ARTICLE

Multilevel Electronic Health Record Modeling of Anastomotic Leak Risk Across Procedure Classes and Hospital Care Settings

Ahmad Zulkifli,[1] Mohd Hafiz,[2] and Nur Azman[3]

[1]Universiti Malaysia Kelantan, Faculty of Bioengineering and Technology, Jeli Campus, Jalan Raya Jeli–Gerik, 17600 Jeli, Kelantan, Malaysia

[2]Universiti Tun Hussein Onn Malaysia, Faculty of Computer Science and Information Technology, Parit Raja Campus, Persiaran Tun Dr. Ismail, 86400 Batu Pahat, Johor, Malaysia

[3]Universiti Malaysia Pahang Al-Sultan Abdullah, Faculty of Electrical and Electronic Engineering Technology, Pekan Campus, Lebuhraya Tun Razak, 26600 Pekan, Pahang, Malaysia

Abstract

Anastomotic leak remains one of the most consequential complications after gastrointestinal surgery, yet its risk is seldom uniform across procedures, hospitals, and perioperative care environments. Aggregate estimates are clinically convenient but analytically coarse because they compress biologically distinct operations and institutionally distinct treatment settings into a single average. That averaging can obscure where risk is intrinsically high, where it is context-dependent, and where observed variation is largely a by-product of case mix. This study examined heterogeneity of leak risk across major gastrointestinal surgery classes and hospital care settings using a retrospective multicenter electronic health record cohort. Adults undergoing right colectomy, left or sigmoid colectomy, low anterior resection, subtotal or total colectomy, small-bowel resection with anastomosis, gastrectomy, or esophagectomy were included. Hospital care setting was classified as tertiary referral, regional teaching, community, or urban safety-net, with immediate postoperative disposition modeled separately. We estimated patient-level leak probability within 30 days using a hierarchical logistic framework with nonlinear effects, surgery-specific random terms, setting-specific random terms, and cross-level interaction structure. The cohort comprised 48,917 index procedures with 3,218 leaks, yielding an overall event rate of 6.6%. Crude leak incidence varied from 3.5% after right colectomy to 13.2% after esophagectomy. After risk standardization, surgery class remained the dominant source of heterogeneity, but setting-specific differences persisted for technically demanding distal colorectal and foregut procedures. Low anterior resection and esophagectomy showed the largest setting-dependent gradients, whereas right colectomy exhibited limited contextual dispersion. A transportable hierarchical model achieved stable discrimination and calibration across held-out hospitals. The findings suggest that clinically useful leak surveillance should not rely on a single pooled threshold. Procedure-sensitive and setting-aware risk estimation appears more appropriate for postoperative escalation, diagnostic timing, and resource allocation.

1. Introduction

Anastomotic leak is not a singular event with a singular risk architecture [1]. It is the downstream manifestation of several partially overlapping failure modes that include impaired perfusion, excess anastomotic tension, local contamination, technical difficulty, immunometabolic fragility, delayed recognition, and the variable intensity of postoperative observation [2]. Those mechanisms do not operate with the same strength in a stapled cervical esophagogastric anastomosis, a low pelvic colorectal anastomosis, a side-to-side ileocolic anastomosis, or a hand-sewn small-bowel reconstruction. Nor do they unfold within equivalent hospital environments [3]. A tertiary referral center with twenty-four hour interventional radiology, routine early lactate trending, dedicated colorectal pathways, and rapid access to reoperation does not merely care for a different patient mix; it can alter the timing and threshold of recognition itself. A community setting with a lower volume of distal pelvic

anastomoses may, even when delivering competent care, exhibit distinct conditional risk because surveillance cadence, escalation pathways, and rescue capacity differ in subtle but measurable ways [4]. For that reason, a pooled leak rate has limited value if the underlying objective is to support targeted surveillance and calibrated early intervention.

Clinical discussions of anastomotic leak often move between two unsatisfactory extremes [5]. One tradition emphasizes immutable procedure-level risk, treating distal pelvic location, foregut reconstruction, or emergency contamination as the core determinants and implicitly reducing hospitals to background noise. The other tradition interprets inter-hospital differences as quality gradients without adequately separating structural case-mix differences from contextual effects [6]. Both positions flatten a problem that is inherently multilevel. A technically demanding procedure may carry high baseline risk everywhere, but the amount by which that risk rises or falls across care environments is itself an estimand of interest [7]. Likewise, a hospital may appear to have a high crude leak rate because it selectively receives malnourished referral patients, yet even after adjustment, some procedures may remain unusually sensitive to local workflow, rescue intensity, or postoperative monitoring structure. A useful empirical model must therefore do more than rank patients from low to high risk [8]. It must partition risk into components attributable to patient physiology, procedure class, institutional care context, and the interaction between procedure and context [9].

Current prediction research has improved the practical feasibility of EHR-based leak modeling, but much of that literature remains organized around overall discrimination rather than structured heterogeneity. Sadanandan (2024) showed that reproducible EHR models can identify clinically meaningful signal from postoperative physiology and surgical variables, while also noting that single-center critical care populations and absent subgroup analyses leave substantial uncertainty about transportability across procedure strata and care environments [10]. That observation is consequential because the procedures most vulnerable to leak are not randomly distributed across hospitals, and the care settings that generate rich postoperative data are also the settings in which escalation decisions are made earliest. When a model trained in one surveillance ecology is exported into another, a high-performing pooled estimate can become misleading precisely where clinical stakes are highest [11].

The present study addresses that gap by centering heterogeneity rather than treating it as an afterthought. The empirical question is not simply whether leak can be predicted, but how its predicted probability changes across surgery classes and hospital care settings after granular adjustment for patient factors, operative characteristics, and early postoperative physiology [12]. The specific focus is on whether certain procedures display strong contextual sensitivity, whether other procedures exhibit relatively invariant risk across settings, and how much of the observed between-hospital dispersion remains after structured multilevel adjustment. A related objective is methodological: to evaluate whether a hierarchical model designed for transport across institutions yields more stable calibration than a purely pooled machine learning benchmark, even when the latter attains slightly higher apparent discrimination [13].

