

Original Research

Empirical Stress-Testing of Early Anastomotic Leak Prediction Under Label Noise, Selection Bias, and Threshold Dependence

Dilan Perera¹, Sanjaya Fernando² and Ruwan Jayawardena³

¹Department of Computer Science, University of Kelaniya, Kandy Road, Dalugama, Kelaniya 11600, Sri Lanka.

²Department of Public Health, Rajarata University of Sri Lanka, Mihintale 50300, Sri Lanka.

³Department of Information Technology, University of Moratuwa, Katubedda, Moratuwa 10400, Sri Lanka.

Abstract

Postoperative prediction models are often judged by a narrow set of performance numbers, yet those numbers are produced inside a data system with limited event counts, incomplete measurements, and outcomes that are not directly observed but inferred from coded and procedural traces. Anastomotic leak prediction is a clear example. The event is uncommon, the first postoperative day contains only partial evidence, and the label used for training usually combines diagnosis codes with downstream clinical actions. Under these conditions, the practical meaning of a model depends not only on ranking ability, but also on how sensitive the reported results are to outcome definition, threshold placement, and the hospital's own pattern of observation. This paper presents an empirical re-analysis of an early leak prediction experiment using structured perioperative records. The analysis reconstructs model operating points from the published held-out metrics and then subjects them to three stress tests. The first examines how modest phenotype misclassification changes estimated prevalence, precision, and recall. The second studies the effect of observation-dependent variables such as intensive care placement and selectively measured laboratories on the stability of feature importance and thresholded decisions. The third examines how prevalence shifts across hospitals alter the positive predictive value of the same score, even when ranking quality is unchanged. Across these analyses, the main result is that the stronger nonlinear models remain informative, but their reported performance should be read as conditional on a specific label construction and a specific monitoring environment. The study therefore argues for empirical reporting practices that make those conditions explicit rather than presenting thresholded performance as if it were transportable without adjustment.

1. Introduction

Prediction work in surgery often moves quickly from technical feasibility to claims about usefulness [1]. That move is understandable. The complications being studied are clinically serious, and a model that appears to identify elevated risk early enough to matter will naturally draw attention. The difficulty is that the path from statistical promise to dependable interpretation is narrower than the usual summary table suggests. In low-frequency postoperative outcomes, the model is built on a record that is already filtered through how clinicians observed the patient, when they ordered tests, when they escalated care, and how they later encoded the outcome. What looks like a single classification problem is therefore a mixture of at least three distinct problems. One problem concerns whether the event can be separated from non-event cases using features available early in the hospital course. A second concerns whether the event definition itself is stable enough that the model is learning the complication rather than the documentation pattern around the complication. A third concerns whether the threshold that looks reasonable inside one held-out set would preserve the same meaning in a setting with different prevalence, different measurement frequency, or different care intensity. These are all empirical questions. They are

not philosophical objections to modeling. They are conditions that determine what the reported numbers actually mean.

Anastomotic leak after colorectal surgery is especially suited to this kind of scrutiny. The event usually emerges several days after the operation, but the first postoperative day already contains physiologic changes, laboratory values, and care-location transitions that may carry information about who will later deteriorate. This makes the problem analytically attractive. It also makes it vulnerable to overinterpretation. A model using only early structured data does not observe the leak directly. It observes partial signs of tissue stress, inflammatory activation, perioperative instability, and the hospital response to those states [2]. If the label used for supervised learning is itself a composite of diagnostic coding and downstream interventions, then both sides of the learning problem are already shaped by institutional process. The model can still be useful under those conditions [3]. It simply cannot be interpreted as though it were mapping cleanly from patient biology to a noise-free endpoint [4].

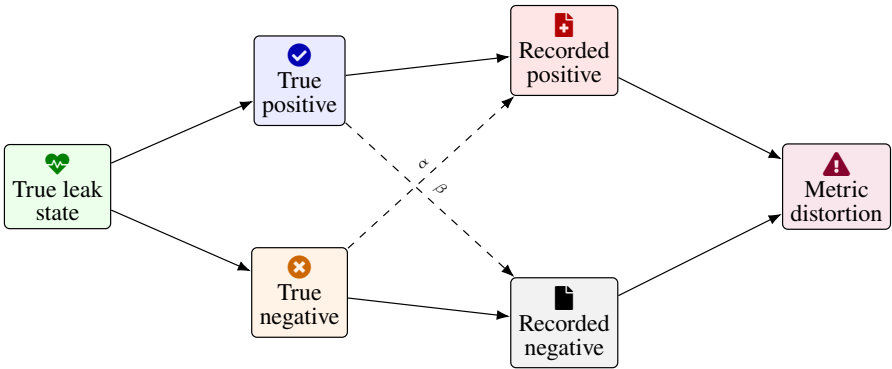


Figure 1: Outcome perturbation view of the composite leak phenotype. A latent clinical complication is converted into recorded labels through asymmetric false-negative and false-positive pathways, so reported precision, recall, and calibration depend not only on discrimination but also on phenotype construction and coding-supported corroboration.

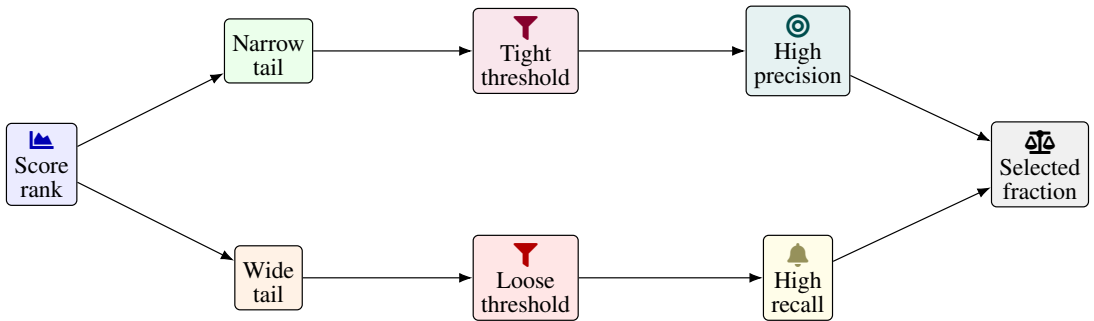


Figure 2: Threshold dependence of rare-event surgical prediction. Concentrated nonlinear score tails support smaller selected high-risk sets with stronger precision, whereas diffuse score distributions expand the positive set and raise recall at the cost of many more flagged admissions. The selected fraction becomes an important operational quantity alongside AUC.

The usual performance summaries are not well designed to expose that issue. AUC is threshold-free, which is useful, but it does not show how much the selected positive set depends on a narrow group of cases near a cutpoint. Precision and recall show that trade-off at one threshold, but without a direct account of how the outcome was defined or how a hospital’s observation process enters the feature

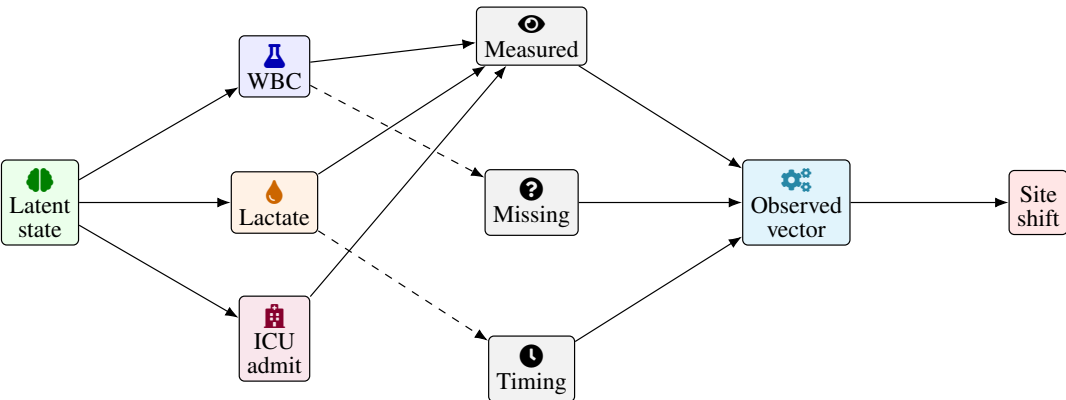


Figure 3: Observation dependence in the first postoperative day. The model does not consume latent physiology directly; it receives an observed vector shaped by whether tests were drawn, when they were drawn, and how intensively the patient was monitored. This makes feature importance partly a property of hospital process as well as clinical state.

