Academy for Global Science and Innovation (AGSI)

ARTICLE

Temporal Signal Concentration and Misclassification Structure in Early Anastomotic Leak Prediction From Routine Perioperative Hospital Records

Chen Yong,[1] Liang Hao,[2] and Zhao Ming[3]

[1]School of Information Science and Engineering, Shandong Jianzhu University, No. 1000 Fengming Road, Licheng District, Jinan 250101, China

[2]Department of Computer Engineering, Zhejiang University of Science and Technology, No. 318 Liuhe Road, Xihu District, Hangzhou 310023, China

[3]College of Computer and Information Engineering, Inner Mongolia University of Technology, No. 49 Aimin Street, Xincheng District, Hohhot 010051, China

Abstract

Postoperative risk models are commonly summarized by a single discrimination score, yet that convention hides the practical structure of error. In anastomotic leak prediction, this omission is consequential because the event is infrequent, the clinical stakes are high, and modest changes in precision or false-alarm burden can alter whether a model is operationally credible. This paper presents an empirical secondary analysis of an early postoperative leak prediction experiment based on structured perioperative records. Rather than asking only which classifier attains the best overall ranking, the study examines where informative signal is concentrated, how false positives accumulate across model families, and how threshold behavior reshapes the balance between detection and workload. The analysis treats reported test-set performance as an observable object from which implied alert rates, event enrichment, and miss-to-capture trade-offs can be derived. Particular attention is given to the relation between nonlinear modeling, probability quality, and the asymmetry created by low event prevalence. The study finds that model differences are not exhausted by changes in area under the receiver operating characteristic curve. Some models compress a meaningful share of leak events into a relatively small alert surface, whereas others recover additional cases only by expanding false-positive volume substantially. The dominant predictors are also notable: early inflammatory and physiologic response variables outweigh many static baseline factors, suggesting that the first postoperative day already contains a concentrated but imperfect signature of later leak. The findings support a more granular empirical language for surgical prediction, one centered on error topology and operational yield rather than on discrimination alone.

1. Introduction

Anastomotic leak prediction sits in a difficult corner of surgical analytics because it is neither a purely mechanistic problem nor a purely statistical one [1]. The complication emerges from tissue healing failure, contamination burden, perfusion quality, operative stress, and host reserve, but the evidence available to a model is generated through ordinary hospital documentation. Laboratory values are ordered because someone decides to order them. Care location changes because clinicians already suspect instability. Even apparently simple variables, such as postoperative white blood cell count or intensive care unit admission, carry both physiologic content and organizational meaning. As a result, a prediction model is never reading an untouched biological signal. It is reading an event through the surface of perioperative measurement. This matters because performance is often reported as though the model

had solved a neutral classification task when in fact it has learned from a record shaped by timing, surveillance, and care intensity [2].

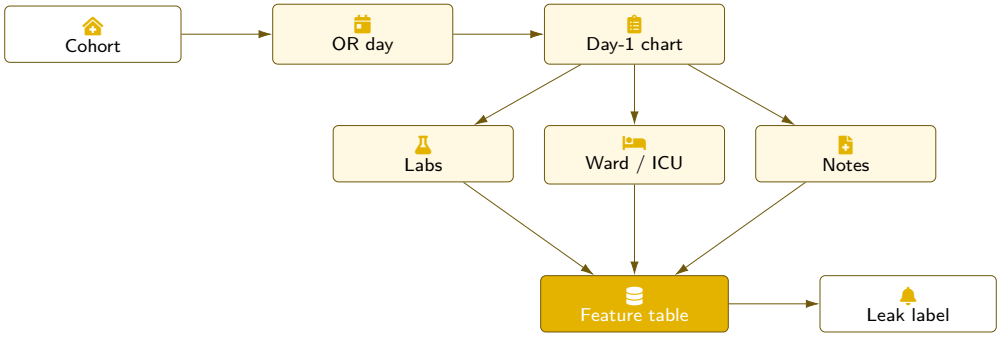


Figure 1. Cohort assembly and outcome construction within the first postoperative day. Structured perioperative observations are consolidated into a compact feature table and then linked to the downstream leak endpoint.

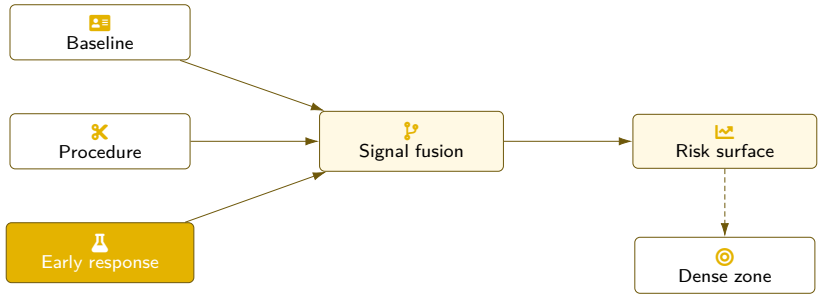


Figure 2. Conceptual decomposition of learnable signal. Dynamic day-1 response variables occupy the steepest part of the risk gradient, while baseline and procedural context shape how that response is interpreted.

Most of the literature on clinical prediction still privileges a narrow set of summary statistics. Area under the receiver operating characteristic curve remains the default headline, sometimes accompanied by precision-recall area under the curve and a calibration metric. Those summaries are useful, but they do not fully describe what a model is doing. Two classifiers can have similar discrimination and yet behave very differently once translated into case-finding burden. A model may achieve a respectable AUC while spreading its true positives across such a broad prediction surface that real use would demand large volumes of unnecessary review. Another may capture fewer events overall but do so within a sharply enriched subset that is much easier to act upon. In a rare postoperative outcome, this distinction is not cosmetic. It determines whether the model is merely interesting or meaningfully selective.

That distinction becomes even more important in leak prediction because the positive class is sparse. When prevalence is low, an algorithm can appear statistically competent while remaining operationally blunt. A clinician does not experience the model as a continuous curve in abstract metric space [3]. The clinician experiences it as a sequence of alerts, non-alerts, explanations, escalations, and misses. How many patients must be flagged to recover a clinically meaningful fraction of events? How many false alarms accompany each true case identified? How much enrichment above baseline prevalence does the alerting subset achieve?

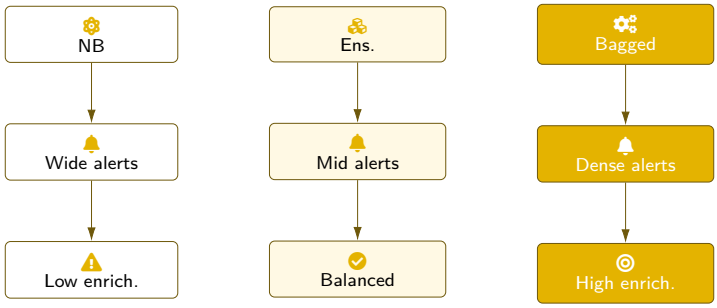


Figure 3. Operational error topology across representative models. A wider alert surface can recover more events but dilutes the flagged subset, whereas a denser alert surface preserves enrichment and lowers unnecessary review volume.

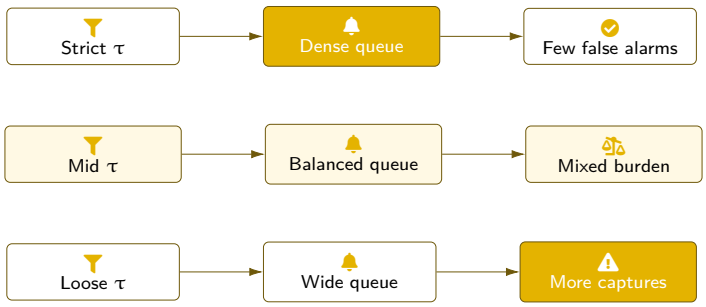


Figure 4. Threshold movement reshapes workload and case capture. Tight decision rules concentrate attention into a small review queue, while looser rules broaden the queue and recover additional events at the cost of extra burden.

These questions are often left implicit even though they speak directly to the practical geometry of the classifier. The language of ranking alone is insufficient for a complication whose clinical burden is concentrated in a small subset of admissions.

The empirical motivation of the present paper comes from that gap. Rather than proposing a new training pipeline, the paper re-examines an existing early postoperative leak prediction experiment through the lens of signal concentration and misclassification structure. This change in emphasis is deliberate. There are circumstances in which a secondary quantitative reading of reported results is more informative than building another model with a minor architectural variation. If the published experiment already reports prevalence, held-out sample size, precision, recall, discrimination, probability error, and ranked predictors, then those outputs themselves can be treated as analyzable data. One can derive approximate confusion geometry, prevalence-adjusted alert burden, and event enrichment without pretending to know more than the experiment disclosed. Such a strategy does not replace primary modeling. It extracts additional empirical meaning from what has already been measured.

This paper therefore studies early postoperative leak classification as a sparse-signal allocation problem. The central concern is not merely whether the models separate positive and negative admissions better than chance. It is whether they compress the available signal into a sufficiently concentrated region of the population to justify selective attention. The notion of concentration is useful here because it links statistical performance to operational selectivity. If a model doubles or triples the prevalence of leak in the flagged subset, that matters.