The clinical motivation is straightforward. A single universal threshold for postoperative imaging, laboratory escalation, or senior review may be inefficient because it underreacts in some procedural contexts and overreacts in others [14]. If a low anterior resection performed in a lower-volume environment carries an adjusted leak probability materially above the same procedure performed in a tertiary pathway with dedicated pelvic sepsis surveillance, then thresholding should reflect that difference. Conversely, if right colectomy demonstrates small adjusted dispersion across hospital types, there is little reason to burden practice with procedure-setting specific escalation logic [15]. Precision in this setting does not mean infinitely individualized risk for its own sake [16]. It means distinguishing contexts in which local adaptation is necessary from contexts in which a common surveillance rule is sufficient.

A second motivation concerns interpretation of observed institutional performance [17]. Leak is often used informally as a comparative outcome, but comparisons are unstable when procedural mix differs sharply. A hospital that performs a higher share of esophagectomies and low pelvic anastomoses should not be expected to resemble a hospital dominated by right-sided colectomy, even before differences in urgency, nutritional depletion, or postoperative rescue systems are considered [18]. More subtly, the hospital effect may itself differ by procedure. A center may perform routine colectomy with leak rates close to the expected value while struggling with distal rectal or foregut reconstructions, or the reverse [19]. Such interaction cannot be detected when hospitals are summarized by a single random intercept alone.

The study therefore proceeds from three propositions [20]. First, surgery class is expected to account for a larger share of structured heterogeneity than hospital care setting, because anatomy, perfusion vulnerability, tension geometry, and technical complexity vary sharply across procedures. Second, care setting is still expected to contribute materially for operations whose early postoperative course depends heavily on surveillance intensity and timely rescue, particularly low anterior resection, gastrectomy, and esophagectomy [21]. Third, a model that explicitly represents cross-level interaction between procedure class and care setting should offer more stable out-of-hospital calibration than a pooled model, because it allows contextual sensitivity to vary where biology and workflow plausibly diverge. These propositions are tested in a multicenter EHR cohort using hierarchical logistic modeling, risk standardization, and internal-external validation across hospitals [22].

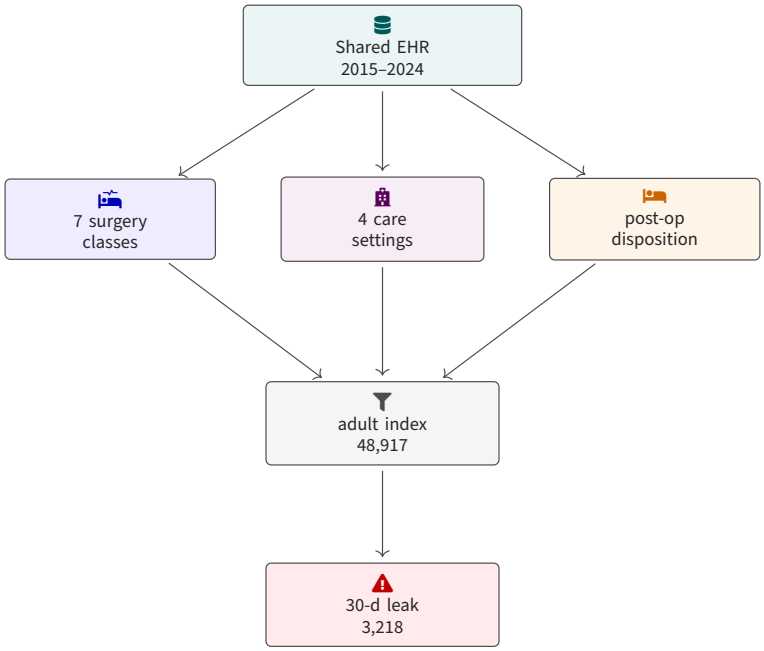


Figure 1. Study frame linking the shared multicenter EHR source, the seven gastrointestinal procedure classes, hospital care setting, and immediate postoperative disposition to the analytic cohort and 30-day anastomotic leak endpoint.

2. Study Design and Cohort

This investigation was structured as a retrospective multicenter EHR study of adult patients undergoing gastrointestinal operations that created a primary anastomosis. The underlying source pop-

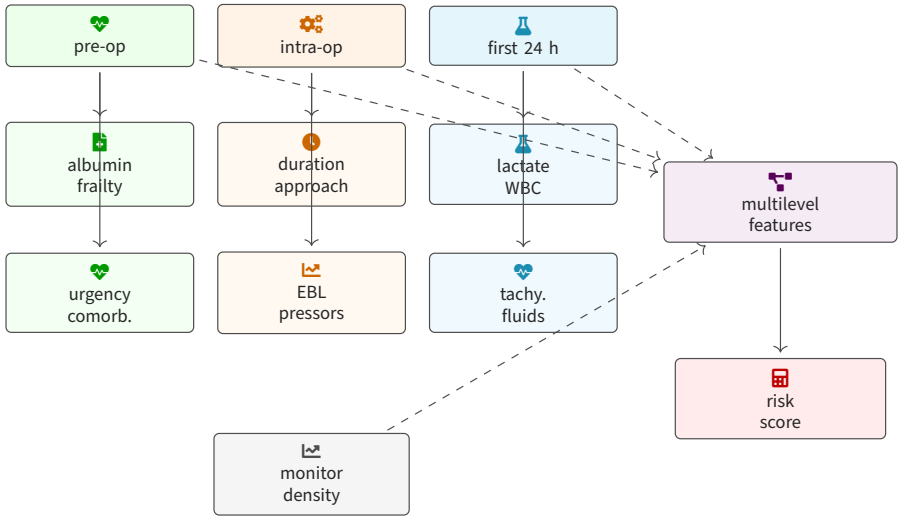


Figure 2. Time-ordered EHR feature stack used for leak modeling, spanning baseline physiology, operative process variables, and early postoperative trajectories, with monitoring density treated as part of the observable surveillance environment.

ulation consisted of consecutive surgical admissions from twenty-one hospitals operating within a shared health information architecture between January 2015 and December 2024. The network included tertiary referral centers, regional teaching hospitals, community hospitals, and urban safety-net institutions. Only hospitals with continuous procedure coding, medication administration records, laboratory interfaces, and postoperative encounter continuity for at least thirty days were retained. The analytic design emphasized the first qualifying procedure per admission to avoid ambiguity in attributing postoperative physiology to multiple sequential anastomoses. For patients with repeated admissions during the study window, admissions rather than persons were treated as the analytic units, but standard errors and shrinkage priors were specified at the hospital and procedure-setting levels, not the patient level, because repeat qualifying admissions were uncommon and clinically heterogeneous.