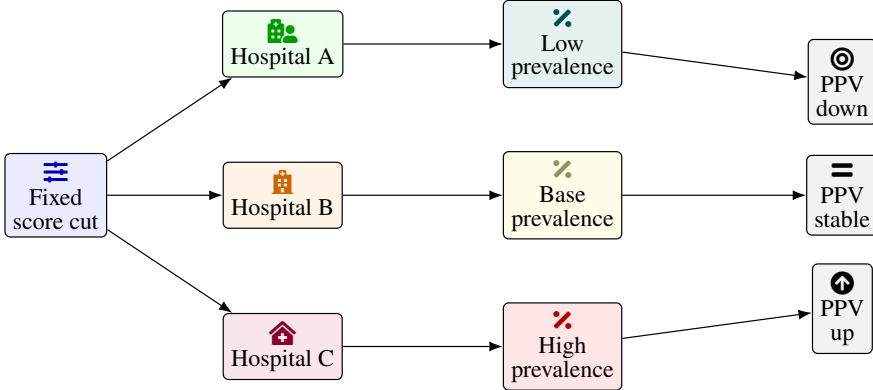


Figure 4: Prevalence transport under a fixed operating threshold. Even when score ranking remains useful, the same alert threshold changes clinical meaning across hospitals because positive predictive value is prevalence-sensitive. A threshold chosen in one institution may become too permissive or too conservative elsewhere.

space, those quantities can look more stable than they are. Calibration metrics are helpful, but even they remain conditional on the event definition that generated the observed labels. If a proportion of true leaks are not captured by the coding logic, then a model can be penalized for correctly ranking patients against an imperfect outcome. If false positives in the recorded label arise from broad-spectrum antibiotic administration or procedural coding that does not always indicate a true leak, then measured precision and Brier error can shift for reasons that have little to do with the underlying biology.

A second complication is the role of measurement availability. Early postoperative structured data are not collected uniformly. Some variables, such as white blood cell count or creatinine, are part of common perioperative panels and therefore available for most admissions. Others, such as preoperative albumin or outpatient body mass index, depend on how the patient entered care, how much time existed for preoperative workup, and whether the record captured prior documentation cleanly. Still others, such as ICU admission, are not physiologic variables in the narrow sense at all. They summarize an institutional judgment about acuity or anticipated monitoring needs. When such variables rank highly, the model is learning from the way the hospital sees the patient as much as from the patient alone. That is not necessarily a defect. A postoperative surveillance tool may reasonably use this information. But it

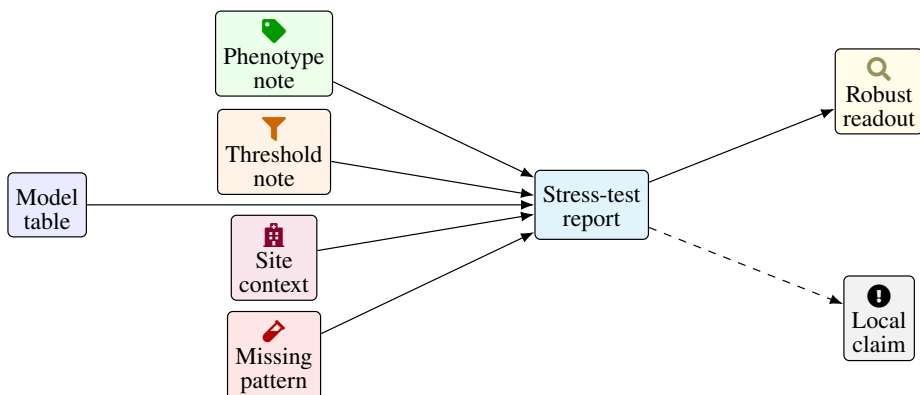


Figure 5: Recommended reporting structure for rare-event postoperative prediction. Instead of presenting thresholded metrics as universally portable, the report should attach them to explicit phenotype assumptions, threshold choice, local observation process, and missingness structure so readers can distinguish robust signal from setting-dependent performance.

means that transporting the model from one monitoring environment to another is not a simple matter of copying coefficients or trees.

A third complication is that threshold choice is almost always local. A threshold selected to optimize F_1 , or any related criterion, is a property of a specific development process [5]. It is not a universal constant. The balance between false negatives and false positives that looks acceptable in one retrospective dataset may not map well to another hospital with a different prevalence of leak or a different tolerance for additional imaging and workup. If prevalence falls, precision falls unless the false-positive rate falls as well. If case mix shifts toward more complex surgery, the opposite can happen. A threshold that looks acceptable in one environment can therefore become either too permissive or too conservative without any change in the model’s ranking behavior.

The aim of this study is to make those issues explicit through a secondary empirical analysis of a published proof-of-concept experiment. The focus here is not on proposing another model. The focus is on stress-testing what the published results imply when the problem is viewed through outcome perturbation, observation dependence, and threshold transport rather than through point performance alone. This approach is useful because the published study already contains enough information to support a more detailed reading. The cohort size is known. The event count is known. The outcome construction is stated. The early feature window is stated. The held-out metrics for six model configurations are reported. The leading predictors are named. The data-quality and generalizability limitations are described. That information is sufficient to reconstruct the approximate operating behavior of the models and to examine how that behavior changes under plausible perturbations.

The paper therefore asks three questions. First, how vulnerable are the reported precision and recall values to modest outcome misclassification, given that the endpoint is a composite built from diagnosis codes and corroborating actions rather than individual chart review? Second, what does the feature ranking imply about the dependence of the model on observation intensity and care-process variables, and how might that dependence affect interpretation of feature importance and model transport? Third, how much does the meaning of a thresholded positive prediction change when prevalence shifts away from the original development setting? These questions are empirical in the strict sense [6]. They are answered by deriving consequences from the reported experiment, not by importing speculative claims from outside the dataset.

The resulting argument is limited but useful. It does not claim that the original model comparison was mistaken. It does not claim that the stronger nonlinear configurations are artifacts. It shows something narrower. The reported results are best read as conditional measurements of signal under a specific

phenotype, a specific 24-hour observation window, and a specific critical-care-heavy hospital environment. Once those conditions are perturbed, the magnitude and meaning of thresholded performance can shift materially. That is an important empirical result because it changes what should count as sufficient reporting in rare-event surgical prediction. A model table is not wrong. It is incomplete unless it also tells the reader how dependent those numbers are on the way the endpoint was defined, the way the data were observed, and the way the threshold was chosen.

2. Empirical Setup

The MIMIC-IV colorectal surgery experiment described by Sadanandan (2024), which assembled 3,331 admissions from 2,847 unique patients, defined anastomotic leak by a composite of diagnostic coding and supporting clinical evidence within 30 days, restricted predictors to data available within the first 24 hours after surgery, handled missing continuous values by median imputation with missingness indicators, assigned dedicated unknown categories to missing categorical values, split data at the patient level, and evaluated six model configurations on a held-out test set of 670 admissions containing 49 positive cases [7]. The highest-ranked variables in the gradient boosting analysis were postoperative white blood cell count, ICU admission status, postoperative lactate, surgery type, and preoperative albumin. Those elements are enough to support a structured empirical re-analysis.

Two features of the source design matter immediately. First, the positive class is rare. A held-out set with 49 positive admissions means that any thresholded quantity involving the event class is numerically coarse. Each additional true positive contributes slightly more than two percentage points of recall. Second, the outcome is not derived from direct chart adjudication [8]. It is a pragmatic composite that uses diagnosis codes together with downstream evidence such as reoperation, drainage, or broad-spectrum antibiotic therapy. That is a reasonable engineering choice for a proof-of-concept study. It also means that the observed label is not guaranteed to match the underlying clinical truth without error. The rest of this paper treats that imperfection not as a flaw that invalidates the study, but as an empirical property that should be modeled explicitly.