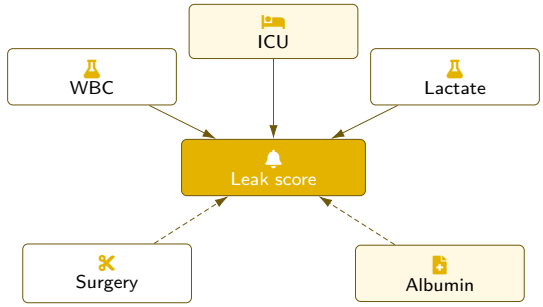


Figure 5. Predictor pattern suggested by the reported importance ranking. Immediate postoperative markers sit closest to the center of the score, while surgery type and preoperative albumin behave as contextual modifiers around that dynamic core.

If it raises enrichment fivefold while keeping alert volume modest, that matters even more. Conversely, if it recovers additional cases only by scattering alerts widely across the negative class, its practical worth becomes more questionable even when recall improves [4].

A second motivation concerns the interpretation of predictors. In many surgical datasets, there is an unresolved tension between baseline risk factors and early response variables. Preoperative nutritional reserve, procedure type, urgency, and body habitus are clearly relevant to healing. Yet once the first postoperative day unfolds, dynamic inflammatory and physiologic variables may dominate the discriminative landscape because they encode the earliest observable manifestations of deviation from routine recovery. When such variables outrank static baseline covariates, the model is not necessarily telling us that baseline biology is unimportant. It may instead be showing that the record becomes much more informative once the operation has already induced a measurable host response. This raises a substantive empirical question: is the learnable signal for leak prediction concentrated more strongly in immediate postoperative reaction than in preoperative predisposition? That is not a philosophical point. It affects where one would expect prediction to be strongest and what kind of surveillance instrument the model most resembles.

A third motivation lies in the asymmetry between discrimination and probability quality. In low-prevalence settings, a model can sort patients reasonably well while still issuing poorly scaled probabilities. That problem is especially serious if predicted risk is interpreted too literally. A clinician may tolerate a ranking score that says one patient deserves closer monitoring than another. It is harder to justify a probability estimate that is numerically unstable or badly calibrated. The empirical contrast between models that preserve usable probability error and models that do not is therefore important even before any deployment discussion begins. Probability quality shapes how confidently a score can be interpreted and how severely threshold choices will amplify or dampen false alarms.

The argument that follows is structured around these concerns. The paper first states the empirical substrate and the way this secondary analysis is framed. It then formalizes a set of derived quantities that expose the relation among prevalence, precision, recall, and alert burden. The subsequent sections examine the reported models through that lens, compare their implied operational regimes, and consider what the variable ranking suggests about where early leak signal actually resides in the perioperative record [5]. The closing discussion remains cautious. It does not claim that a secondary analysis of published results can settle the question of clinical readiness. It claims something narrower and more defensible: a better understanding of error structure can materially sharpen how one reads a surgical

prediction study, especially when event rarity makes operational burden as important as headline discrimination.

2. Study Setting and Outcome Construction

The empirical substrate for this paper is the structured early-postoperative leak prediction experiment reported by Sadanandan (2024) [6], built from the MIMIC-IV critical care database, restricted to colorectal resection admissions, and summarized by a cohort of 3,331 admissions from 2,847 patients with a leak prevalence of 7.6%, a patient-level split into training, validation, and held-out test partitions, a test set of 670 admissions with 49 positive outcomes, six model configurations, and a predictor set assembled from demographic information, admission context, procedure descriptors, laboratory measurements, and baseline anthropometrics drawn within the first 24 hours after surgery. That study is particularly suitable for secondary empirical examination because it reports not only discrimination but also precision, recall, Brier score, missing-data handling, and a ranked list of influential features. The present paper does not alter the underlying cohort or claim access to additional patient-level records. Instead, it treats the reported test-set behavior of the models as a fixed experimental output and examines the structure implied by those outputs.

Several elements of that design deserve emphasis because they shape what can and cannot be inferred. The cohort arises from a single tertiary academic center embedded within a large critical care environment. This matters because both case complexity and surveillance density are likely to be higher than in lower-acuity surgical settings. The outcome is defined as a composite of diagnostic coding for intestinal wound disruption or postprocedural intestinal fistula together with corroborating evidence such as return to the operating room, drainage, or broad-spectrum antibiotic administration within 30 days. That is an analytically sensible attempt to reduce isolated coding noise, but it also means the target is a clinically filtered event rather than a pure histopathologic fact. The model is therefore learning a serious postoperative failure process as recognized through routine hospital actions. This is entirely reasonable for a prediction experiment, yet it should be remembered when interpreting both sensitivity and feature importance. The positive class is not an unmediated biological truth. It is an event made visible through the hospital record.

The temporal window used for feature extraction is equally important. All predictors are limited to information available within 24 hours of surgery [7]. This choice excludes later postoperative signals that might strengthen classification, but it preserves the central empirical question of whether the first postoperative day already contains a nontrivial leak signature. Such a window is narrow enough to remain clinically early and broad enough to include immediate inflammatory, metabolic, and care-location responses. In this setting, a feature like postoperative lactate is not just a metabolic marker; it is a compressed summary of tissue stress, perfusion disturbance, and the intensity of acute resuscitative attention. Similarly, intensive care unit admission within 24 hours may reflect operative difficulty, early instability, monitoring preferences, or all three. The empirical value of these variables lies precisely in their composite nature. They are not clean mechanistic readouts, but they may be powerful early discriminators because they absorb several dimensions of postoperative deviation at once.

The studys missing-data strategy also shapes interpretation. Continuous variables were imputed using training-set medians, and categorical variables were given explicit unknown categories, while imputation indicators allowed models to learn from absence itself. In a perioperative dataset, this is not a trivial engineering detail. Missing albumin may indicate

lack of outpatient optimization, emergency presentation, or fragmented preoperative workup. Missing anthropometrics may reflect documentation asymmetry rather than random omission. By preserving missingness indicators, the original experiment allowed nonavailability to function as information. That decision is methodologically important because it aligns the model with the reality of hospital records, where what is not measured can correlate with how the case entered care. For the present secondary analysis, it also means that the reported model performance already incorporates a limited but meaningful handling of measurement asymmetry rather than a complete-case simplification.

The original paper compared six model configurations: logistic regression, naive Bayes, gradient boosting, gradient boosting with target encoding, a bagged gradient boosting ensemble of five learners, and a three-way ensemble. This range is analytically useful because the models differ in how they handle interaction, correlation, and class imbalance. Logistic regression provides an additive baseline. Naive Bayes imposes strong conditional independence assumptions and exposes what happens when probability structure is simplified aggressively. Gradient boosting accommodates threshold effects and interaction patterns without manual specification [8]. The bagged variant can reduce variance and stabilize decision boundaries. The three-way ensemble can redistribute errors across learners. Because the same held-out test set was used for all six, the results offer a compact empirical laboratory in which to study not only who ranks best, but how different inductive biases change the allocation of true and false positives.

The test-set size and event count are large enough to permit a useful, though still finite-sample, operational reading. With 670 admissions and 49 positives, each percentage point of recall corresponds to roughly half an event, and each percentage point of false-positive rate corresponds to slightly more than six negative admissions. This scale encourages modesty. Differences that appear visually large in performance plots can correspond to a relatively small number of admissions, while seemingly small changes in precision may produce noticeable differences in false-alarm burden. The present paper therefore reports several derived quantities as approximations, acknowledging that the original performance metrics were rounded. That caveat does not undermine the exercise. It simply marks the precision appropriate to a secondary quantitative reading rather than to a raw-data re-estimation.

The original descriptive results also provide outcome context that is easy to overlook if one focuses only on model metrics. Patients with leak had markedly longer hospital stays, more frequent early intensive care admission, and higher in-hospital mortality. Those differences do not prove that an early classifier would change outcomes, but they do explain why even moderate enrichment can matter. In a complication with substantial downstream burden, a model need not be perfect to warrant interest. It must instead demonstrate that it can isolate a subset in which the event rate is meaningfully higher than baseline and in which the cost of additional review is not absurdly inflated. The empirical question is therefore less absolute than perfect prediction and more relational: relative to the base rate, how sharply can the first postoperative day concentrate concern?

The ranked predictor list reported in the source study also gives this paper a substantive foothold. Postoperative white blood cell count, intensive care unit admission, postoperative lactate, surgery type, and preoperative albumin were the most influential variables [9]. This ordering is notable because it mixes dynamic postoperative response with static procedural and nutritional context. The result suggests that early leak classification is neither purely an acute physiologic detection task nor purely a preoperative risk stratification exercise. It is a hybrid. The immediate host response seems to supply the strongest discriminative gradient, while procedure type and nutritional reserve shape that gradients background. The sections

that follow use the published performance metrics and this predictor ordering to reconstruct the empirical topology of the problem more explicitly.