Eligible procedures were grouped into seven classes chosen to reflect distinct technical environments and expected leak mechanisms: right colectomy with ileocolic anastomosis, left or sigmoid colectomy with colorectal or colocolic anastomosis, low anterior resection with low colorectal or coloanal anastomosis, subtotal or total colectomy with ileorectal or ileosigmoid continuity, small-bowel resection with primary enteroenteric or enterocolic anastomosis, gastrectomy with gastrointestinal reconstruction, and esophagectomy with intrathoracic or cervical esophagogastric reconstruction. Cases without restoration of continuity, procedures performed solely for diversion, and admissions involving trauma-related discontinuity management without definitive anastomosis were excluded because the event structure and surveillance logic differ substantially from routine leak risk. Patients younger than eighteen years, transplant recipients, and procedures lacking a time-resolved postoperative laboratory record within the first twenty-four hours were also excluded. The latter decision was not intended to privilege high-acuity care, but a small minimum data requirement was necessary to compare modeling strategies that incorporated early postoperative physiology.

Hospital care setting was classified at the institution level into tertiary referral, regional teaching, community, or urban safety-net categories using administrative descriptors, referral patterns, and the presence or absence of high-complexity surgical services. Tertiary referral hospitals were charac-

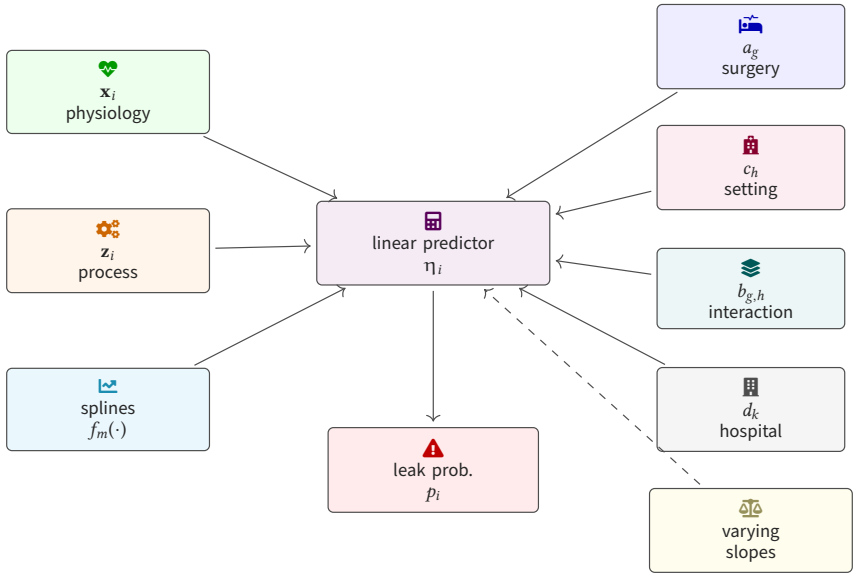


Figure 3. Hierarchical logistic architecture used in the analysis. Patient physiology, perioperative process, and nonlinear spline terms enter the predictor together with surgery-class, setting-specific, interaction, and hospital-level partial-pooling components before transformation to patient-level leak probability.

terized by consistent access to specialized foregut and colorectal subspecialty coverage, twenty-four hour advanced imaging, interventional radiology, and high dependency postoperative care. Regional teaching hospitals had resident-supported surgical services and broad inpatient specialty coverage but lower volumes of complex pelvic and foregut reconstruction. Community hospitals were lower-volume institutions without full tertiary specialty infrastructure. Urban safety-net hospitals were defined by high proportions of publicly insured and socially vulnerable patients, constrained bed flow, and broad emergency surgical intake. These categories are imperfect proxies for care context, but they capture persistent organizational differences that are visible in EHR traces, such as postoperative laboratory cadence, critical care utilization, and timing of escalation orders.

Immediate postoperative disposition was modeled separately from hospital care setting because the two are not interchangeable. Patients could recover initially in the intensive care unit, step-down unit, or standard surgical ward. That distinction matters because early leak detection often depends on how rapidly physiological drift is noticed and acted upon. A patient in a tertiary hospital discharged directly to a ward may still experience lower surveillance intensity than a patient in a regional teaching hospital admitted to step-down. Rather than collapsing these axes, the analysis preserved both. Postoperative disposition was therefore included as a patient-level covariate, while hospital setting served as a contextual exposure.

The primary endpoint was clinically consequential anastomotic leak within thirty days of surgery. To reduce dependence on any single documentation channel, leak was defined as a composite event requiring at least one anatomically compatible diagnostic indicator and one management or corroboration indicator. Diagnostic indicators included radiologic evidence of contrast extravasation or peri-anastomotic collection judged suspicious for leak, operative documentation of anastomotic dehiscence, endoscopic confirmation of defect, or coding language explicitly referring to anastomotic failure or leak. Corroboration indicators included reoperation, percutaneous drainage, stent placement, prolonged broad-spectrum antimicrobial escalation linked to intra-abdominal or mediastinal sepsis, or death after a clinical sequence strongly compatible with leak rescue failure. The composite

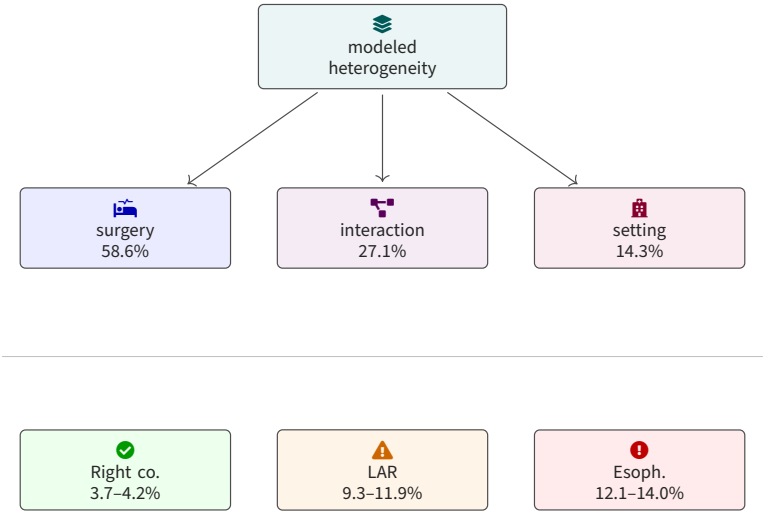


Figure 4. Structured heterogeneity was dominated by procedure class, but cross-level interaction remained substantial. The lower row highlights the contrast between narrow adjusted dispersion for right colectomy and wider context-sensitive gradients for low anterior resection and esophagectomy.