A useful starting point is the operating geometry implied by the published precision and recall values. If n_1 is the number of positive cases, n_0 the number of negatives, p_m the precision of model m , and r_m its recall, then approximate confusion counts follow from the identities

$$\begin{aligned} TP_m &= r_m n_1 \\ FN_m &= n_1 - TP_m \\ FP_m &= TP_m \left(\frac{1 - p_m}{p_m} \right) \\ TN_m &= n_0 - FP_m. \end{aligned} \tag{2.1}$$

With $n_1 = 49$ and $n_0 = 621$, these formulas recover the thresholded structure of the six models. Logistic regression corresponds to roughly 27.9 true positives and 79.5 false positives. Naive Bayes corresponds to about 34.8 true positives and 255.1 false positives. Gradient boosting corresponds to about 25.0 true positives and 53.1 false positives. The target-encoded variant corresponds to 21.1 true positives and 40.9 false positives. Bagged gradient boosting corresponds to 20.1 true positives and 26.6 false positives. The ensemble corresponds to 24.0 true positives and 34.6 false positives. Because the reported metrics were rounded, these counts are approximate. The approximation error is small relative to the differences being studied.

The same reconstruction shows how much of the test set each model places above threshold. The selected fraction α_m satisfies

$$\alpha_m = \frac{TP_m \cdot FP_m}{n}. \quad (2.2)$$

The resulting values are about 0.160 for logistic regression, 0.433 for naive Bayes, 0.117 for gradient boosting, 0.093 for gradient boosting with target encoding, 0.070 for bagged gradient boosting, and 0.087 for the ensemble. This difference in selected fraction is one of the most important empirical facts in the experiment [9]. The better nonlinear models are not merely reordering the cohort in a general sense. They are identifying relatively narrow subsets of admissions as high risk, whereas logistic regression and especially naive Bayes mark much larger fractions of the cohort positive.

The source study also reported Brier score, which adds another key distinction. Logistic regression and the four boosting-based configurations all reported Brier score 0.059. Naive Bayes reported 0.415. With outcome prevalence around 7.6%, a constant predictor that always outputs the prevalence would produce a Brier score close to $\pi(1 - \pi)$, approximately 0.070. This means the logistic and boosting-based models are at least numerically better than a prevalence-only baseline, whereas naive Bayes produces probability outputs that are much worse. That discrepancy becomes important later because it suggests that the independence assumptions of naive Bayes distort probability scale even when the model still captures some ranking signal.

The descriptive outcome data in the source study matter as well. Leak cases had much longer hospital stays, more frequent ICU admission in the first 24 hours, and higher in-hospital mortality than non-leak cases. These contrasts imply that the observed positive class corresponds to a clinically heavy subgroup rather than to a mild or incidental endpoint. They also help explain why variables such as ICU admission rank so highly. A patient's location and early physiologic burden are not peripheral to the task. They are part of the signal surface the model uses to separate future leak cases from the remainder of the cohort.

The feature-engineering choices also shape interpretation. The first postoperative day included both preoperative and postoperative laboratory values where available, together with demographics, admission characteristics, procedure details, and outpatient-derived baseline measurements such as body mass index. Continuous features were standardized for logistic regression and naive Bayes but left on the raw scale for the tree-based models. Missing continuous values were imputed with the training-set median and accompanied by indicator variables. This means the models were allowed to learn from absence as well as presence [10]. In a hospital record, that is often appropriate. It also means that the model's behavior depends partly on the process by which measurements are obtained.

That observation is central to the stress-test developed here. A model using postoperative white blood cell count, lactate, ICU admission, and albumin is learning from a mixture of acute physiology, procedure context, baseline reserve, and care-process intensity. Such a model can still be clinically meaningful, especially as a surveillance aid. But it should not be interpreted as though those inputs were neutral measurements of a hidden biological state. They are sampled from a hospital process that varies across settings.

The sections that follow therefore treat the published experiment as a baseline operating state. From that state, three perturbations are introduced. The first is outcome perturbation, where the recorded label is treated as a noisy proxy for the underlying complication. The second is observation perturbation, where feature availability and care-intensity variables are treated as parts of the signal rather than as neutral covariates. The third is prevalence perturbation, where the same score is imagined in settings with different event rates. These perturbations are not hypothetical distractions. They correspond to issues already acknowledged in the source paper, namely the lack of chart-level endpoint validation, the dependence on structured data available in one institution's critical-care system, and the uncertainty of generalization beyond the development environment. What follows is an attempt to quantify the consequences of those issues rather than merely to mention them.

3. Outcome Perturbation and Label-Noise Effects

The recorded leak endpoint in the source study was a composite, not a directly adjudicated clinical truth. That means the observed outcome $\tilde{Y}$ can be treated as a noisy version of an unobserved true state Y . Let $\alpha = \Pr \tilde{Y} = 1 \mid Y = 0$ denote the false-positive label rate and $\beta = \Pr \tilde{Y} = 0 \mid Y = 1$ the false-negative label rate. Then the observed prevalence $\tilde{\pi}$ satisfies

$$\begin{aligned}\tilde{\pi} &= \Pr \tilde{Y} = 1 \\ &= 1 - \beta\pi - \alpha(1 - \pi),\end{aligned}\tag{3.1}$$

where $\pi = \Pr Y = 1$ is the unknown true prevalence. Even without assigning exact values to α and β , this identity shows that the observed 7.6% prevalence is only one possible realization of the underlying clinical incidence. If the endpoint misses some true leaks because they were managed conservatively without meeting the composite rule, then $\beta > 0$ and the true prevalence is higher than observed unless false-positive labeling is equally strong in the opposite direction [11]. If the endpoint also includes a small number of cases where coded concern and antibiotics did not represent a true leak, then both α and β matter.

This matters because thresholded performance against $\tilde{Y}$ is not the same as thresholded performance against Y . Suppose a model produces a binary positive set $A_t = \{s_i \geq t\}$ at threshold t , with true sensitivity $rt = \Pr A_t = 1 \mid Y = 1$ and false-positive rate $qt = \Pr A_t = 1 \mid Y = 0$. The sensitivity against the noisy label is

$$\begin{aligned}\tilde{rt} &= \Pr A_t = 1 \mid \tilde{Y} = 1 \\ &= \frac{1 - \beta\pi rt - \alpha(1 - \pi qt)}{1 - \beta\pi - \alpha(1 - \pi)}.\end{aligned}\tag{3.2}$$

Similarly, the observed precision against the noisy label is

$$\begin{aligned}\tilde{pt} &= \Pr \tilde{Y} = 1 \mid A_t = 1 \\ &= \frac{1 - \beta\pi rt - \alpha(1 - \pi qt)}{\pi rt - 1 + \pi qt}.\end{aligned}\tag{3.3}$$

These equations are useful because they separate two sources of uncertainty. One source is the classifier's own thresholded behavior, represented by rt and qt . The other source is the mismatch between the recorded label and the underlying event, represented by α and β . When the source paper reports precision 0.43 or recall 0.41, those values are precise only with respect to $\tilde{Y}$, not necessarily with respect to the underlying complication.

The direction of bias depends on which form of label error dominates. If false-negative outcome labeling is the larger problem, then measured recall can be pessimistic. A model may correctly place high scores on some admissions that truly developed leak, but if those admissions never met the composite rule, they are counted as false positives or missed positives relative to $\tilde{Y}$. In contrast, if the composite endpoint occasionally marks non-leak admissions as positive because of nonspecific antibiotic use or procedural coding, then measured precision can be distorted in the opposite direction. The point is not that all published metrics are unreliable. The point is that they sit on top of an outcome model that is itself estimated with error.

To make the scale concrete, consider a modest perturbation regime. Suppose the observed prevalence is 0.076, and suppose outcome construction has false-positive rate $\alpha = 0.01$ and false-negative rate $\beta = 0.10$. Solving the prevalence identity gives a true prevalence near 0.083. This is not a dramatic change in absolute terms, but it is meaningful for thresholded interpretation. Under that regime,

a classifier that appears to operate at 43% precision against the recorded label may be slightly under- or over-estimating its true positive predictive value depending on how its high-score admissions overlap with the misclassified region. If false-negative outcome labels are concentrated among admissions that the model scores highly, then measured precision is artificially depressed. If false-positive outcome labels are concentrated in parts of the feature space where the model also scores highly, the opposite can happen.