3. Modeling Strategy and Error Decomposition

The central methodological move in this paper is simple: reported test-set performance is converted into a set of derived operational quantities. Because prevalence, test-set size, precision, and recall are known, one can approximate the confusion structure of each model at its reported operating point. This is not a substitute for patient-level re-estimation, but it is sufficient to expose how differently the models allocate attention. Let n denote the total number of test admissions, n_1 the number of leak cases, n_0 the number of non-leak cases, p_m the precision of model m, and r_m its recall. The implied counts satisfy

$$\begin{aligned} n &= n_1 + n_0 \\ \widehat{TP}_m &= r_m n_1 \\ \widehat{FN}_m &= n_1 - \widehat{TP}_m \\ \widehat{FP}_m &= \widehat{TP}_m \left(\frac{1 - p_m}{p_m} \right) \\ \widehat{TN}_m &= n_0 - \widehat{FP}_m. \end{aligned} \tag{3.1}$$

Because the published precision and recall values are rounded, the resulting counts are approximate and need not be integers. Their value lies in revealing comparative scale. A model with recall 0.41 and precision 0.43 in a test set of 49 positives and 621 negatives is operating in a very different regime from a model with recall 0.71 and precision 0.12, even if both might be described casually as moderately sensitive. The equations above make that difference concrete by translating summary metrics into implied case volumes.

From these counts, several additional descriptors follow naturally. The alert rate is the proportion of all admissions labeled positive:

$$a_m = \frac{\widehat{TP}_m + \widehat{FP}_m}{n}. \tag{3.2}$$

This is a clinically intuitive quantity because it approximates how frequently the model would demand extra attention if all positive predictions triggered review. In a sparse-outcome setting, alert rate can differ sharply across models with superficially similar ranking performance. A second useful quantity is event enrichment, defined as the ratio of precision to baseline prevalence $\pi = n_1/n$:

$$\kappa_m = \frac{p_m}{\pi}. \tag{3.3}$$

If $\kappa_m = 1$, the models positive set is no more event-dense than random selection [10]. If $\kappa_m = 5$, the flagged subset carries five times the baseline event rate. This number is especially revealing in surgical prediction because it measures concentration directly. A model may miss

many cases and still have high enrichment if the cases it does flag are packed into a small, dense subset. Conversely, a model with high recall but low enrichment may simply be casting a wide net.

A third quantity is the false-alarm ratio, the number of false positives per true positive detected:

$$\phi_m = \frac{\widehat{FP}_m}{\widehat{TP}_m} = \frac{1 - p_m}{p_m}. \tag{3.4}$$

This ratio is not identical to false-positive rate; it speaks to the local quality of the positive set rather than to its relation to the negative class as a whole. In practice, it approximates how many non-events must be investigated or tracked for each event found. When discussing rare surgical complications, this ratio often communicates burden more directly than specificity because it keeps the numerator and denominator tied to the clinicians experience of review.

A fourth descriptor is miss-to-capture burden, defined as the number of leak cases missed for each leak case detected:

$$\mu_m = \frac{\widehat{FN}_m}{\widehat{TP}_m} = \frac{1 - r_m}{r_m}. \tag{3.5}$$

This quantity is helpful because it restores asymmetry to recall. The difference between 0.41 and 0.49 recall may appear modest on paper, but in miss-to-capture terms it means moving from roughly 1.44 missed leaks per detected leak to about 1.04. That is a substantial change in the balance between discovered and undiscovered event mass. It also clarifies why improving recall at stable precision can matter disproportionately when the event is clinically serious.

The paper also uses a prevalence-standardized version of the test set to express burden per 1,000 comparable admissions. If the observed prevalence is $\pi = 0.076$, then 1,000 admissions contain approximately 76 leak cases and 924 non-leak cases. For each model, one can therefore estimate

$$\begin{aligned} \widetilde{TP}_m &= 1000 \pi r_m \\ \widetilde{FP}_m &= 1000(1 - \pi) \left(\frac{\widehat{FP}_m}{n_0} \right) \\ \widetilde{A}_m &= \widetilde{TP}_m + \widetilde{FP}_m. \end{aligned} \tag{3.6}$$

This translation into a common operating scale is useful because hospital teams think in service volumes rather than in abstract fractions. Saying that one model would generate approximately 71 alerts per 1,000 cases while another would generate approximately 163 makes the practical contrast more vivid than quoting precision and recall alone [11]. The same move also allows marginal comparisons: how many extra true cases are recovered per additional 100 alerts when moving from one model family to another?

Probability quality is treated through the Brier score, but the present analysis interprets it relative to a trivial baseline. If one were to predict the prevalence π for every admission,

the Brier score would equal $\pi(1-\pi)$. For $\pi = 0.076$, this constant-probability baseline is approximately 0.070. Thus, any model with Brier score 0.059 is improving on a prevalence-only forecast, whereas a model with Brier score 0.415 is not merely underperforming; it is issuing probability estimates whose squared-error burden is several-fold worse than a constant guess. This contrast matters because it separates ranking competence from probability usefulness. A model may still order patients tolerably while assigning probabilities that are too distorted to communicate or threshold reliably.

The secondary analysis also introduces a simple concentration efficiency score:

$$\eta_m = \frac{\widehat{TP}_m / n_1}{a_m} = \frac{r_m}{a_m}. \tag{3.7}$$

This quantity measures how much positive-class recovery is achieved per unit alert coverage of the full cohort. It is not a standard metric and is not proposed as a replacement for established summaries. Its purpose is descriptive. In a low-prevalence problem, a model with higher η_m is packing recovered events into a smaller share of the population. That notion aligns closely with selective surveillance. A classifier with modest recall can still have favorable η_m if it retrieves its true positives with unusual compactness.

Beyond scalar summaries, the paper treats the ranked predictor list as evidence about signal placement within the feature space. Let the observed predictor vector be partitioned into baseline context b_i, procedural context s_i, and early postoperative response q_i. A general nonlinear score can be expressed abstractly as

$$\begin{aligned} \ell_i = \alpha &+ f_b(b_i) + f_s(s_i) + f_q(q_i) \\ &+ f_{bs}(b_i, s_i) + f_{bq}(b_i, q_i) + f_{sq}(s_i, q_i). \end{aligned} \tag{3.8}$$

The dominant variables reported in the source study suggest that the gradient of ℓ_i with respect to q_i is steeper than the gradients with respect to most elements of b_i, while certain elements of s_i and b_i, notably surgery type and albumin, retain meaningful influence [12]. Interpreted empirically rather than causally, this means that early postoperative response variables concentrate more classification signal than many static covariates, but that their interpretation is sharpened by procedural and nutritional context. This decomposition will matter later when discussing why nonlinear models outperform additive baselines only modestly in AUC but much more noticeably in enrichment and false-alarm control.

The final analytic decision concerns approximation discipline. Because patient-level probability outputs are unavailable, one cannot reconstruct full receiver operating curves, calibration plots, or threshold sweeps. The paper therefore restricts itself to quantities that are logically identifiable from the published summaries. This constraint is not a weakness so much as a guardrail. It prevents unsupported elaboration while still allowing a surprisingly rich empirical reading. By focusing on what precision, recall, prevalence, and Brier score jointly imply, the analysis turns a familiar performance table into a more interpretable map of selective case finding.

Table 1. Cohort Summary Statistics :contentReference[oaicite:0]index=0

Metric	Value	Notes
Total admissions	3331	colorectal resections
Patients	2847	unique individuals
Leak prevalence	7.6%	30-day outcome
Test set size	670	held-out
Positive cases	49	leaks

Table 2. Model Discrimination and Calibration

Model	AUC	Brier	PR-AUC
Logistic Regression	0.738	0.059	–
Naive Bayes	0.699	0.415	–
Grad. Boosting	0.754	0.059	–
Target Encoded GB	0.764	0.059	–
Bagged GB	0.772	0.059	–
Ensemble	0.762	0.059	best

4. Empirical Results

The reported models occupy a narrow band in conventional discrimination but a much wider band in operational character. Logistic regression achieved AUC 0.738, naive Bayes 0.699, gradient boosting 0.754, gradient boosting with target encoding 0.764, bagged gradient boosting 0.772, and the three-way ensemble 0.762. Read casually, these values might suggest a modest progression from additive linearity toward nonlinear ensembling, with the strongest tree-based model posting only a limited gain over logistic regression. Yet the same set of models looks quite different when one considers probability error and positive-set composition. Logistic regression and all four boosting-based variants share Brier scores of approximately 0.059, which is better than the prevalence-only baseline of roughly 0.070. Naive Bayes, by contrast, has a Brier score of 0.415, a value so poor that its probability outputs can scarcely be regarded as useful in absolute terms. This already indicates that the primary divide is not only between weaker and stronger ranking. It is between models whose numerical risk estimates remain within a plausible scale and one whose independence assumptions badly distort probability space.