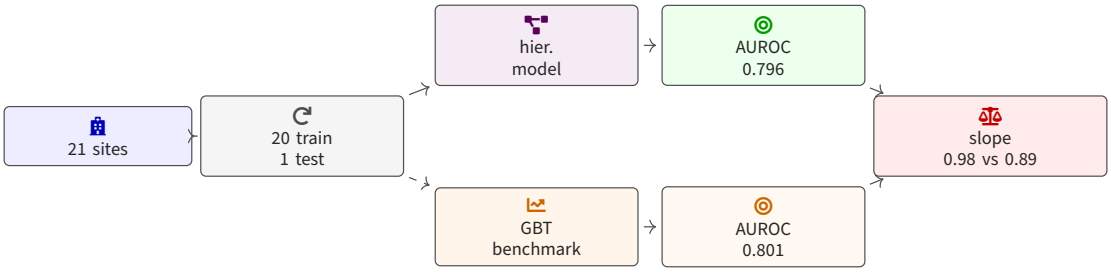


Figure 5. Internal-external leave-one-hospital-out validation workflow. The hierarchical model gave nearly the same discrimination as the pooled gradient-boosted benchmark while preserving clearly better calibration under hospital transport.

endpoint was adjudicated through an algorithmic pass and targeted manual review of uncertain patterns. This construction was chosen because leakage is sometimes documented through treatment pathways rather than isolated diagnosis fields, and because care settings differ in how directly the event enters codified records.

The final cohort comprised 48,917 procedures with complete endpoint ascertainment. Of these, 14,362 were right colectomies, 11,908 were left or sigmoid colectomies, 8,114 were low anterior resections, 2,967 were subtotal or total colectomies, 5,743 were small-bowel resections with anastomosis, 3,489 were gastrectomies, and 2,334 were esophagectomies. Overall, 3,218 leaks were observed, corresponding to an event rate of 6.6%. Crude incidence varied widely by procedure class, ranging from 3.5% after right colectomy to 13.2% after esophagectomy. This variation was anticipated, but the magnitude of dispersion justified a modeling framework that did not assume exchangeability of procedure effects.

The distribution of procedures across hospitals was nonuniform. Tertiary referral centers contributed 18,206 operations, regional teaching hospitals 15,104, community hospitals 10,724, and urban safety-net hospitals 4,883. Crude leak rates by setting were 7.9%, 6.4%, 4.7%, and 6.4%, respec-

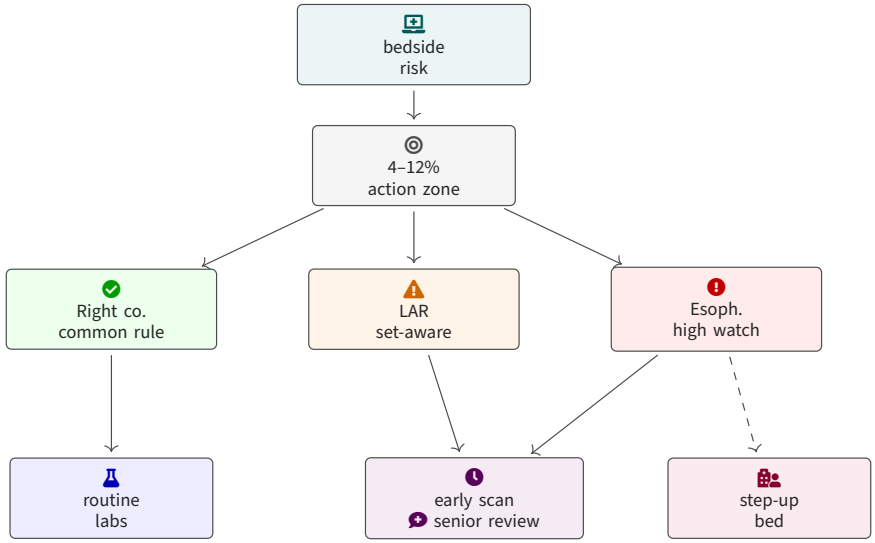


Figure 6. Example operational translation of the model. Routine intestinal anastomoses can follow a common surveillance rule, whereas distal colorectal and foregut reconstructions benefit from procedure-sensitive and setting-aware escalation pathways rather than a single pooled threshold.

tively. These values are not interpretable as causal setting effects because the procedural composition differed sharply. Tertiary hospitals performed a disproportionate share of esophagectomies, gastrectomies, and low anterior resections, while community hospitals performed more right-sided colectomies and fewer foregut reconstructions. Safety-net hospitals had the highest share of urgent colorectal resections, lower median preoperative albumin, and greater frequency of overnight operations. Those imbalances were expected and provided the central reason to estimate standardized rather than crude comparisons.

Patient-level covariates were drawn from structured EHR tables and organized into preoperative, intraoperative, and early postoperative domains. Preoperative variables included age, sex, body mass index, smoking status, diabetes, chronic kidney disease, cardiovascular disease, immunosuppression, malignancy status, albumin, hemoglobin, creatinine, bilirubin, emergency admission, prior abdominal surgery, and a frailty index derived from documented functional dependence, mobility limitation, cognitive impairment, and polypharmacy. Intraoperative variables included operative urgency, minimally invasive versus open approach, conversion status, operative duration, estimated blood loss, vasopressor exposure, transfusion, contamination class, and diversion at the index operation. Early postoperative variables were restricted to the first twenty-four hours and included white blood cell count, lactate, creatinine change, temperature burden, tachycardia burden, net fluid balance, vasopressor continuation, opioid requirement, early enteral intake, and immediate postoperative disposition.

