp. 048–054, 7 2016.
- [22] S. P. Fortin, W. Petraiuolo, G. Cafri, G. Scapini, P. Agarwal, D. Chakke, S. Johnston, B. H. Johnson, P. M. Coplan, and S. Zhang, "Comparison of clinical outcomes of gripping surface technology staple reloads versus standard staple reloads used with manual linear surgical staplers.," *Medical devices (Auckland, N.Z.)*, vol. 15, pp. 385–399, 12 2022.
- [23] T. Hossain, A. Jackson, F. Youssef, M. Trotter, F. Yanni, D. Zanotti, J. Catton, N. Welch, R. Vohra, and S. L. Parsons, "Ogc p25 outcomes of laparoscopic vs open oesophagectomy in a single high volume centre," *British Journal of Surgery*, vol. 109, 12 2022.
- [24] S. Joglekar, A. Huynh, W. Zhou, L. Chong, A. Sayed-Hassen, M. Hii, and R. Cade, "398. the hybrid intra-thoracic oesophago-gastric anastomosis—contemporary performance with low benign anastomotic stricture formation," *Diseases of the Esophagus*, vol. 36, 8 2023.
- [25] M. C. Haksal, N. Okkabaz, A. E. Atici, O. Civil, Y. Ozdenkaya, A. Erdemir, N. Aksakal, and M. Oncel, "Fortune of temporary ileostomies in patients treated with laparoscopic low anterior resection for rectal cancer," *Annals of surgical treatment and research*, vol. 92, pp. 35–41, 12 2016.
- [26] 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.