The implied confusion geometry sharpens the contrast. In the held-out test set of 49 positives and 621 negatives, logistic regression corresponds to approximately 28 true positives, 80 false positives, 21 false negatives, and 541 true negatives. Naive Bayes corresponds to about 35 true positives, 257 false positives, 14 false negatives, and 364 true negatives [13]. Plain gradient boosting yields about 25 true positives, 53 false positives, 24 false negatives, and 568 true negatives. Target-encoded boosting yields roughly 21 true positives, 41 false positives, 28 false negatives, and 580 true negatives. Bagged gradient boosting implies about 20 true positives, 27 false positives, 29 false negatives, and 594 true negatives. The three-way ensemble implies about 24 true positives, 35 false positives, 25 false negatives, and 586 true negatives. These are rounded reconstructions, but they capture the central empirical fact: the models differ far more in how they distribute false positives than in how they distribute true positives.

This can be expressed another way. Logistic regression recovers more positives than the bagged gradient boosting model, but it does so by tripling the false-positive volume. Naive

Table 3. Precision and Recall Comparison

Model	Precision	Recall	F1
Logistic Regression	0.43	0.57	–
Naive Bayes	0.12	0.71	–
Grad. Boosting	0.32	0.51	–
Target Encoded GB	0.34	0.43	–
Bagged GB	0.43	0.41	–
Ensemble	0.41	0.49	best

Table 4. Approximate Confusion Structure (Test Set)

Model	TP	FP	FN
Logistic Regression	28	80	21
Naive Bayes	35	257	14
Grad. Boosting	25	53	24
Target Encoded GB	21	41	28
Bagged GB	20	27	29
Ensemble	24	35	25

Bayes recovers still more positives, yet it does so at a scale of false alarms that overwhelms any plausible selective surveillance use. The boosting family, especially the bagged and target-encoded forms, moves in the opposite direction. It gives up some recall in exchange for much tighter concentration of the positive set. The ensemble occupies an intermediate position, recovering more positives than the bagged model while keeping false alarms far below the logistic baseline. In other words, the empirical spread among models is driven less by who can identify an extra handful of leak cases and more by how diffusely or compactly each model must cast its net to do so.

Event enrichment makes this even clearer. With prevalence at 7.6%, a positive prediction from logistic regression has about 3.42 times the baseline event density. Naive Bayes raises density only about 1.58-fold above baseline despite its broader recall. Plain gradient boosting yields approximately 4.21-fold enrichment, target-encoded boosting about 4.47-fold, bagged gradient boosting about 5.66-fold, and the three-way ensemble about 5.39-fold. These numbers do not describe absolute correctness; they describe packing efficiency. Bagged gradient boosting is not the most sensitive model, but its positive set is the densest [14]. The ensemble is slightly less dense but retrieves more events. Logistic regression sits lower on enrichment because its additional captures are purchased through broader dispersion into the negative class. Naive Bayes is weakest in this regard. Its detected cases are diluted within such a large predicted-positive set that the model scarcely compresses signal relative to random prevalence.

The false-alarm ratio tells a similar story. Logistic regression produces roughly 2.85 false positives for every true positive detected. Naive Bayes produces about 7.33. Plain gradient boosting produces approximately 2.13, target-encoded boosting 1.94, bagged gradient boosting 1.33, and the three-way ensemble 1.44. These ratios matter because they map directly onto review burden. A model with ratio 1.33 is not free of error, but it asks clinicians to examine roughly one to one-and-a-third non-events for every event retrieved. A model with ratio 7.33 asks for an entirely different style of use, more akin to broad screening than to selective escalation. In a postoperative complication where clinical attention is finite, that

Table 5. Event Enrichment and Alert Rate

Model	Enrichment	Alert Rate	Ratio
Logistic Regression	3.42	0.16	moderate
Naive Bayes	1.58	0.44	high
Grad. Boosting	4.21	0.12	improved
Target Encoded GB	4.47	0.09	strong
Bagged GB	5.66	0.07	highest
Ensemble	5.39	0.09	balanced

Table 6. False Alarm and Miss Ratios

Model	FP/TP	FN/TP	Interpretation
Logistic Regression	2.85	0.75	moderate burden
Naive Bayes	7.33	0.41	very high FP
Grad. Boosting	2.13	0.96	balanced
Target Encoded GB	1.94	1.33	conservative
Bagged GB	1.33	1.44	dense alerts
Ensemble	1.44	1.04	trade-off

distinction is substantial.

Miss-to-capture burden adds the other side of the ledger. Logistic regression misses roughly 0.75 leaks for every leak it detects. Naive Bayes misses only about 0.41 per detected leak, which is the natural arithmetic consequence of its higher recall. Plain gradient boosting misses about 0.96, target-encoded boosting 1.33, bagged gradient boosting 1.44, and the three-way ensemble 1.04. These values show why no single metric resolves the comparison. If the sole objective were to minimize missed leaks without regard to burden, naive Bayes would appear attractive, followed by logistic regression and the ensemble. But once one accounts for false-positive volume and probability instability, the interpretation changes sharply [15]. The ensemble becomes more attractive than logistic regression because it recovers almost as many true cases while keeping the review burden far lower. Bagged gradient boosting becomes attractive if one values concentrated alerts over raw retrieval. The important empirical lesson is not that one model wins universally. It is that the models occupy distinct operating regimes that conventional AUC reporting compresses too heavily.

The comparison between bagged gradient boosting and the three-way ensemble is particularly informative because it reveals a genuine trade-off within the stronger model family. Bagged gradient boosting posts the highest AUC and the best precision, while the ensemble posts the best precision-recall area under the curve and higher recall. The difference between them is not a matter of one model being uniformly stronger. Bagging compresses the alert surface more aggressively, yielding a denser positive subset. The ensemble relaxes this compression slightly, admitting more positives and more false positives, but in a proportion that improves event recovery relative to alert expansion. This is precisely the kind of nuance that disappears when model comparison is reduced to one scalar. The ensemble is not worse because its AUC is slightly lower. It is better if one values a different balance of retrieved cases and tolerated review burden.

The calibration contrast with naive Bayes deserves additional attention. A Brier score of 0.415 against a prevalence-only baseline of about 0.070 implies a model whose probability

Table 7. Service-Level Estimates (per 1000 Admissions)

Model	TP	FP	Alerts
Logistic Regression	43	119	163
Naive Bayes	54	383	437
Grad. Boosting	39	79	117
Target Encoded GB	33	61	94
Bagged GB	31	40	71
Ensemble	37	52	89

Table 8. Top Predictive Features

Rank	Feature	Type	Role
1	WBC (post-op)	lab	inflammation
2	ICU admission	care	intensity marker
3	Lactate (post-op)	lab	perfusion
4	Surgery type	procedural	context
5	Albumin (pre-op)	lab	reserve

estimates are severely mis-scaled. Yet naive Bayes still reaches recall 0.71. This combination is a reminder that ranking and probability quality can pull apart dramatically. The model is evidently able to identify a wide swath of admissions in which leaks are overrepresented, but the way it translates feature evidence into numerical probabilities is so poor that its score space cannot be interpreted at face value. In a practical setting, this would matter even if the model were used only for ranking [16]. Thresholds chosen on the basis of probability magnitude would be unstable, and communication of absolute risk would be misleading. The source studys inclusion of Brier score is therefore especially valuable. It prevents the model comparison from drifting into a false equivalence between broad sensitivity and meaningful risk estimation.

The variable ranking enriches the performance comparison by showing where the learnable signal is concentrated. Postoperative white blood cell count appears first, followed by intensive care unit admission and postoperative lactate. These are not subtle long-horizon markers. They are immediate reflections of inflammatory activation, physiologic strain, and acute clinical intensity. Their dominance suggests that by the end of the first postoperative day, the record already contains a strong deviation profile in at least some future leak cases. Surgery type and preoperative albumin then enter as contextual variables. One can read this combination empirically as a two-layer structure. The acute layer captures the earliest measurable departure from ordinary recovery. The contextual layer modulates how that departure should be interpreted. A given white blood cell count or lactate concentration does not carry identical meaning across all procedures or nutritional states. The nonlinear models likely gain their advantage by exploiting precisely these context-dependent inflections.

The clinical outcome contrasts reported for leak and non-leak admissions reinforce why this structure matters. Leak cases had far longer stays, greater early critical-care exposure, and substantially higher mortality. The immediate predictive problem is difficult, but the downstream burden of the event is large enough that even partial enrichment may be meaningful. A classifier that identifies only one-third to one-half of leak cases may still matter if the flagged subset is dense enough to justify more careful review. Conversely, a classifier that identifies two-thirds of cases but floods the service with low-yield false alarms may not be viable despite

Table 9. Model Trade-off Efficiency

Comparison	ΔTP	ΔAlerts	Yield (%)
Bagged → Ensemble	+6	+18	33
Ensemble → LogReg	+6	+74	8
GB → Bagged	-8	-46	–
Naive → LogReg	-11	-274	–

Table 10. Missingness and Data Characteristics

Variable	Missingness	Distribution	Note
WBC	low	right-skewed	strong signal
Creatinine	low	right-skewed	stable
Hemoglobin	low	symmetric	baseline
Albumin	moderate	left-skewed	reserve marker
BMI	moderate	mixed	documentation bias

better recall [17]. The empirical findings here, particularly the differences among the boosting family, suggest that the sharpest question is not whether early postoperative structured data contain signal. They clearly do. The sharper question is how that signal should be traded against the cost of scanning a far larger pool of non-events.