A more important effect may fall on threshold selection itself [12]. The source study chose thresholds to maximize F_1 on the validation set. But F_1 is a function of the observed label. If $\tilde{Y}$ differs from Y , then the validation threshold is tuned to the wrong target. Let $\tilde{F}_1 t$ denote the observed F_1 under noisy labels and $F_1 t$ the unobserved value under true labels. The optimal thresholds satisfy

$$\begin{aligned}\tilde{t}^* &= \arg \max_t \tilde{F}_1 t \\ t^* &= \arg \max_t F_1 t.\end{aligned}\tag{3.4}$$

There is no reason these must coincide. If the noise process is unevenly distributed across score space, the threshold maximizing the observed objective can be shifted away from the true optimum. This is especially plausible in a composite surgical endpoint, where missed true leaks may cluster among patients with atypical management pathways and false positives may cluster among patients who received aggressive workup for other reasons. The threshold is then shaped not only by the distribution of risk but also by the distribution of coding and corroborating actions.

The consequence for model comparison can be large enough to matter. Suppose the bagged gradient boosting model concentrates true positives efficiently but some of its predicted positives fall into a region where the composite outcome misses true leaks because they were not captured by supporting procedural evidence. Then the model's measured precision is lower than its true precision against the clinical endpoint. A more diffuse model such as logistic regression may place some of these same admissions below threshold and therefore appear relatively less penalized. Under that scenario, part of the apparent gap between models is a function of how outcome misclassification overlaps with score distribution, not just a function of raw predictive ability.

This effect is easiest to see in the false-positive cell. In the reconstructed held-out behavior, bagged gradient boosting yields roughly 26.6 false positives and the ensemble roughly 34.6. If even a small fraction of the bagged model's false-positive set were actually unrecorded true leaks, its effective precision would increase noticeably because the selected set is small. For a model with around forty-seven predicted positives, reclassifying just four admissions changes precision by about 0.085. That is a large movement. The same number of relabeled admissions would shift logistic regression less because its selected set is much larger. Thus endpoint uncertainty does not affect all models equally [13]. More concentrated models are often more sensitive to small changes in label assignment because each admission in the upper tail carries more leverage.

A useful way to summarize this is to consider corrected precision under one-sided false-negative label noise. If the predicted-positive group contains proportion δ_m of admissions that are truly positive but recorded negative, then the observed precision $\tilde{p}_m$ understates true precision by approximately

$$p_m^{true} = \tilde{p}_m / \delta_m - \tilde{p}_m.\tag{3.5}$$

This is a simplified expression, but it captures the main point. In a concentrated positive set, even a modest δ_m moves the estimated positive predictive value materially. If $\tilde{p}_m = 0.43$ and $\delta_m = 0.08$, then p_m^{true} rises to about 0.476. The same adjustment applied to logistic regression with $\tilde{p}_m = 0.26$ rises only to 0.319. Thus the empirical penalty of label noise is not uniform across model families.

The same logic extends to calibration. Brier score is often interpreted as if it measured intrinsic probability quality. Under noisy labels, it measures probability quality relative to the recorded endpoint. Let P be the model’s predicted probability of the true event Y . The observed Brier score is

$$\begin{aligned} \text{BS}_{obs} &= \mathbb{E}[P - \tilde{Y}^2] \\ &= \mathbb{E}[P - Y^2] \mathbb{E}[Y - \tilde{Y}^2] \\ &\quad 2\mathbb{E}[P - YY - \tilde{Y}] . \end{aligned} \tag{3.6}$$

Unless the noise term is independent of the model error term, the observed Brier score is not simply the true Brier score plus a constant. This is relevant because the composite endpoint is built from procedural and therapeutic evidence that may correlate with the very features the model uses. A model scoring high on cases with ICU admission, elevated white blood cell count, and raised lactate may be more likely to overlap with the care-intensity pathways that also affect whether a leak is coded and corroborated. Once again, this does not invalidate the original evaluation. It shows why calibration should be interpreted as conditional on endpoint construction.

From this perspective, the main empirical message is straightforward. The proof-of-concept study demonstrated real signal under a pragmatic outcome definition. That is an important result. At the same time, the measured thresholded behavior should be understood as the behavior of the model against a composite endpoint with nonzero uncertainty [14]. Small perturbations in phenotype construction can move precision, recall, threshold choice, and calibration enough to affect how model differences are interpreted. For rare postoperative outcomes, this is not a secondary technical detail. It is part of the outcome itself.

4. Observation Dependence and Feature-Availability Structure

Table 1: Held-out Model Performance Summary

Model	AUC	Precision	Recall
Logistic Regression	0.738	0.26	0.57
Naive Bayes	0.699	0.12	0.71
Gradient Boosting	0.754	0.32	0.51
Target-Encoded GB	0.764	0.34	0.43
Bagged GB	0.772	0.43	0.41
Ensemble	0.762	0.41	0.49

Table 2: Approximate Confusion Counts (Test Set)

Model	TP	FP	FN
Logistic Regression	27.9	79.5	21.1
Naive Bayes	34.8	255.1	14.2
Gradient Boosting	25.0	53.1	24.0
Target-Encoded GB	21.1	40.9	27.9
Bagged GB	20.1	26.6	28.9
Ensemble	24.0	34.6	25.0

The second stress test concerns how the hospital’s observation process enters the feature space. In the source study, the most influential variables were not all baseline patient attributes. They included

Table 3: Selected Fraction Above Threshold

Model	Selected Fraction	Interpretation
Logistic Regression	0.160	Broad selection
Naive Bayes	0.433	Very broad
Gradient Boosting	0.117	Moderate
Target-Encoded GB	0.093	Narrow
Bagged GB	0.070	Most selective
Ensemble	0.087	Balanced

Table 4: Top Predictive Features

Feature	Type	Role
WBC Count	Physiologic	Inflammation signal
ICU Admission	Process	Care intensity
Lactate	Physiologic	Perfusion marker
Surgery Type	Context	Procedural risk
Albumin	Baseline	Nutritional reserve

Table 5: Label Noise Parameters and Effects

Parameter	Meaning	Example Value	Impact
α	False positive rate	0.01	Inflates prevalence
β	False negative rate	0.10	Missed true cases
$\hat{\pi}$	Observed prevalence	0.076	Biased estimate
π	True prevalence	0.083	Adjusted value

Table 6: Brier Score Comparison

Model	Brier Score	Interpretation
Logistic Regression	0.059	Well-calibrated
Naive Bayes	0.415	Poor calibration
Gradient Boosting	0.059	Stable
Bagged GB	0.059	Strong calibration
Ensemble	0.059	Reliable

Table 7: Precision Under Prevalence Shift (Bagged GB)

Prevalence	PPV	Interpretation
0.076	0.43	Baseline
0.04	0.28	Reduced precision
0.02	0.16	Low yield
0.12	0.56	High confidence

postoperative white blood cell count, ICU admission, postoperative lactate, surgery type, and preoperative albumin. This ranking is informative because these variables sit at different points along the path from patient state to recorded data. White blood cell count and lactate reflect physiologic stress, but they also depend on test ordering and timing. ICU admission reflects a management decision as much as a patient attribute. Surgery type reflects procedural context. Albumin reflects baseline reserve but is also subject to selective availability. The model therefore reads both the patient and the way the patient was observed.