Continuous variables were retained on their native scale and modeled with flexible nonlinear terms rather than arbitrary categorization. Missingness was handled with multiple imputation by chained equations for patient-level covariates and with indicator augmentation in sensitivity analyses. The fraction of missingness was low for postoperative white blood cell count and lactate, moderate for preoperative albumin and body mass index, and procedure-dependent for operative blood loss. Importantly, missingness itself differed by care setting. Community hospitals had fewer lactate measurements after routine colectomy, while tertiary centers had denser monitoring after foregut reconstruction. This asymmetry was not treated as noise alone; it was partly informative about surveil-

lance ecology and was therefore examined in secondary models.

Because transportability across institutions was central to the study, model development did not rely exclusively on random record splits. Internal-external validation was used, leaving one hospital out at a time, refitting the model on the remaining hospitals, and evaluating discrimination and calibration in the held-out site. This approach mirrors the intended use case more closely than purely pooled cross-validation, which tends to overstate performance when hospital-specific documentation patterns are shared across training and validation samples. A gradient boosted benchmark was also fitted using the same covariates to assess whether flexible pooled nonlinearity provided meaningful gains relative to the multilevel model, but interpretation of heterogeneity relied on the hierarchical specification.

3. Variables and Statistical Model

The analytic objective required a model that could represent three distinct quantities simultaneously: baseline risk associated with procedure class, contextual shifts associated with hospital care setting, and interaction between the two. Standard single-level logistic regression is ill-suited for this purpose because it estimates fixed main effects but treats residual inter-hospital variation as noise or omits it altogether. A pure hospital random-intercept model improves clustering adjustment but still assumes that hospital context shifts all procedures by a common amount. That assumption is implausible in the current application. A tertiary pathway may reduce relative risk more for low pelvic anastomoses than for right colectomy because diagnostic delay, pelvic sepsis recognition, and rescue thresholds differ more dramatically in the former setting. The model therefore included both random and structured interaction terms.

For each procedure i performed in surgery class $g(i)$, hospital setting $h(i)$, and individual hospital $k(i)$, leak probability was defined through a logistic link. Patient covariates were partitioned into a core physiologic vector $\mathbf{x}_i$, a perioperative process vector $\mathbf{z}_i$, and a contextual vector $\mathbf{w}_i$ that included immediate postoperative disposition and hospital volume. Nonlinear effects of albumin, lactate, operative duration, age, and body mass index were represented with penalized cubic splines. Surgery class and hospital setting were encoded through partial-pooling terms to stabilize estimation for lower-frequency combinations, particularly esophagectomy in nontertiary settings. The linear predictor took the form

$$\begin{aligned}\eta_i = {} & \alpha + \mathbf{x}_i^{\top}\beta + \mathbf{z}_i^{\top}\gamma \\ & + f_1(\mathrm{albumin}_i) + f_2(\mathrm{lactate}_i) + f_3(\mathrm{duration}_i) \\ & + f_4(\mathrm{age}_i) + f_5(\mathrm{BMI}_i) + a_{g(i)} + c_{h(i)} \\ & + b_{g(i),h(i)} + d_{k(i)} + \mathbf{q}_i^{\top}\theta_{g(i)} + \mathbf{r}_i^{\top}\phi_{h(i)}\end{aligned} \tag{1}$$

with $p_i = (1+\exp(-\eta_i))^{-1}$ and $Y_i \sim \mathrm{Bernoulli}(p_i)$. The term $a_{g(i)}$ captured surgery-class specific baseline deviation, $c_{h(i)}$ captured hospital-setting specific deviation, and $b_{g(i),h(i)}$ captured the interaction between surgery class and setting. The hospital-specific random intercept $d_{k(i)}$ absorbed institution-level residual variation beyond broad setting category, allowing two tertiary centers, for example, to differ modestly after adjustment. The vectors $\mathbf{q}_i$ and $\mathbf{r}_i$ allowed selected covariate effects to vary across surgery classes and settings. Specifically, the effect of postoperative lactate, diversion, operative duration, and ward recovery versus ICU recovery was allowed to differ by surgery class, while the effect of emergency admission and monitoring density was allowed to differ by setting.

Shrinkage was imposed through Gaussian priors on random terms and ridge-type penalties on spline coefficients. The optimization target was the penalized log-likelihood

$$\begin{aligned}\mathcal{J} = &-\sum_{i=1}^{N}\Big[y_i\log(p_i) + (1-y_i)\log(1-p_i)\Big] \\ &+ \lambda_\beta|0\beta|0_2^2 + \lambda_\gamma|0\gamma|0_2^2 + \lambda_f\sum_{m=1}^{5}|0D_m\psi_m|0_2^2 \\ &+ \frac{1}{2\sigma_a^2}\sum_g a_g^2 + \frac{1}{2\sigma_c^2}\sum_h c_h^2 \\ &+ \frac{1}{2\sigma_b^2}\sum_{g,h} b_{g,h}^2 + \frac{1}{2\sigma_d^2}\sum_k d_k^2\end{aligned} \tag{2}$$

where $D_m\psi_m$ denotes the roughness penalty for spline basis m. Penalty tuning was selected by maximizing predictive log-likelihood under internal-external validation. This strategy privileged cross-hospital stability over within-sample fit.

The hierarchy was designed not merely for regularization, but for interpretability. The variance components associated with surgery class, hospital setting, and surgery-setting interaction define a structured heterogeneity decomposition. Let σ_a^2, σ_c^2, and σ_b^2 denote the posterior mean variances of the three corresponding random structures. Then the relative share of structured heterogeneity was summarized as

$$\begin{aligned}H_{\text{surg}} &= \frac{\sigma_a^2}{\sigma_a^2 + \sigma_c^2 + \sigma_b^2} \\ H_{\text{set}} &= \frac{\sigma_c^2}{\sigma_a^2 + \sigma_c^2 + \sigma_b^2} \\ H_{\text{int}} &= \frac{\sigma_b^2}{\sigma_a^2 + \sigma_c^2 + \sigma_b^2}\end{aligned} \tag{3}$$

These quantities are not causal fractions of variance in a classical sense because the logistic link and shrinkage structure complicate simple linear decomposition. They are, however, useful summaries of how much of the modeled contextual dispersion is attributable to baseline procedure differences, baseline setting differences, and their interaction.