Another notable pattern is the relative modesty of the AUC gap between logistic regression and the stronger nonlinear models compared with the much larger gap in false-positive burden. This implies that many leak cases are not radically re-ranked by nonlinear modeling. Instead, the main advantage appears to lie in the tightening of the decision boundary around a more selective subset. In other words, the tree-based models may not be discovering entirely new event regions so much as trimming away negative cases that the linear baseline cannot cleanly separate because of interaction and threshold effects. That interpretation is consistent with the reported feature structure. If risk depends on elevated white blood cell count in combination with procedure type or albumin status rather than as an additive main effect alone, then a boosting model can preserve recall while discarding negatives that logistic regression continues to flag.

Taken together, these empirical results support a more differentiated reading of the underlying experiment. The original study demonstrated that early postoperative structured records can support meaningful leak stratification. The present analysis adds that the informative differences among models lie mainly in concentration, burden, and probability quality rather than in dramatic shifts of ranking ability. The event signal is real but sparse, and stronger models earn their advantage by shaping that sparse signal into a tighter subset rather than by transforming the problem into a high-recall task. That finding is clinically sober rather than triumphant. It suggests a prediction problem in which gains are possible, but only by respecting the arithmetic of rarity.

5. Threshold Geometry and Clinical Yield

The most direct way to appreciate the differences among the six models is to express them on a common service scale. Suppose 1,000 admissions resemble the test population, so that approximately 76 develop leak and 924 do not. Then logistic regression, at the reported operating point, would detect about 43 leak cases while generating roughly 119 false positives,

for a total of about 163 alerts [18]. Naive Bayes would detect about 54 leak cases but would generate about 383 false positives, producing approximately 437 alerts. Plain gradient boosting would detect about 39 leaks with about 79 false positives, or roughly 117 alerts. Target-encoded boosting would detect about 33 leaks with about 61 false positives, for about 94 alerts. Bagged gradient boosting would detect about 31 leaks with about 40 false positives, for about 71 alerts. The three-way ensemble would detect about 37 leaks with about 52 false positives, for about 89 alerts. These numbers are approximate, but the ordering is robust. The models populate distinct alert-volume regimes.

This translation exposes the essential geometry of the comparison. Naive Bayes sits in a high-recall, very high-burden corner. Logistic regression sits in a moderate-recall, moderate-to-high-burden corner. The pure boosting variants and their bagged form move steadily toward lower alert burden and higher event density. The three-way ensemble shifts slightly back toward recall, accepting a moderate expansion of the alert surface relative to the bagged model. What becomes visible here is the shape of the trade-off. The stronger nonlinear models are not producing a dramatic jump in the number of leaks found. They are changing how expensive each additional recovered leak becomes in terms of non-event review.

Marginal comparison is especially useful. Moving from bagged gradient boosting to the three-way ensemble increases expected detections from about 31 to about 37 per 1,000 admissions, a gain of roughly 6 additional leak cases. The same move increases false positives from about 40 to about 52, a rise of roughly 12, and total alerts from about 71 to about 89, a rise of about 18 [19]. This is a relatively efficient trade. The model family remains within a compact alert surface while recovering a clinically noticeable increment of true cases. By contrast, moving from the three-way ensemble to logistic regression increases expected detections from about 37 to about 43, again a gain of roughly 6 leak cases, but false positives jump from about 52 to about 119, an increase of about 67, and total alerts increase by roughly 74. That is a much less favorable exchange. The linear model recovers additional cases, but it does so by diluting the alert set far more aggressively.

The comparison between bagged gradient boosting and logistic regression is even starker. Logistic regression finds about 12 more leaks per 1,000 admissions than the bagged model, but this improvement comes with roughly 80 additional false positives and about 92 additional alerts. Whether that trade is acceptable depends on the clinical context, yet the arithmetic itself is unavoidable. To favor logistic regression over the bagged model, one must believe that the value of recovering those 12 additional leaks outweighs the cost of tracking almost another hundred flagged admissions. The secondary analysis does not dictate that answer. It clarifies what answer one would have to endorse.

These margins can be summarized through an incremental yield ratio. If model m_2 is compared with m_1, define

$$\begin{aligned} \Delta_{21}^{+} &= \widetilde{TP}_{m_2} - \widetilde{TP}_{m_1} \\ \Delta_{21}^{A} &= \widetilde{A}_{m_2} - \widetilde{A}_{m_1} \\ \psi_{21} &= 100\,\frac{\Delta_{21}^{+}}{\Delta_{21}^{A}}. \end{aligned} \tag{5.1}$$

Here ψ_{21} gives the number of extra true leak detections gained per 100 additional alerts when moving from one model to another. Between bagged gradient boosting and the three-way ensemble, ψ is roughly 33. Between the ensemble and logistic regression, ψ falls to about

8. This is a meaningful empirical distinction. It says that expanding the alert surface from the bagged model to the ensemble buys a relatively dense cluster of additional events, whereas expanding again to logistic regression buys a much more dilute set.

Another way to view the same phenomenon is through concentration efficiency. Using $\eta_m = r_m/a_m$, the bagged gradient boosting model achieves the highest value because it recovers a moderate portion of the positive class while covering the smallest share of the population [20]. The ensemble is close behind. Logistic regression falls lower because its broader alert rate erodes compactness. Naive Bayes performs worst despite strong recall, since the share of the population it flags is so large. Conceptually, η_m can be read as selective surveillance power. A model with high η_m is efficient at packing recovered events into a narrow band of admissions. This is not identical to precision, which looks only at the predicted-positive set. Efficiency includes cohort coverage, thereby penalizing expansive alerting more directly.

The threshold geometry also highlights the asymmetry between event recovery and false-alarm control in low-prevalence conditions. Because only 76 of every 1,000 comparable admissions develop leak, even a model that doubles or triples prevalence in its positive set is still identifying a subset in which non-events remain common. That is why the false-alarm ratio remains above 1 even for the best-performing model family. The arithmetic of rarity does not permit a clean separation with the current signal strength. Any interpretation of the results must therefore avoid a hidden binary fantasy in which the model either finds leaks or fails. The real output is a selective overrepresentation of leaks in a flagged subset. The degree of overrepresentation differs by model, and those differences matter, but none of the models eliminate ambiguity.

This has implications for how one reads recall itself. The difference between 31 and 43 detected leaks per 1,000 admissions may seem substantial, and in some circumstances it is. Yet what matters clinically is not merely that 12 additional leaks are detected. It is also who those 12 cases are. If the additional retrieved cases lie near the diagnostic boundary and require large expansion of the alert set, the service-level burden rises sharply [21]. If they belong to a moderately enriched fringe that can be reviewed cheaply, the trade may be worthwhile. The present analysis cannot resolve this patient-level heterogeneity, but it shows that threshold geometry is where such questions live. AUC cannot answer them.

The Brier score comparison adds an important caution. The bagged model, the ensemble, and logistic regression all share Brier values around 0.059. This suggests that once probability error is averaged over the cohort, the main practical difference among them is not global probability quality but the shape of their thresholded operating points. Put differently, the stronger models may not be much better than logistic regression at minimizing average squared probability error, but they are better at concentrating positive predictions into a denser subset. This is why threshold geometry deserves separate treatment. A clinician selecting a model for selective review may rationally prefer one with similar Brier score but markedly lower false-alarm burden. The relevant improvement lies in positive-set composition rather than in global calibration loss.

The threshold discussion also clarifies how the original studys precision-recall findings should be interpreted. The three-way ensemble achieved the best precision-recall area under the curve even though bagged gradient boosting had the highest point precision and AUC. That pattern is consistent with a model whose favorable behavior extends across a broader range of thresholds rather than being concentrated at a single highly selective point. This makes the ensemble appealing if one expects operating thresholds to change across settings or if one values robustness in the precision-recall plane more than peak concentration at one chosen

cutpoint. Bagged gradient boosting, by contrast, appears particularly attractive if the goal is to operate in a tightly controlled alert regime.

None of this implies that the reported operating points are themselves ready-made clinical thresholds. They were chosen to optimize F1 within the experiment, and F1 is only one way to balance precision and recall. The value of the present analysis is different. It shows what those chosen points imply about the burden structure of the models and thereby illustrates why threshold choice cannot be treated as an afterthought [22]. In leak prediction, where prevalence is low and the cost of each false positive is not negligible, small movements along the threshold axis can have large operational consequences. The source study supplied enough information to reveal that fact even without a full threshold sweep. Once the results are translated into service-scale counts, the importance of operating geometry becomes unmistakable.

6. Sensitivity, Missingness, and Stability

Prediction from routine perioperative records is always partially prediction from the habits of record generation. This is particularly clear in the source experiment because several of the strongest variables are measured in environments where data density is itself clinically informative. A patient admitted to intensive care shortly after surgery has, by definition, entered a more intensively observed state than a patient transferred directly to a routine ward. A patient with several early laboratory values available may already have triggered more concern than a patient with sparse measurement. This does not make the model invalid. It means that part of the signal resides in the observation pattern. The stability of the classifier is therefore linked not only to physiology, but also to missingness structure and care intensity.