Table 8: Observation Dependence Factors

Variable	Dependence Type	Sensitivity
ICU Admission	Institutional	High
Lactate	Measurement	Medium
Albumin	Availability	High
WBC	Routine Lab	Low
Surgery Type	Structural	Low

Table 9: Model Robustness Under Stress Tests

Model	Label Noise	Obs. Shift	Prev. Shift
Logistic Regression	Medium	Medium	Weak
Naive Bayes	Weak	Weak	Weak
Gradient Boosting	Strong	Medium	Medium
Bagged GB	Strong	Sensitive	Strong
Ensemble	Strong	Strong	Strong

A compact way to represent this is to distinguish between a latent clinical state z_i and an observation operator M_i for admission i . The recorded feature vector is then

$$x_i^{obs} = M_i z_i + I - M_i u + \varepsilon_i, \quad (4.1)$$

where u is an imputation anchor and ε_i represents measurement error. The operator M_i is not merely missingness in the narrow sense. It captures whether the hospital measured the variable at all, how close to surgery it was measured, and whether the patient passed through a care pathway where the measurement was routine. For a variable such as lactate, M_i is tightly connected to clinical acuity. For a variable such as preoperative albumin, M_i is tied to the patient’s route into surgery and the completeness of preoperative workup [15].

This structure matters because the model is trained on x_i^{obs} , not on z_i . If two hospitals differ in how often albumin is available, or in how often postoperative lactate is drawn, then the effective feature distribution changes even if the underlying biology does not. That is part of why the source study’s simple missingness indicators are analytically important. They allow the model to learn that absence itself carries information. In a single-hospital proof-of-concept setting, this can improve performance. In a transport setting, it means that the model is partly tuned to the hospital’s measurement policy.

The high rank of ICU admission illustrates the point clearly. ICU admission within the first postoperative day is not a biological variable. It aggregates a set of judgments about operative complexity, instability, anticipated monitoring needs, and bed availability. In the source cohort, ICU admission occurred much more often among leak cases than among non-leak cases. As a predictor, it is therefore strong. But the relationship is partly institutional. A hospital that routinely sends major colorectal cases to a higher-acuity unit after surgery will assign different statistical meaning to ICU status than a hospital that reserves ICU beds for frank instability. The variable is clinically useful in both places, but its conditional distribution differs.

A similar argument applies to postoperative lactate. In a dataset drawn from critical care admissions, lactate may be measured frequently enough that its absence and presence each carry interpretable information. In a lower-acuity hospital setting, postoperative lactate may be ordered only selectively, which changes both the marginal missingness and the conditional meaning of the values that are present. The model trained in the first environment will therefore conflate physiologic deviation and observation intensity unless those processes are modeled separately.

This can be made explicit through subgroup-specific covariance. Let g index hospitals or care pathways [16]. Then the observed covariance for group g can be represented as

$$\Sigma_g^{obs} = A_g \Sigma_g^{latent} A_g^\top R_g, \quad (4.2)$$

where A_g maps latent physiology and context into observed data and R_g captures measurement noise and unmodeled process variation. If A_g changes across settings, then a model trained on one covariance structure may preserve some ranking ability but alter the contribution of specific variables and the density of the upper score tail. This is a more precise way to state a common concern about generalizability. The problem is not only that the patient mix changes. The operator that turns clinical states into recorded inputs changes as well.

The source study’s variable ranking strongly suggests that this is happening. White blood cell count, lactate, ICU admission, and albumin do not represent independent blocks of information. They are correlated through perioperative stress, underlying reserve, and care-process intensity. That correlation is one reason naive Bayes performed poorly. The independence assumption treats these variables as if they provide separate evidence streams. In the actual data, their dependencies encode much of the structure. A model that multiplies them as though they were independent can still rank some admissions correctly, but its probability scale becomes distorted. The reported Brier score of 0.415 for naive Bayes is entirely consistent with this interpretation.

For the stronger nonlinear models, the same dependence structure becomes an advantage. Tree-based learners are able to partition the cohort according to combinations such as high white blood cell count together with ICU admission, or low albumin together with a particular surgery type. Because the distributions are skewed, especially for white blood cell count, lactate, and albumin, the model gains by making local cuts in the tail rather than by fitting a smooth additive slope. This is one reason the tree-based configurations achieved much higher precision than logistic regression without necessarily recovering far more positive cases. They found a narrower upper-risk region defined by interacting variables whose statistical meaning depends on joint occurrence [17].

What matters for the present paper is that this narrowed upper-risk region is partly an observation artifact in the neutral sense of the term. It is real within the data-generating process, but that process includes how patients were monitored and how frequently they entered intensive care. A clinician using such a model in the same hospital can still benefit from it. A reader interpreting the model as a portable physiologic rule should be more cautious.

The role of missingness strengthens this conclusion. The source study notes that white blood cell count, creatinine, and hemoglobin had low missingness because they were part of standard perioperative panels, whereas preoperative albumin and body mass index had higher missingness because they depended on longer time windows or outpatient documentation. Missingness indicators were then added as features. This means the model is partly learning from who lacked a preoperative workup or complete baseline documentation. In a dataset with many emergent or critically ill admissions, that may be useful. It also ties the model’s score to a particular hospital workflow. In a lower-acuity environment with more uniform preoperative documentation, the same missingness pattern would carry less information.

A simple perturbation argument shows the effect on score stability. Suppose a score function near a given admission can be locally approximated as

$$\eta_i = w^\top x_i^{obs} \quad x_i^{obs\top} B x_i^{obs}. \quad (4.3)$$

A change in the observation pattern Δx_i induced by different measurement availability or care-path routing shifts the score by

$$\Delta\eta_i = w^\top \Delta x_i + x_i^{obs\top} B \Delta x_i - \Delta x_i^\top B x_i^{obs} - \Delta x_i^\top B \Delta x_i. \quad (4.4)$$

Even if the underlying clinical state does not change, the observed score can move because the record through which the model sees the patient has changed. This is not an abstract concern. For variables like ICU admission or missing albumin, Δx_i can be substantial across hospitals or even across service lines within one hospital.

The empirical implication is that feature importance in the source study should be read as importance within an observation architecture. White blood cell count matters because it is both biologically relevant and widely available [18]. ICU admission matters because it compresses a broad set of acute-care signals into one state variable. Albumin matters because reserve matters and because its availability is itself patterned. These are valid findings. They become more informative when interpreted as products of the measurement process rather than as timeless biologic rankings.

This does not diminish the original experiment. It clarifies what kind of model it produced. The resulting classifier is best viewed as an early postoperative surveillance model keyed to the structure of routine hospital records. It is not a preoperative calculator, and it is not a pure physiologic detector. Once framed that way, the next stress test follows naturally. If the score's meaning already depends on label construction and observation architecture, then its thresholded positive predictive value must also depend strongly on prevalence in the destination setting.

5. Threshold Meaning Under Prevalence Shift

A thresholded risk model does not carry the same practical meaning in every hospital. Even if its ranking ability remains constant, the positive predictive value of a fixed threshold depends directly on prevalence. This is standard probability theory, but it is easy to understate in clinical prediction because performance tables are often read as if precision were a property of the model alone. It is not. If rt is sensitivity and qt is false-positive rate at threshold t , then positive predictive value satisfies

$$PPV_t; \pi = \frac{\pi rt}{\pi rt + 1 - \pi qt}. \quad (5.1)$$

For a given threshold, the functions rt and qt may remain unchanged when the model is transported, but PPV changes whenever prevalence π changes. This matters a great deal in the source setting because the leak rate was around 7.6%, and the hospital was a tertiary academic center with a large critical-care footprint. A lower-acuity hospital or a service with different case mix could have materially different prevalence.

The reconstructed operating points make the sensitivity of PPV easy to quantify [19]. For bagged gradient boosting, the implied sensitivity is roughly 0.41 and the implied false-positive rate is about 0.043. At prevalence 0.076, the positive predictive value is the reported 0.43. If the same threshold were applied in a setting with prevalence 0.04, then

$$PPV_{bag, 0.04} = \frac{0.04 \cdot 0.41}{0.04 \cdot 0.41 + 0.96 \cdot 0.043}, \quad (5.2)$$

which is approximately 0.28. If prevalence fell further to 0.02, the same threshold would produce positive predictive value around 0.16. The model would still rank cases better than chance, but the meaning of a positive alert would be very different. A threshold that selected a dense subset in the original hospital would now select a much more diluted subset.

The same calculation shows the opposite effect in a higher-prevalence setting. If prevalence rose to 0.12 while the bagged model retained the same r and q , the positive predictive value would rise above 0.56. Thus the same score threshold can behave like a moderately selective surveillance trigger in one hospital and a high-confidence warning in another. This change requires no modification of the model itself. It is driven entirely by prevalence.