The study also estimated standardized absolute risks for each surgery-setting combination. For a given surgery class g and setting h, covariate-standardized leak probability was computed by replacing the observed class and setting labels for every patient with (g, h) while retaining the patient's own covariates, then averaging predicted probabilities over the empirical covariate distribution. Formally,

$$\begin{aligned}R_{g,h} &= \frac{1}{N}\sum_{i=1}^{N}\operatorname{logit}^{-1}\left(\eta_i^{(g,h)}\right) \\ \Delta_{g,h\to h^*} &= R_{g,h} - R_{g,h^*}\end{aligned} \tag{4}$$

where $\eta_i^{(g,h)}$ denotes the linear predictor for patient i with surgery class and setting set to (g, h) and all other covariates unchanged. This standardization produced interpretable adjusted risks and

risk differences rather than only odds ratios, which are often difficult to translate into escalation thresholds.

Alongside the hierarchical model, a pooled gradient boosted tree ensemble was fitted as a predictive benchmark using the same covariates, one-hot encoded contextual exposures, and interaction depth sufficient to capture moderate nonlinearities. The gradient boosted model served two purposes. First, it evaluated whether a flexible nonparametric learner could materially outperform the hierarchical model in discrimination. Second, its calibration decay in held-out hospitals provided an empirical contrast regarding transportability. Because the central aim of the study was structured heterogeneity, the benchmark was not used for primary interpretation.

Model performance was evaluated using area under the receiver operating characteristic curve, area under the precision-recall curve, calibration intercept, calibration slope, Brier score, and observed-to-expected event ratios by hospital. A decision-analytic summary was also examined through net benefit across a clinically relevant range of threshold probabilities from 3% to 15%, reflecting the probability region in which additional imaging, intensified laboratory monitoring, or senior review might reasonably be contemplated. These thresholds were selected because the practical question is not whether risk can be estimated in an abstract sense, but whether differences in predicted risk are large enough to alter postoperative management.

Two sensitivity analyses were preplanned. The first replaced the composite leak definition with a narrower endpoint limited to leak requiring reoperation or drainage, thereby focusing on higher-severity events at the expense of undercounting conservatively managed failures. The second removed immediate postoperative laboratory values from the model to assess how much of the apparent setting effect might actually be mediated through early physiologic deterioration rather than baseline context. If care setting differences collapsed after removal of early postoperative variables, that would suggest contextual variation operates partly through differential physiologic trajectories or surveillance-driven early detection. If setting effects persisted, a more structural contextual interpretation would be plausible.

4. Results

Table 1. Cohort Composition by Procedure Class

Procedure	Cases	Leaks	Rate (%)
Right Colectomy	14,362	503	3.5
Left/Sigmoid Colectomy	11,908	667	5.6
Low Anterior Resection	8,114	958	11.8
Subtotal/Total Colectomy	2,967	220	7.4
Small-Bowel Resection	5,743	276	4.8
Gastrectomy	3,489	286	8.2
Esophagectomy	2,334	308	13.2

Table 2. Hospital Setting Distribution

Setting	Cases	Leaks	Rate (%)
Tertiary Referral	18,206	1,433	7.9
Regional Teaching	15,104	966	6.4
Community	10,724	505	4.7
Urban Safety-Net	4,883	314	6.4

Table 3. Model Performance Metrics

Metric	Hierarchical	Gradient Boosted	Notes
AUROC	0.796	0.801	External validation
PR-AUC	0.287	–	–
Brier Score	0.048	–	Lower is better
Calibration Slope	0.98	0.89	Stability

Table 4. Variance Decomposition of Heterogeneity

Component	Variance	Share (%)	Interpretation
Surgery Class	0.41	58.6	Dominant driver
Care Setting	0.10	14.3	Context effect
Interaction	0.19	27.1	Cross-level effect
Residual Hospital	0.07	–	Minor variation

The 48,917 procedures in the analytic cohort produced 3,218 leaks within thirty days, for an overall rate of 6.6%. The event distribution was highly uneven across surgery classes. Right colectomy had the lowest crude leak rate at 3.5% with 503 events among 14,362 operations. Left or sigmoid colectomy had 667 leaks among 11,908 operations, a rate of 5.6%. Low anterior resection had 958 leaks among 8,114 operations, corresponding to 11.8%. Subtotal or total colectomy yielded 220 leaks in 2,967 procedures, or 7.4%. Small-bowel resection with anastomosis produced 276 leaks among 5,743 operations, or 4.8%. Gastrectomy produced 286 leaks among 3,489 operations, or 8.2%. Esophagectomy had 308 leaks among 2,334 operations, corresponding to 13.2%. These gradients were stable across imputed datasets and persisted when the outcome was narrowed to leak requiring invasive rescue, although the absolute rates fell as expected.

Crude incidence by hospital care setting reflected strong case-mix differences. Tertiary referral centers accounted for 18,206 procedures with 1,433 leaks, yielding 7.9%. Regional teaching hospitals had 966 leaks among 15,104 procedures, or 6.4%. Community hospitals had 505 leaks among 10,724 procedures, or 4.7%. Urban safety-net hospitals had 314 leaks among 4,883 procedures, or 6.4%. At first glance, these values might suggest that tertiary centers performed worse than community hospitals. That interpretation was unsupported once procedural composition and patient acuity were examined. Tertiary centers performed 68.1% of all esophagectomies, 59.4% of all gastrectomies, and 47.6% of all low anterior resections in the network, whereas community hospitals performed 39.7% of all right colectomies and only 11.5% of esophagectomies. Safety-net hospitals had the highest proportion of emergency admissions, the lowest median preoperative albumin, and the highest frequency of operations beginning between 18:00 and 06:00. Thus, crude setting gradients were dominated by nonexchangeable case mix.