The original data-quality description indicates that white blood cell count, creatinine, and hemoglobin had relatively low missingness, while features dependent on longer windows or outpatient documentation, such as albumin and body mass index, were missing more frequently. This asymmetry is exactly what one would expect in a hospital dataset that includes urgent admissions and critical-care trajectories. The modeling consequence is subtle. Variables with low missingness tend to behave as broad carriers of common postoperative information. Variables with high missingness often become conditionally informative because their very absence marks a different clinical path. A model that includes imputation indicators can therefore extract signal from both value and nonavailability. In matrix form, if x_i is the original feature vector and d_i is a diagonal matrix of observability indicators, then the effective input is

$$\begin{aligned} x_i^* &= d_i x_i + (I - d_i)\bar{x} \\ z_i &= \begin{bmatrix} x_i^* \\ \operatorname{diag}(d_i) \end{bmatrix}, \end{aligned} \tag{6.1}$$

where $\bar{x}$ is an imputed reference vector and z_i is the augmented predictor state. The imputation indicator component permits the classifier to treat missingness as structure rather than as pure nuisance [23].

This matters empirically because the high-ranking predictors include both routinely measured acute variables and more selectively available baseline markers. Preoperative albumin, for example, is clinically meaningful as an indicator of nutritional reserve, but in a critical-care-derived dataset its availability is itself patterned. A patient who arrives for emergency surgery

without prior optimization may have no recent albumin value, whereas a planned elective case may have a fuller preoperative panel. If the model uses both the imputed albumin value and the indicator marking whether it was observed, then it is learning in a space where nutritional deficit and pathway irregularity can interact. That can improve predictive power, but it also means the resulting classifier is sensitive to documentation culture. A hospital that measures albumin more uniformly could change the models operating characteristics even if patient biology were unchanged.

The same logic applies to body mass index and other baseline descriptors. Missingness in these features is not uniformly random because anthropometric completeness depends on where and how patients enter the system. The original papers use of unknown categories and indicator variables is therefore methodologically coherent, but it highlights the classifiers dependence on the measurement regime. Stability of performance across institutions or even across time within one institution would depend partly on the persistence of these missingness patterns. The secondary analysis cannot test that directly, yet the model behavior already suggests that some of the strongest discrimination arises from features whose observability is tightly coupled to clinical workflow.

Distributional shape further affects stability. The source study reports right skew in white blood cell count, lactate, and creatinine, with high-end values concentrated among leak cases, and left skew in albumin with low values overrepresented among positives. These are not benign departures from normality. They create sharp tail regions in which additive models may struggle unless the relevant thresholds and interactions are hand specified. Tree-based learners thrive in precisely this setting because they can place splits near extreme concentrations without assuming linear response. The modest AUC advantage of boosting over logistic regression can therefore understate the structural importance of these tails. What the nonlinear models appear to gain is not a wholesale reordering of all admissions, but a cleaner exclusion of negatives that share average values but not the same pattern of threshold crossing across correlated variables [24].

One can formalize this intuition by writing the score as a sum over region indicators rather than as a purely linear map. For a tree-based ensemble,

$$f(x_i) = \sum_{t=1}^{T} \sum_{r=1}^{R_t} \gamma_{tr} \mathbb{I}(x_i \in \mathcal{R}_{tr}), \tag{6.2}$$

where each $\mathcal{R}_{tr}$ is a rectangular partition of feature space. When the positive class concentrates in irregular combinations such as high white blood cell count with elevated lactate and low albumin, a partition-based model can isolate that region while leaving large average-looking portions of the negative class untouched. Logistic regression cannot do this unless such interactions are explicitly encoded. The practical consequence is improved stability of the positive set: the tree model can maintain enrichment without expanding broadly into ambiguous regions.

The behavior of naive Bayes is instructive in the opposite direction. Its strong recall and disastrous Brier score indicate that the model is sensitive to the same informative variables but unable to preserve a sensible probability scale under heavy feature dependence. White blood cell count, lactate, albumin, and intensive care unit admission are not independent evidence streams. They are connected through operative severity, acute inflammation, and postoperative instability. A classifier that multiplies their likelihood contributions as though they were conditionally separate will over-amplify some patterns and distort absolute risk.

That is exactly what the probability error suggests. The model is not blind. It is unstable in how it accumulates evidence. This is a useful empirical reminder that more sensitivity does not necessarily mean more reliable synthesis.

The original experiments simple imputation scheme also deserves comment from a stability perspective. Median imputation is deliberately conservative. It avoids the variance introduced by more elaborate imputation models and keeps the pipeline transparent. At the same time, it compresses uncertainty. When a missing continuous feature is replaced by a typical value, the model receives a pseudo-observation that must be corrected by the missingness indicator if the absence is informative [25]. This can work reasonably well, but it places substantial weight on the models ability to learn interactions between value and observability. Tree-based models often handle this gracefully because they can split on the indicator itself or on the imputed value conditional on the indicator. Linear models can also exploit the augmentation, but only additively unless interactions are introduced. That may be one reason the tree family shows better positive-set concentration without dramatic global metric superiority.

The stability question is not limited to institutions. Even within one center, case mix and ordering practice can drift. If postoperative lactate becomes more routinely measured over time, the signal content of missingness changes. If critical-care admission criteria shift, the meaning of that variable changes as well. A model trained on a period with one measurement regime may then behave differently later even if the biological incidence of leak stays similar. This possibility is especially relevant when strong predictors are not purely biological but partly organizational. The empirical findings of the source study, read through the present analysis, suggest that some of the apparent strength of the early postoperative signal may arise from exactly this fusion of physiology and care intensity. That does not diminish the value of the model; it does mean that stability should never be assumed from internal performance alone.

The papers broader implication is that missingness and measurement structure are not side conditions. They are part of the actual prediction problem. A classifier that performs well on such data has learned not just from bodies, but from the hospitals pattern of seeing those bodies. Any account of model behavior that ignores this point risks misdescribing what the prediction really encodes. The importance of white blood cell count, lactate, albumin, and intensive care unit admission is clinically plausible, yet their combined value is inseparable from when they are measured, how often they are missing, and in what context they acquire meaning. The resulting model is best understood as a detector of early postoperative deviation as documented in routine care, not as a purely pathophysiologic instrument detached from the observation system [26].

7. Interpretation of Predictive Signal

The ranked predictor profile reported in the original experiment offers a compact empirical answer to a question that often remains vague in surgical prediction studies: where, exactly, is the model finding useful information? The prominence of postoperative white blood cell count, intensive care unit admission, and postoperative lactate indicates that the strongest gradient in the first postoperative day lies in immediate systemic response. This does not mean the model is detecting fully developed leak. Most leaks declare themselves later, and the outcome window extends well beyond the first day. It means that future leak cases already tend to occupy a different region of early postoperative physiology and care intensity. The difference may be small at the level of individual patients, but aggregated across the cohort it becomes strong enough to shape classification.

One way to conceptualize this is to separate the feature space into a baseline reserve component and an acute-response component. Let b_i contain nutritional and demographic information, let s_i encode surgery type and related procedural context, and let q_i collect immediate postoperative measurements. A flexible score can then be written as

$$\begin{aligned}\ell_i = {} & \beta_b^\top b_i + \beta_s^\top s_i + \beta_q^\top q_i \\ & + b_i^\top \Gamma_{bq} q_i + s_i^\top \Gamma_{sq} q_i + \varepsilon_i. \end{aligned} \tag{7.1}$$

In this representation, the dominance of white blood cell count and lactate implies that β_q and the interaction blocks involving q_i carry a substantial share of the signal. The continued importance of surgery type and albumin indicates that the contextual blocks are not negligible. Elevated inflammatory markers are more informative when interpreted against a background of high-risk procedure profile or depleted nutritional reserve. The nonlinear learners likely derive their advantage by representing these context-sensitive slopes more effectively than a purely additive linear baseline.

This interpretation helps explain why the studys strongest variables are not the ones most closely tied to long-term patient identity. Age and body mass index contributed, but they were less influential than early physiologic and organizational measures. That pattern is empirically coherent. Baseline covariates distribute risk over a broad horizon, but they may not create sharp local separation once the operation is complete. By contrast, the first postoperative day contains concentrated information about whether the immediate recovery trajectory is ordinary or perturbed. Inflammation, perfusion stress, metabolic disturbance, and intensive monitoring together form a more selective signature than generic demographic risk. The classifier is therefore reading a postoperative deviation surface rather than merely aggregating preoperative vulnerability [27].

The presence of surgery type among the top variables is also important. It suggests that the model does not interpret early physiology in a vacuum. Different colorectal operations carry different technical demands, contamination patterns, and anastomotic contexts. A lactate elevation after one type of resection need not mean what it means after another. Procedure category therefore acts as a geometric conditioning variable, tilting the decision surface so that the same physiologic measurement can map to different risk increments. This is one reason additive models may underperform even when their overall discrimination remains respectable. Without explicit interaction terms, they tend to apply a common slope to all procedural contexts. The boosting family can bend that slope locally.