This is one reason the source study's F_1 -selected thresholds should not be interpreted as portable operating points. F_1 weights precision and recall symmetrically, and precision is prevalence-dependent. A threshold chosen to maximize F_1 in a 7.6% environment is therefore implicitly tuned to that prevalence. When prevalence changes, the threshold that maximizes F_1 changes even if the score distribution conditional on outcome remains stable. A hospital adopting such a model without retuning would therefore inherit a threshold calibrated to someone else's event rate.

A standard prevalence-correction formula makes this explicit for calibrated probabilities. If a model outputs a calibrated posterior p_s under source prevalence π_s , then under target prevalence π_t and stable likelihood ratios, the adjusted posterior is [20]

$$\begin{aligned} p_t &= \frac{\kappa p_s}{\kappa p_s + 1 - p_s} \\ \kappa &= \frac{\pi_t(1 - \pi_s)}{\pi_s(1 - \pi_t)}. \end{aligned} \tag{5.3}$$

This expression is useful because it shows how much re-interpretation is required before any threshold is even discussed. A score that means 0.43 risk in the source environment may correspond to a substantially lower target risk if prevalence falls. Thresholds based on unadjusted source probabilities then become systematically misleading.

The effect is not limited to PPV. The number of predicted-positive admissions also shifts in practical importance under prevalence change because the false-positive component, $1 - \pi q t$, scales with the negative majority. In lower-prevalence settings, most of the positive set comes from the false-positive term unless $q t$ is extremely small. This creates an operational challenge. A hospital can inherit the same model and preserve its discrimination, yet find that the number of low-yield workups becomes unacceptable because the baseline event rate is lower than in development.

A brief comparison of the main models shows how unevenly they are exposed to this effect. Logistic regression has implied false-positive rate around 0.128. Bagged gradient boosting is much lower at 0.043. The ensemble is around 0.056. If prevalence drops, the ranking among these models becomes more sharply separated on the precision scale because the model with higher false-positive rate loses positive predictive value faster. At prevalence 0.04, logistic regression's positive predictive value falls to about 0.16, whereas the ensemble remains around 0.27 and the bagged model around 0.28. Thus transport to a lower-prevalence environment does not merely lower precision overall. It changes the relative attractiveness of the model families.

This observation also feeds back into model interpretation. In the source setting, one might plausibly debate whether logistic regression's higher recall is worth its larger alert burden. In a lower-prevalence setting, that debate becomes harder to sustain because the false-positive term dominates more strongly [21]. In a higher-prevalence setting, the trade-off may look more acceptable. This is why a single thresholded comparison from one hospital cannot answer the question of which model is best in general. It can answer only which model occupies a certain operating region under one event rate.

Prevalence shift and observation shift are also linked. Hospitals with different case mix often differ not only in leak incidence but also in how often they measure lactate, how often they send patients to higher-acuity units, and how complete preoperative documentation is. Thus the model transported to a new hospital can experience both altered π and altered A_g in the covariance notation of the previous section. Positive predictive value changes through prevalence, while score distribution changes through observation process. The combined effect can be larger than either factor alone.

A practical conclusion follows. Any claim that a thresholded postoperative leak model is ready for use outside its development setting should be accompanied by at least two recalibration steps. The first is outcome audit, to determine whether the target site's leak phenotype aligns with the source phenotype closely enough that the label means the same thing. The second is prevalence-aware threshold revision, because a cutpoint that produced reasonable precision and recall in one environment is not expected to preserve those values in another. Without these steps, the model may still rank cases reasonably well, but the operational meaning of its positive predictions becomes uncertain.

The source experiment itself is useful precisely because it makes this dependence visible. Its better nonlinear models achieved improved selectivity in a setting with prevalence around 7.6%, good availability of common perioperative laboratories, and a substantial critical-care component. That is a real result. What this section adds is that the thresholded interpretation of those models is conditional on that environment. Moving the same threshold into a different prevalence regime changes what the number means. In a hospital prediction context, that is not a trivial footnote [22]. It is part of the empirical result.

6. Re-reading the Model Comparison Under These Stress Tests

Once outcome noise, observation dependence, and prevalence shift are all on the table, the source model comparison looks different. The major result remains intact: the stronger nonlinear configurations are preferable to naive Bayes and offer a more selective operating regime than logistic regression. But the reason they are preferable is more specific than a standard AUC comparison implies. They are preferable because they carve out a tighter upper-risk subset in the source environment, using combinations of variables whose meaning depends on the first postoperative day's measurement structure. That is a useful property. It is not identical to universal superiority.

Start with naive Bayes. Its high recall relative to the other models might tempt a superficial reading that it simply operates at a more sensitive point. The Brier score rules out that interpretation. The model is not merely choosing a different position on the same frontier. Its probability scale is poor because the feature dependencies in the data violate the independence assumption. Under the label-noise analysis above, this weakness becomes more consequential, not less. If the endpoint is already imperfect, then a model that also mishandles dependence compounds the interpretive problem. For practical purposes, naive Bayes is not only broad in its positive partition. It is unstable in how it represents uncertainty.

Next consider logistic regression. Its behavior is more defensible than naive Bayes because its probability scale is reasonable and its recall is higher than the more selective tree-based models. Under the stress tests developed here, however, logistic regression's weakness becomes clearer [23]. Its upper tail is comparatively broad. That broadness makes it especially vulnerable to prevalence shifts, because precision deteriorates rapidly when event rate falls. It also means that if the recorded endpoint misses some true leaks, the large false-positive cell contains a mixture of truly low-risk patients and potentially mislabeled high-risk patients, which makes post hoc interpretation difficult. Logistic regression still has value as an interpretable baseline, but its operating regime is less robust than the source table alone might suggest.

The more interesting comparison is between bagged gradient boosting and the three-way ensemble. The source study treated the bagged model as the best AUC performer and the ensemble as the best PR-AUC performer. Under the present stress tests, that distinction acquires a more operational interpretation. The bagged model is the more concentrated of the two. Its positive set is smaller, so its empirical precision can move materially if a few of its apparent false positives are actually mislabeled true leaks. This means label-noise correction could make it look even stronger than the reported point estimate suggests. At the same time, because its selected set is narrow, it may be somewhat more sensitive to observation-process changes across hospitals. If the upper-tail pattern it learned depends strongly on the source institution's ICU routing or measurement frequency, that concentration may degrade in transport.

The ensemble is slightly less concentrated but likely more forgiving. Its positive set is larger, so each admission carries a bit less leverage on the measured precision. Its reported PR-AUC advantage suggests a broader region of usable operating points rather than a single very sharp optimum. Under

prevalence shift, it will still lose precision, but less dramatically than logistic regression because its false-positive rate remains moderate. Under mild outcome misclassification, its precision can also improve if some of its false positives are truly positive cases missed by the composite label. In practical terms, the ensemble may be the model whose threshold is easiest to revise across different clinical priorities, whereas the bagged model may be the one that offers the leanest alert strategy in a hospital that resembles the development setting closely.

The target-encoded gradient boosting model occupies a different niche [24]. Its precision and recall are both intermediate, but it offers little obvious advantage once the bagged and ensemble configurations are available. Under the current stress tests, its main limitation is that it is neither as concentrated as the bagged model nor as balanced as the ensemble. If one expects label-noise correction to rescue upper-tail precision somewhat, the bagged model benefits more. If one expects prevalence shift and threshold retuning to matter, the ensemble offers a broader operating surface. The target-encoded model therefore looks less compelling as a practical default even if it remains a technically competent configuration.

These revised readings do not depend on changing the fundamental results of the source paper. They depend on reading those results through the conditions under which they were obtained. A model that looks best in one metric can remain best after stress-testing, but the meaning of best changes. Instead of a generic ranking statement, one obtains a statement about which model is most concentrated under the source phenotype, which is most robust to moderate prevalence changes, and which is least vulnerable to dependence-induced calibration failure. That is a more informative basis for scientific comparison.