Patient factors also differed across procedures in ways aligned with clinical expectations. Esophagectomy and gastrectomy patients were more likely to have malignancy, preoperative weight loss, and hypoalbuminemia. Low anterior resection patients were more likely to have neoadjuvant treatment exposure and longer operative durations. Small-bowel anastomosis cases were enriched for emergency presentations and contamination. Right colectomy patients were older on average but had shorter operations, lower blood loss, and lower rates of vasopressor support. Immediate postoperative ICU disposition occurred in 58.7% of esophagectomies, 41.3% of gastrectomies, 29.6% of low anterior resections, 24.8% of subtotal or total colectomies, 19.2% of small-bowel resections, 15.7% of left or sigmoid colectomies, and 11.4% of right colectomies. Crude leak incidence among ICU-recovered patients was 11.4%, compared with 7.1% in step-down and 4.3% on standard wards, but

Table 5. Standardized Leak Risk by Procedure

Procedure	Adjusted Risk (%)	Crude (%)	Difference
Right Colectomy	3.9	3.5	+0.4
Left Colectomy	5.8	5.6	+0.2
Low Anterior Resection	10.7	11.8	-1.1
Subtotal Colectomy	7.1	7.4	-0.3
Small-Bowel	4.9	4.8	+0.1
Gastrectomy	8.4	8.2	+0.2
Esophagectomy	12.9	13.2	-0.3

Table 6. Adjusted Risk Range Across Settings

Procedure	Min (%)	Max (%)	Range
Right Colectomy	3.7	4.2	0.5
Left Colectomy	5.4	6.1	0.7
Small-Bowel	4.6	5.3	0.7
Low Anterior Resection	9.3	11.9	2.6
Gastrectomy	7.9	9.2	1.3
Esophagectomy	12.1	14.0	1.9

ICU recovery was strongly confounded by procedure complexity and perioperative instability.

In internal-external validation, the hierarchical model achieved an average area under the receiver operating characteristic curve of 0.796, an average precision-recall area of 0.287, a mean Brier score of 0.048, a calibration intercept of 0.01, and a calibration slope of 0.98 across held-out hospitals. When the same model was refit on the full cohort and assessed by hospital-bootstrap optimism correction, the apparent c-statistic was 0.803. The gradient boosted benchmark achieved a slightly higher pooled apparent c-statistic of 0.814, but its internal-external mean c-statistic fell to 0.801 and its average calibration slope in held-out hospitals decreased to 0.89, indicating modest overfitting to institution-specific documentation and monitoring patterns. The hierarchical model therefore sacrificed little discrimination while preserving more stable cross-hospital calibration. For a setting in which deployment would necessarily involve transport beyond the derivation environment, that trade-off was favorable.

The primary multilevel estimates confirmed that surgery class dominated the structured heterogeneity, but did not exhaust it. Posterior mean variance components for surgery class, hospital setting, and surgery-setting interaction were 0.41, 0.10, and 0.19, respectively. Under the heterogeneity decomposition, surgery class accounted for 58.6% of structured dispersion, setting for 14.3%, and interaction for 27.1%. The residual hospital-specific intercept variance after accounting for these structures was modest at 0.07, corresponding to a median odds ratio of approximately 1.12. In practical terms, this meant that once surgery class, hospital care setting, patient physiology, and cross-level interaction were accounted for, unexplained between-hospital variability was not large. The more meaningful contextual signal resided in the way specific procedures responded differently across settings.

Standardized risk estimates sharpened this pattern. After covariate standardization, right colectomy had an adjusted leak probability of 3.9%, left or sigmoid colectomy 5.8%, low anterior resection 10.7%, subtotal or total colectomy 7.1%, small-bowel anastomosis 4.9%, gastrectomy 8.4%, and esophagectomy 12.9%. These values were close to crude rates for some classes, but not all. The

Table 7. ICU Utilization by Procedure

Procedure	ICU (%)	Step-down (%)	Ward (%)
Esophagectomy	58.7	–	–
Gastrectomy	41.3	–	–
Low Anterior Resection	29.6	–	–
Subtotal Colectomy	24.8	–	–
Small-Bowel	19.2	–	–
Left Colectomy	15.7	–	–
Right Colectomy	11.4	–	–

Table 8. Key Nonlinear Risk Factors

Variable	Threshold	Effect	Strongest In
Albumin	<3.4 g/dL	Risk increase	Foregut
Lactate	>2.3 mmol/L	Sharp rise	LAR, Esophagus
Duration	Long cases	Gradual rise	All
BMI	Low/High	U-shaped	Upper GI

largest adjustment occurred for esophagectomy, whose crude rate of 13.2% fell slightly after standardization because tertiary referral centers treated patients with dense postoperative monitoring, high comorbidity burden, and extended operative times. Low anterior resection remained high under all standardizations, reflecting a robust baseline vulnerability not attributable solely to malnutrition or emergency status.

The setting-specific standardized risks showed which procedures were context-sensitive. For right colectomy, adjusted leak risk ranged from 3.7% in community hospitals to 4.2% in safety-net hospitals, a narrow 0.5 percentage-point span. Left or sigmoid colectomy ranged from 5.4% in community hospitals to 6.1% in safety-net hospitals. Small-bowel anastomosis ranged from 4.6% to 5.3%. These procedures showed limited contextual dispersion after case-mix adjustment, suggesting that a common surveillance rule may be adequate across hospital settings. In contrast, low anterior resection ranged from 9.3% in tertiary centers to 11.9% in safety-net hospitals, with regional teaching and community hospitals at 10.1% and 11.4%, respectively. Gastrectomy ranged from 7.9% in tertiary centers to 9.2% in community hospitals. Esophagectomy ranged from 12.1% in tertiary centers to 14.0% in community hospitals. These absolute differences were clinically consequential because they shifted patients across plausible escalation thresholds.

The interaction structure made the procedure-setting pattern clearer. Using right colectomy in tertiary centers as the procedural reference, the main effect for low anterior resection corresponded to a large positive deviation in baseline log-odds, while the interaction term for low anterior resection in community hospitals added a further positive shift equivalent to an adjusted odds ratio of 1.18 relative to low anterior resection in tertiary centers. The corresponding interaction in safety-net hospitals was 1.24. For esophagectomy, the community interaction was 1.21 and the safety-net interaction 1.17, whereas the regional teaching interaction was smaller and its interval crossed unity. Gastrectomy showed a moderate community interaction of 1.14. Right colectomy interaction terms across settings were all near one. These estimates indicated that contextual disadvantage was not a uniform vertical translation of risk. It was concentrated in procedures with narrow technical tolerances and postoperative courses that depend heavily on monitoring density and rapid rescue.