Preoperative albumin introduces a different type of conditioning. Unlike the acute variables, albumin is a slower reflection of host reserve, nutritional adequacy, and chronic illness burden. Its presence among the top predictors suggests that immediate postoperative response alone is not enough. Patients who begin from a depleted baseline may move into the high-risk region of the postoperative feature space more readily or may do so at lower absolute inflammatory thresholds. In geometric terms, albumin can be thought of as shifting the origin from which postoperative deviation is measured. A white blood cell count or lactate value that is only moderately concerning in one nutritional context may be more informative in another. Again, a flexible model can encode such modulation more naturally than a linear one.

Intensive care unit admission is perhaps the most revealing variable of all because it is not purely a laboratory observation or a fixed patient trait. It is a decision-dependent state marker. It reflects operative complexity, hemodynamic instability, anticipated monitoring

needs, or institutional recovery practice [28]. Its high rank indicates that the earliest informative signal for future leak is partly organizational. Cases destined for leak are, on average, already causing enough concern or enough objective instability to alter immediate postoperative placement. This is not a flaw in the predictor. It is evidence that the first postoperative day contains a composite risk imprint in which physiology and care intensity are intertwined. From a statistical point of view, such a variable can be highly predictive precisely because it summarizes many unmeasured details. From an interpretive point of view, it reminds us that the classifier is reading the hospitals response to the patient as well as the patient directly.

The variable profile also clarifies the likely limits of preoperative-only prediction. If the strongest discriminators arise after surgery, then a model restricted to preoperative data would probably lose much of the signal concentration observed here. That does not make preoperative assessment unimportant. Albumin and surgery type still matter, and baseline risk remains clinically relevant. But the empirical pattern indicates that the problem becomes substantially more learnable once the body has already produced an early response to the operation. In practical terms, the model resembles an early surveillance instrument more than a purely anticipatory screen. It captures deviation from expected postoperative behavior rather than forecasting leak solely from static predisposition.

This reading is reinforced by the modest gap between logistic regression and the stronger nonlinear models. If leak risk were driven primarily by a small number of strong additive covariates, the linear baseline might be closer to optimal and the gain from boosting would be minimal in both AUC and concentration. Instead, the nonlinear family achieves much better enrichment and false-positive control, implying that the decisive structure lies in interaction and thresholding. White blood cell count may become especially informative only beyond a certain range and only in combination with other features. Lactate may matter most when paired with intensive care admission or certain procedures. Albumin may act chiefly by modifying the slope of postoperative markers [29]. These are precisely the kinds of feature relationships that shallow boosted trees absorb naturally.

It is also worth noticing what the model is not relying on most heavily. The leading predictors are not exotic biomarkers, genomic profiles, or free-text embeddings. They are ordinary structured measures already present in routine hospital data. Empirically, this means that the learnable signal for early postoperative leak is not hidden in an inaccessible corner of the record. It is visible, though imperfectly, in standard perioperative documentation. That should not be romanticized. The signal remains weak enough that none of the models approaches high-certainty classification. But it does mean that improved prediction need not wait for entirely new data modalities before making analytic progress. Much of the difficulty lies not in absence of measurable information, but in the scarcity of the event and the heterogeneous way ordinary measurements combine.

The broader interpretation is therefore restrained. The source study does not show that leak can be identified cleanly within 24 hours. It shows that the first postoperative day contains a distorted but meaningful echo of the processes that will later become clinically manifest leak. The leading variables suggest that this echo is strongest in acute inflammatory and physiologic response, sharpened by procedural category and host reserve, and amplified by organizational reactions such as intensive care admission. The nonlinear models succeed to the extent that they carve a dense alert subset out of this mixed signal. The empirical contribution of the present paper is to show that this carving is the main action. The classifiers are not creating certainty. They are compressing an early, noisy, clinically entangled signature into a smaller region of the cohort where leak is substantially more common than baseline [30].

8. Conclusion

This paper has examined early postoperative anastomotic leak prediction from a different empirical angle. Instead of centering the analysis on headline discrimination alone, it treated the reported test-set performance of six classifiers as a map of how event signal is distributed, concentrated, and diluted across the cohort. That perspective changes the reading of the experiment in useful ways. The most consequential differences among models are not dramatic shifts in AUC. They are changes in alert density, false-positive volume, enrichment above baseline prevalence, and the burden required to recover additional leak cases. Once those quantities are made explicit, the comparison among model families becomes more clinically intelligible.

The strongest lesson is that the first postoperative day contains a real but limited concentration of leak-related information. Dynamic inflammatory and physiologic variables, especially white blood cell count and lactate, carry more discriminative weight than many static baseline factors, while surgery type and albumin provide contextual shaping of that acute signal. The resulting predictive problem is therefore neither purely preoperative nor purely mechanistic. It is a classification task built on the early geometry of postoperative deviation as captured in ordinary hospital records.

The secondary quantitative analysis also shows why low-prevalence complications demand more than discrimination summaries. A model that retrieves more events at the cost of a much larger false-positive field may not be preferable to a model with slightly lower recall but much denser positive predictions. Bagged gradient boosting and the three-way ensemble illustrate this clearly. Their main strength lies in packing true events into a relatively compact alert surface. Logistic regression recovers more events only by broadening that surface substantially, while naive Bayes couples broad recall to severe probability instability.

A sober conclusion follows. Early structured perioperative data are informative enough to support selective leak risk stratification, but the useful scientific question is not simply whether signal exists. It is how tightly that signal can be compressed, how stable the resulting probabilities remain, and how much non-event burden must be accepted for each additional event retrieved. Those are the empirical terms in which surgical prediction studies are most likely to become genuinely interpretable [31].

References

[1] Nadir Adam, Shahrzad Rahbar, and Ruby Skinner. “Outcomes of Colostomy Reversal in a Public Safety Net Hospital: The End or Beginning of a New Problem?” In: *The American surgeon* 81.10 (Oct. 1, 2015), pp. 1084–1087. DOI: 10.1177/000313481508101034.

[2] Felix J Hüttner, Solveig Tenckhoff, Katrin Jensen, Lorenz Uhlmann, Yakup Kulu, M.W. Büchler, Markus K. Diener, and Alexis Ulrich. “Metaanalysis of reconstruction techniques after low anterior resection for rectal cancer”. In: *The British journal of surgery* 102.7 (Mar. 31, 2015), pp. 735–745. DOI: 10.1002/bjs.9782.

[3] Sarath Sujatha-Bhaskar, Mehraneh D. Jafari, Mark H. Hanna, Christina Y. Koh, Colette S. Inaba, Steven Mills, Joseph C. Carmichael, Ninh T. Nguyen, Michael J. Stamos, and Alessio Pigazzi. “An endoscopic mucosal grading system is predictive of leak in stapled rectal anastomoses.” In: *Surgical endoscopy* 32.4 (Sept. 15, 2017), pp. 1769–1775. DOI: 10.1007/s00464-017-5860-y.

[4] Chengkui Liu, Xiaoqing Li, and Qingfeng Wang. "Postoperative complications observed with robotic versus laparoscopic surgery for the treatment of rectal cancer: An updated meta-analysis of recently published studies." In: *Medicine* 100.36 (Sept. 10, 2021), e27158–. DOI: 10.1097/md.0000000000027158.

[5] Affan Adib Sungkar and Budhi Ida Bagus. "The Short-term Outcome of Early Oral Feeding as Recovery Protocols on Resectable Gastric Cancer". In: *Open Access Macedonian Journal of Medical Sciences* 10.B (Jan. 2, 2022), pp. 78–81.

[6] Binesh Sadanandan. "Data-Driven Risk Stratification for Anastomotic Leak: A Proof-of-Concept Study Using Electronic Health Records". In: *Journal of Data Science, Predictive Analytics, and Big Data Applications* 9.6 (2024), pp. 1–27.

[7] Keegan Guidolin, Andrea Marie Covelli, Tyler R. Chesney, Arman Draginov, Sami A Chadi, and Fayez A. Quereshy. "Apical lymphadenectomy during low ligation of the IMA during rectosigmoid resection for cancer." In: *Surgery open science* 5 (June 23, 2021), pp. 1–5. DOI: 10.1016/j.sopen.2021.06.002.

[8] Francesco Vito Mandarino et al. "Endoscopic Vacuum Therapy (EVT) versus Self-Expandable Metal Stent (SEMS) for Anastomotic Leaks after Upper Gastrointestinal Surgery: Systematic Review and Meta-Analysis." In: *Life (Basel, Switzerland)* 13.2 (Jan. 19, 2023), pp. 287–287. DOI: 10.3390/life13020287.

[9] Dumitru Radulescu et al. "Neutrophil/Lymphocyte Ratio as Predictor of Anastomotic Leak after Gastric Cancer Surgery". In: *Diagnostics (Basel, Switzerland)* 10.10 (Oct. 9, 2020), pp. 799–. DOI: 10.3390/diagnostics10100799.