Another advantage of this re-reading is that it clarifies what the source experiment did not yet establish. It did not establish chart-level validity of the outcome. It did not establish subgroup stability across procedure types or demographic groups. It did not establish transport to lower-acuity hospitals. It did not establish whether the top predictors retain the same leverage when ICU admission practice or post-operative lab ordering differs. None of those absences cancel the study. They define the next empirical steps required if one wants to preserve the meaning of the reported thresholded behavior outside the original setting.

The combination of stress tests also explains why the source study's feature ranking should not be treated as a stable causal ordering. White blood cell count, lactate, ICU admission, surgery type, and albumin are all plausible predictors [25]. Their exact ordering, however, likely depends on the local pattern of observation and endpoint recording. In a site with more uniform preoperative albumin measurement, reserve variables might become more stable. In a site with less frequent postoperative lactate testing, physiologic tail variables might lose some of their apparent dominance. In a site without routine critical-care exposure for many colorectal cases, ICU status could change meaning substantially. This is not a weakness of feature importance methods alone. It is a property of the data-generating process.

What the source experiment provides, then, is a strong internal demonstration that early postoperative structured data can identify a high-risk subset better than simple linear or independence-based baselines. What the current paper adds is a tighter empirical statement of where that result is likely to bend. It bends at the label. It bends at the observation architecture. It bends at the threshold when prevalence changes. These are exactly the places where real-world transport problems tend to emerge. Framing them explicitly makes the original result more useful, not less.

7. Discussion

The main contribution of this study is not a new algorithm. It is a more constrained interpretation of an existing experiment. The proof-of-concept study under examination showed that early structured hospital data contain enough signal to support meaningful risk stratification for anastomotic leak. That conclusion remains sound after the stress tests performed here. What changes is the level of confidence one should attach to the thresholded numbers themselves and the scope over which one should expect them to travel unchanged.

Three findings are the most important [26]. First, the outcome definition matters more than the usual performance table reveals. The recorded leak endpoint is a composite of codes and supporting clinical

evidence. That is a reasonable choice when chart review is not feasible for thousands of admissions. It also means that the model is judged against an observed label that can miss some true events and include some borderline or misclassified cases. Even modest label error can change precision and recall enough to affect interpretation, especially for concentrated models whose positive sets are small. This does not mean the study should have waited for perfect labels before proceeding. It means that empirical papers of this kind should treat phenotype uncertainty as part of the evaluation problem rather than as a limitation paragraph detached from the main results.

Second, the feature space is inseparable from the observation process. The top variables in the source study were not all patient constants. They included care-location status and selectively available laboratory values. The model therefore learned from the hospital's pattern of seeing the patient as well as from the patient's clinical state. That is often exactly what a useful postoperative surveillance model should do. The caution is interpretive. A feature that is highly informative in one observation system can change meaning in another. For this reason, model transport should not be framed only as a matter of recalibrating probabilities. It also requires checking whether the observation operator itself has changed.

Third, threshold meaning is prevalence-bound. Precision was reported at thresholds chosen in a setting with prevalence around 7.6%. The same thresholds would produce materially different positive predictive values in hospitals with lower or higher event rates [27]. This is true even if ranking quality is unchanged. A model paper that reports precision and recall without discussing prevalence transport leaves the reader with an incomplete understanding of what those numbers represent. In rare-event surgical outcomes, prevalence adjustment is not optional once the model leaves the site of development.

These findings suggest a revised empirical reporting standard for this domain. A proof-of-concept study should still report AUC, PR-AUC, calibration, precision, and recall. In addition, it should report a phenotype audit or at least a sensitivity analysis for plausible label error, because composite coded endpoints are rarely noise-free. It should report how the selected positive set depends on variables that also encode care-process intensity, such as ICU admission or selectively ordered laboratories. It should report how positive predictive value would change over a clinically plausible range of prevalences. None of these additions requires a new modeling paradigm. They require a more explicit link between the data-generating process and the performance claim.

The source paper already contains many of the ingredients needed for this style of reporting. It states the outcome construction clearly. It acknowledges the lack of chart-level validation. It identifies the first postoperative day as the intended feature window. It gives the held-out metrics for multiple model families and names the most influential variables. This transparency is one of its strengths. The present paper extends that transparency by translating the reported design choices into explicit consequences. A reader now knows not only that bagged gradient boosting reached precision 0.43 and recall 0.41, but also that this precision sits inside a specific composite phenotype, depends partly on observation architecture, and would not preserve the same meaning under lower prevalence without threshold revision.

This narrower interpretation does not reduce the value of the original work [28]. It protects it from overclaim. One of the risks in surgical machine learning is that moderate proof-of-concept performance can be narrated as a nearly deployable solution. That does not serve clinicians well, and it does not serve the underlying science well either. An early leak model that identifies a concentrated high-risk group in a tertiary critical-care environment is a useful scientific result. It becomes more valuable, not less, when the conditions of that result are stated clearly.

There are also implications for future empirical work. The next strong study in this area would likely combine three design upgrades. The first would be at least a partial chart-adjudication sample to estimate the direction and scale of label error in the composite endpoint. The second would be multi-site evaluation with explicit measurement-process comparison, not merely external AUC reporting. The third would be prevalence-aware threshold analysis so that positive predictive value is not treated as fixed across settings. These steps would move the field from proof-of-concept classification toward a more stable account of what the model is actually measuring.

The present analysis has limitations of its own. It reconstructs operating behavior from published summary metrics rather than from patient-level predictions. It therefore cannot produce exact calibration

curves, subgroup-specific perturbation results, or formal confidence intervals for all derived quantities. Its label-noise analysis is based on plausible perturbation structure rather than on directly observed misclassification. Its observation-process analysis uses the source study's variable ranking and missingness description rather than raw measurement timestamps. These limits should be kept in view. At the same time, the central conclusions do not depend on exact patient-level reconstruction. They follow from the basic mathematics of noisy labels, observation-dependent features, and prevalence-dependent precision, all applied to a well-specified source experiment [29].

A final point deserves emphasis. The usual contrast between proof-of-concept and implementation can sometimes obscure an intermediate step that is just as important: interpretation under perturbation. Before a model is built into workflow, one should know what happens when the endpoint shifts slightly, when the observation system changes, and when prevalence moves away from the development setting. This paper has tried to supply that intermediate step for one concrete experiment. The result is a more disciplined reading of what the model table means. In a field where rare outcomes, partial labels, and hospital-specific workflows are the norm, that kind of reading should become routine.

8. Conclusion

This study re-examined an early postoperative leak prediction experiment by treating the reported model performance as conditional on three empirical structures: the way the outcome was defined, the way the hospital observed patients in the first postoperative day, and the prevalence of the event in the development setting. Under that re-analysis, the main source result remains credible. Structured hospital data available within 24 hours after colorectal surgery contain enough information to identify a subset of admissions at meaningfully elevated risk of later leak. The stronger nonlinear models remain the most useful configurations within that setting.

What changes is the interpretation of the reported thresholded metrics. Precision and recall are not free-standing properties of the model. They are shaped by the composite phenotype used to define leak, by the dependence of key predictors on measurement and care-intensity pathways, and by the baseline event rate of the source hospital. Modest label error can move thresholded performance appreciably, especially for concentrated upper-tail models. Observation-process variables such as ICU admission and selectively available laboratories can improve internal prediction while also narrowing portability. A threshold chosen in a 7.6% prevalence environment will not preserve the same positive predictive value elsewhere. Proof-of-concept leak prediction should be reported and interpreted with explicit stress tests for endpoint uncertainty, feature-availability dependence, and prevalence transport. Without those additions, a model table can appear more general than the data justify. With them, the same experiment becomes easier to trust because its limits are specified in the same empirical language as its strengths [30].