[10] Byung Jo Choi, Won Jun Jeong, Say-June Kim, and Sang Chul Lee. "Single-port laparoscopic surgery for sigmoid volvulus." In: *World journal of gastroenterology* 21.8 (Feb. 28, 2015), pp. 2381–2386. DOI: 10.3748/wjg.v21.i8.2381.

[11] Carissa L. Garey, Carrie A. Laituri, Adam J. Kaye, Daniel J. Ostlie, Charles L. Snyder, George W. Holcomb, and Shawn D. St. Peter. "Esophageal perforation in children: a review of one institution's experience." In: *The Journal of surgical research* 164.1 (June 13, 2010), pp. 13–17. DOI: 10.1016/j.jss.2010.05.049.

[12] Alaa Alakkari, Ruchit Sood, Simon M Everett, Bjorn J. Rembacken, Jeremy D. Hayden, Abeezar I. Sarela, and Noor Mohammed. "First UK experience of endoscopic vacuum therapy for the management of oesophageal perforations and postoperative leaks". In: *Frontline gastroenterology* 10.2 (Jan. 12, 2019), pp. 200–203. DOI: 10.1136/flgastro-2018-101138.

[13] Christopher M. Dodgion, Bridget A. Neville, Stuart R. Lipsitz, Yue Yung Hu, Deborah Schrag, Elizabeth M. Breen, and Caprice C. Greenberg. "Do older Americans undergo stoma reversal following low anterior resection for rectal cancer". In: *The Journal of surgical research* 183.1 (Dec. 21, 2012), pp. 238–245. DOI: 10.1016/j.jss.2012.11.057.

[14] Erika Andreatta, Alberto Buogo, Emanuele Asti, Sara Boveri, and Luigi Bonavina. "Comparison of pleural drain amylase and serum C-reactive protein for early detection of intrathoracic esophago-gastric anastomotic leaks." In: *Langenbeck's archives of surgery* 407.7 (May 17, 2022), pp. 2715–2724. DOI: 10.1007/s00423-022-02550-4.

[15] Alaa El-Hussuna, Adriano Axel Ceccotti, T. Madsbøll, D. Straarup, and Anders Uth Ovesen. "Association of mechanical bowel preparation with oral antibiotics and anastomotic leak following left sided colorectal resection: an international, multi-centre, prospective audit". In: *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland* 20.S6 (Sept. 25, 2018), pp. 15–32. DOI: 10.1111/codi.14362.

[16] Robert J. Touloukian and John H. Seashore. "Thirty-fiveYear Institutional Experience With End-to-Side Repair for Esophageal Atresia". In: *Archives of surgery (Chicago, Ill. : 1960)* 139.4 (Apr. 1, 2004), pp. 371–374. DOI: 10.1001/archsurg.139.4.371.

[17] Daniel L. Miller, Gerald A. Helms, and William Mayfield. "Evaluation of Esophageal Anastomotic Integrity With Serial Pleural Amylase Levels". In: *The Annals of thoracic surgery* 105.1 (Nov. 27, 2017), pp. 200–206. DOI: 10.1016/j.athoracsur.2017.07.053.

[18] Paul Poirier, Marc-Andre Cornier, Theodore Mazzone, Sasha Stiles, Susan Cummings, Samuel Klein, Peter A. McCullough, Christine Ren Fielding, and Barry A. Franklin. "Bariatric Surgery and Cardiovascular Risk Factors A Scientific Statement From the American Heart Association". In: *Circulation* 123.15 (Mar. 14, 2011), pp. 1683–1701. DOI: 10.1161/cir.0b013e3182149099.

[19] Douglas B. Evans, Peter W.T. Pisters, Jeffrey E. Lee, Richard J. Bold, Chuslip Charnsangavej, Nora A. Janjan, Robert A. Wolff, and James L. Abbruzzese. "Preoperative chemoradiation strategies for localized adenocarcinoma of the pancreas". In: *Journal of hepato-biliary-pancreatic surgery* 5.3 (Nov. 16, 1998), pp. 242–250. DOI: 10.1007/s005340050041.

[20] Moran Slavin, Avigayil Goldstein, Barak Raguan, Yaron Rudnicki, Shmuel Avital, and Ian White. "Postoperative CRP Levels Can Rule out Anastomotic Leaks in Crohn's Disease Patients." In: *Journal of personalized medicine* 12.1 (Jan. 5, 2022), pp. 54–54. DOI: 10.3390/jpm12010054.

[21] Chun Deng, Yang Liu, Zhen-Yu Zhang, Heng-Duo Qi, Zhi Guo, Xu Zhao, and Xiao-Jun Li. "How to examine anastomotic integrity intraoperatively in totally laparoscopic radical gastrectomy? Methylene blue testing prevents technical defect-related anastomotic leaks." In: *World journal of gastrointestinal surgery* 14.4 (Apr. 27, 2022), pp. 315–328. DOI: 10.4240/wjgs.v14.i4.315.

[22] Rajesh C. Mistry, R. Vijayabhaskar, George Karimundackal, Sabita Jiwnani, and C.S. Pramesh. "Effect of short-term vs prolonged nasogastric decompression on major postesophagectomy complications: a parallel-group, randomized trial." In: *Archives of surgery (Chicago, Ill. : 1960)* 147.8 (Aug. 1, 2012), pp. 747–751. DOI: 10.1001/archsurg.2012.1008.

[23] Changjiang Qin, Xue-Qun Ren, Kaiwu Xu, Zhihui Chen, Yulong He, and Xinming Song. "Does Preoperative Radio(chemo)therapy Increase Anastomotic Leakage in Rectal Cancer Surgery? A Meta-Analysis of Randomized Controlled Trials". In: *Gastroenterology research and practice* 2014 (Nov. 12, 2014), pp. 910956–910956. DOI: 10.1155/2014/910956.

[24] Benjamin D. Shogan et al. "Alterations of the Rectal Microbiome Are Associated with the Development of Postoperative Ileus in Patients Undergoing Colorectal Surgery". In: *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract* 24.7 (Apr. 22, 2020), pp. 1663–1672. DOI: 10.1007/s11605-020-04593-8.

[25] Lixiu Liu, Lihuang He, Afang Qiu, and Min Zhang. "Rapid rehabilitation effect on complications, wound infection, anastomotic leak, obstruction, and hospital re-admission for gastrointestinal surgery subjects: A meta-analysis." In: *International wound journal* 19.6 (Feb. 22, 2022), pp. 1539–1550. DOI: 10.1111/iwj.13753.

[26] Jae Y. Kim et al. "Does the Timing of Esophagectomy After Chemoradiation Affect Outcome". In: *The Annals of thoracic surgery* 93.1 (Oct. 1, 2011), pp. 207–213. DOI: 10.1016/j.athoracsur.2011.05.021.

[27] S.W.O. Ogendo. "Follow up pattern for post oesophagectomy patients at a single centre: association with peri-operative variables". In: *Annals of African Surgery* 1.1 (Sept. 8, 2009), pp. 20–28. DOI: 10.4314/aas.v1i1.45792.

[28] Dyllan Landry, Karson R Quinn, Stephen D Helmer, and Noel C Sanchez. "Is Two Better Than One? A Retrospective Study on Colorectal Surgery Outcomes Using the

Da Vinciő Dual-Console Robot." In: *Kansas journal of medicine* 15.3 (Nov. 28, 2022), pp. 418–421. DOI: 10.17161/kjm.vol15.18278.

[29] Ismail Mahmood, Fuad Mustafa, Basil Younis, Khalid Ahmed, Ayman El-Menyar, Mohammad Asim, Ammar Al-Hassani, Ruben Peralta, and Hassan Al-Thani. "Postoperative complications of intestinal anastomosis after blunt abdominal trauma." In: *European journal of trauma and emergency surgery : official publication of the European Trauma Society* 46.3 (Sept. 24, 2018), pp. 599–606. DOI: 10.1007/s00068-018-1013-9.

[30] Oliver Peacock, Peadar S Waters, Joseph C Kong, Satish K Warrier, Christopher Wakeman, Tim Eglinton, Declan G. Murphy, Alexander G. Heriot, Frank A. Frizelle, and Jacob J McCormick. "Complications and 5-year survival after radical resections which include urological organs for locally advanced and recurrent pelvic malignancies: analysis of 646 consecutive cases". In: *Techniques in coloproctology* 24.2 (Jan. 6, 2020), pp. 181–190. DOI: 10.1007/s10151-019-02141-4.

[31] Shafquat Zaman, Pratik Bhattacharya, Ali Yasen Y Mohamedahmed, Fang Yi Cheung, Kamila Rakhimova, Salomone Di Saverio, Rajeev Peravali, and Akinfemi Akingboye. "Outcomes following open versus laparoscopic multi-visceral resection for locally advanced colorectal cancer: Aăsystematic review and meta-analysis." In: *Langenbeck's archives of surgery* 408.1 (Feb. 22, 2023), pp. 98–98. DOI: 10.1007/s00423-023-02835-2.