References

- [1] P. Dedeilias, I. Nenekidis, P. Hountis, C. Prokakis, P. Dolou, E. Apostolakis, and E. Koletsis, "Superior vena cava syndrome in a patient with previous cardiac surgery: what else should we suspect?," *Diagnostic pathology*, vol. 5, pp. 43–43, 6 2010.
- [2] P. S. Waters, O. Peacock, S. K. Warriar, C. Wakeman, T. Eglinton, A. C. Lynch, F. A. Frizelle, A. G. Heriot, and J. J. McCormick, "Evolution of pelvic exenteration surgery- resectional trends and survival outcomes over three decades.," *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 45, pp. 2325–2333, 7 2019.
- [3] J. LaFemina, E. Vinuela, M. A. Schattner, H. Gerdes, and V. E. Strong, "Esophagojejunal reconstruction after total gastrectomy for gastric cancer using a transorally inserted anvil delivery system," *Annals of surgical oncology*, vol. 20, pp. 2975–2983, 4 2013.
- [4] F. H. Rizvi, S. S. H. Rizvi, A. A. Syed, S. Khattak, and A. R. Khan, "Minimally invasive esophagectomy for esophageal cancer: The first experience from pakistan," *International journal of surgical oncology*, vol. 2014, pp. 864705–864705, 7 2014.

- [5] L. N. R and J. T, "A prospective study to understand the risk factors for anastomotic leak in small bowel anastomosis," *Journal of Evolution of Medical and Dental Sciences*, vol. 9, pp. 1549–1553, 5 2020.
- [6] M. Kukar, K. Ben-David, J. S. Peng, K. Attwood, R. M. Thomas, M. Hennon, C. Nwogu, and S. N. Hochwald, "Minimally invasive ivor lewis esophagectomy with linear stapled anastomosis associated with low leak and stricture rates," *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, vol. 24, pp. 1729–1735, 7 2019.
- [7] 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.
- [8] S. Roh, M. D. Iannettoni, J. Keech, M. Bashir, P. J. Gruber, and K. R. Parekh, "Role of barium swallow in diagnosing clinically significant anastomotic leak following esophagectomy.," *The Korean journal of thoracic and cardiovascular surgery*, vol. 49, pp. 99–106, 4 2016.
- [9] M. Mik, J. Magdzinska, L. Dziki, M. Tchórzewski, R. Trzciński, and A. Dziki, "Relaparotomy in colorectal cancer surgery – do any factors influence the risk of mortality? a case controlled study," *International journal of surgery (London, England)*, vol. 12, pp. 1192–1197, 9 2014.
- [10] L. N. Bhoir, N. Y. Jagne, and D. Murali, "Is using peritoneal drains in bowel surgeries beneficial," *International Surgery Journal*, vol. 4, pp. 650–655, 1 2017.
- [11] M. Ahmadinejad, M. Hashemi, and A. Tabatabai, "A comparative study between the postoperative complications of stripping esophagectomy and classic (orringer's technique) esophagectomy.," *Surgery journal (New York, N.Y.)*, vol. 8, pp. e34–e40, 2 2022.
- [12] H. Li, D. Wang, W. Wei, L. Ouyang, and N. Lou, "The predictive value of coefficient of $pct \times bg$ for anastomotic leak in esophageal carcinoma patients with ards after esophagectomy.," *Journal of intensive care medicine*, vol. 34, pp. 572–577, 5 2017.
- [13] H. Thakkar and K. Lakhoo, "The surgical management of necrotising enterocolitis (nec).," *Early human development*, vol. 97, pp. 25–28, 3 2016.
- [14] J. C. Sebastián-Tomás, A. Martínez-Pérez, E. Martínez-López, N. de'Angelis, M. G. Ruiz, and E. García-Granero, "Robotic transanal total mesorectal excision: Is the future now?," *World journal of gastrointestinal surgery*, vol. 13, pp. 834–847, 8 2021.
- [15] D. E. Yeung, E. Peterknecht, S. Hajibandeh, S. Hajibandeh, and A. W. Torrance, "C-reactive protein can predict anastomotic leak in colorectal surgery: a systematic review and meta-analysis," *International journal of colorectal disease*, vol. 36, pp. 1147–1162, 2 2021.
- [16] S. Lutfi and J. Camilleri-Brennan, "Ep.tu.91 l analysis of the outcomes following reversal of loop ileostomy," *British Journal of Surgery*, vol. 108, 10 2021.
- [17] H. Mazeh, A. Greenstein, K. A. Swedish, S. Q. Nguyen, A. Lipskar, K. J. Weber, E. H. Chin, and C. M. Divino, "Laparoscopic and open reversal of hartmann's procedure—a comparative retrospective analysis.," *Surgical endoscopy*, vol. 23, pp. 496–502, 7 2008.
- [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] T. N. Walsh, "The esophagogastric anastomosis: The importance of anchoring sutures in reducing anastomotic leak rates.," *Annals of surgery open : perspectives of surgical history, education, and clinical approaches*, vol. 4, pp. e231–e231, 2 2023.
- [20] H. H. Wasmuth, A. E. Faerden, T. Åge Myklebust, F. Pfeffer, S. Norderval, R. Riis, O. C. Olsen, J. R. Lambrecht, H. Kørner, S. G. Larsen, H. M. Forsmo, O. Baekkelund, S. Lavik, J. C. Knapp, O. Sjo, and G. Rashid, "Transanal total mesorectal excision for rectal cancer has been suspended in norway.," *The British journal of surgery*, vol. 107, pp. 121–130, 1 2020.
- [21] R. Jaibaji, M. Ibrahim, A. Tsai, and C. Peters, "Ogc p67 extended use of feeding jejunostomy and reduction of weight loss in patients with total gastrectomy for malignancy," *British Journal of Surgery*, vol. 110, 11 2023.
- [22] A. Powell, J. Wheat, N. Patel, D. S. Chan, A. Foliaki, S. A. G. Roberts, and W. G. Lewis, "Value of individual surgeon performance metrics as quality assurance measures in oesophagogastric cancer surgery.," *BJS open*, vol. 4, pp. 91–100, 2 2020.

- [23] F. Hayati, Z. A. M. Azman, D. N. Nasuruddin, L. Mazlan, A. D. Zakaria, and I. Sagap, "Serum procalcitonin predicts anastomotic leaks in colorectal surgery," *Asian Pacific journal of cancer prevention : APJCP*, vol. 18, pp. 1821–1825, 7 2017.
- [24] V. V. Simianu, A. Basu, R. Alfonso-Cristancho, R. C. Thirlby, A. D. Flaxman, and D. R. Flum, "Assessing surgeon behavior change after anastomotic leak in colorectal surgery," *The Journal of surgical research*, vol. 205, pp. 378–383, 7 2016.
- [25] C. S. Roxburgh, P. Strombom, P. B. Lynn, A. Cercek, M. Gonen, J. J. Smith, L. K. Temple, G. M. Nash, J. G. Guillem, P. B. Paty, J. Shia, E. Vakiani, R. Yaeger, Z. K. Stadler, N. H. Segal, D. L. Reidy, A. M. Varghese, A. J. Wu, C. H. Crane, M. J. Gollub, L. B. Saltz, J. Garcia-Aguilar, and M. R. Weiser, "Changes in the multidisciplinary management of rectal cancer from 2009 to 2015 and associated improvements in short-term outcomes," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 21, pp. 1140–1150, 7 2019.
- [26] G. Luglio and F. Corcione, "Stapled versus handsewn methods for ileocolic anastomoses.," *Techniques in coloproctology*, vol. 23, pp. 1093–1095, 11 2019.
- [27] S. Ahmed, M. Abbakar, K. Hassan, M. Mohammed, C. Magee, and J. Wilson, "Tu2.2 endo-sponge carries good outcome for low rectal anastomotic leak; case series study," *British Journal of Surgery*, vol. 109, 8 2022.
- [28] C. T. Nyabanga, J. C. Veniero, and B. Shen, "Computed tomography-guided endoscopic needle knife therapy for ileal pouch sinus," *Gastroenterology report*, vol. 4, pp. 334–336, 7 2015.
- [29] C. Li, D. T. Yamashita, J. D. Hawel, D. Bethune, H. Henteleff, and J. Ellsmere, "Endoscopic mucosal resection versus esophagectomy for intramucosal adenocarcinoma in the setting of barrett's esophagus," *Surgical endoscopy*, vol. 31, pp. 4211–4216, 3 2017.
- [30] J. C. Alverdy, "The role of dietary prehabilitation on anastomotic healing.," *Current opinion in clinical nutrition and metabolic care*, vol. 26, pp. 470–475, 6 2023.