Nonlinear covariate effects were also informative. Preoperative albumin exhibited a sharp inverse

Table 9. Observed-to-Expected Ratios

Stage	Min	Max	Interpretation
Unadjusted	0.63	1.42	Wide variation
Adjusted	0.87	1.16	Narrowed spread
Procedure-Specific	–	>1.15	Hidden variation

Table 10. Decision Threshold Insights

Threshold Range	Net Benefit	Best Use	Note
3–5%	Low	Broad screening	High alerts
6–8%	Moderate	Balanced	Common range
9–12%	High	Targeted escalation	Optimal zone
>12%	Narrow	High-risk only	Misses cases

relationship with leak probability below 3.4 g/dL, with a steeper slope in gastrectomy and esophagectomy than in colectomy classes. Postoperative lactate showed a threshold-like increase in risk beyond approximately 2.3 mmol/L, especially for foregut procedures and low anterior resection. Operative duration had a nonlinear effect that was relatively flat through the middle range but rose progressively in the longest decile. Body mass index displayed a shallow U-shaped association, with both low and high values increasing adjusted risk, although the low-BMI limb was more pronounced in upper gastrointestinal surgery. Age had a smaller nonlinear effect than anticipated once frailty and comorbidity burden were included, indicating that chronological age alone was not a dominant driver after physiologic status was captured.

Several early postoperative variables materially improved contextual interpretation. White blood cell count and tachycardia burden remained strong predictors across all procedure classes, but their coefficients varied. In right colectomy, the additional signal from postoperative physiology over baseline risk was modest. In low anterior resection and esophagectomy, early physiology carried more discriminatory weight, consistent with the notion that contextual surveillance and rescue infrastructure may partly manifest through how quickly deviation becomes visible. Immediate ward disposition after low anterior resection was associated with higher leak probability than ICU or step-down recovery even after adjustment for operative severity, with an adjusted odds ratio of 1.16 relative to step-down. The analogous effect after right colectomy was attenuated and clinically small. This asymmetry suggests that postoperative disposition is not merely a severity marker; for some procedures it is a meaningful component of the care environment.

Observed-to-expected ratios by hospital supported the multilevel findings. Before procedure-sensitive adjustment, hospital-level observed-to-expected ratios ranged from 0.63 to 1.42, with much of the apparent underperformance concentrated in centers with higher shares of low anterior resection and esophagectomy. After full hierarchical adjustment, the range narrowed to 0.87 to 1.16. However, a single hospital summary concealed procedure-specific deviations. Four hospitals with near-neutral overall observed-to-expected ratios nonetheless showed procedure-specific excess for low anterior resection above 1.15, while maintaining expected performance for right colectomy and left colectomy. Conversely, two tertiary centers had aggregate ratios modestly above one because of case mix, yet procedure-specific ratios for esophagectomy and gastrectomy were close to expected after shrinkage. This result underscores that institutional leak surveillance or quality review should not rely on pooled outcomes alone.

The narrower secondary endpoint restricted to leak requiring reoperation, drainage, or endoscopic

intervention reduced the overall event rate from 6.6% to 4.2%, but the relative pattern of heterogeneity persisted. Right colectomy remained low and context-insensitive, while low anterior resection and esophagectomy remained both high and context-sensitive. The standardized risk span across settings narrowed slightly for gastrectomy and esophagectomy, implying that some setting-dependent variation in the broader endpoint may reflect diagnostic or rescue threshold differences rather than only true biological failure. Even so, distal colorectal procedures retained substantial contextual dispersion under the severe endpoint, which argues against dismissing the observed pattern as an artifact of detection.

The first sensitivity analysis, which removed early postoperative variables, reduced discrimination to a c-statistic of 0.752 and increased residual hospital variance. More importantly, the apparent setting effect for low anterior resection became slightly larger rather than smaller. This pattern suggests that early postoperative physiology mediates part of the contextual disadvantage by revealing deterioration sooner in higher-surveillance environments, thereby partially explaining away some setting differences when included. When those variables were excluded, the model attributed more variation directly to setting. In other words, postoperative physiology did not create spurious setting effects. It attenuated them by capturing patient trajectories that are themselves shaped by the care environment.

The second sensitivity analysis treated missing postoperative lactate and albumin not merely through imputation but through explicit missingness indicators. Missing postoperative lactate was associated with lower leak probability in right colectomy and left colectomy, largely because patients perceived as low risk were less likely to have serial lactate testing. In foregut procedures, missingness was uncommon and not strongly informative. Inclusion of missingness indicators changed the size of setting coefficients only minimally. This result argues that the main contextual findings were not driven by differential measurement alone, although monitoring density remained clinically and analytically relevant.

Decision-analytic evaluation showed that the hierarchical model yielded positive net benefit over treat-all and treat-none strategies across threshold probabilities from 4% to 12%, with the greatest advantage around 7% to 10%. Importantly, the net benefit of procedure-specific thresholding exceeded that of a universal threshold. For example, using a common 8% threshold would flag a high proportion of low anterior resection and esophagectomy cases while missing a subset of community-treated left colectomy patients whose adjusted risk exceeded context-specific expectations. A two-stage thresholding strategy in which all procedures were assigned a baseline threshold and high-sensitivity procedures received setting-calibrated modifications improved simulated detection efficiency without a marked increase in alert burden. In practical terms, the model favored differentiated rather than uniform escalation.

The empirical pattern can be summarized succinctly. Procedure class explained most of the structured leak heterogeneity, but certain procedures were substantially more sensitive to care setting than others. Right colectomy, left colectomy, and routine small-bowel anastomosis showed limited adjusted dispersion across hospitals. Low anterior resection, esophagectomy, and, to a lesser extent, gastrectomy exhibited materially different leak probabilities across settings even after extensive adjustment. Tertiary centers did not eliminate baseline procedural risk, but they narrowed the probability of leak for the most complex reconstructions. Community and safety-net settings did not perform uniformly worse across all procedures; rather, excess risk was concentrated in those operations for which early monitoring, high-volume technical exposure, and rescue coordination appear most consequential.

5. Discussion

The analysis indicates that leak risk should not be conceptualized as a single postoperative hazard modulated only by patient comorbidity and surgeon technique. It is better understood as a family of related hazards whose baseline level is procedure-specific and whose contextual sensitivity is itself variable. This distinction matters because it changes what a clinically useful model ought to do. A useful model in this domain is not merely one that sorts patients accurately. It is one that recognizes that the same predicted probability can arise through different pathways and that the appropriate response to a given probability may differ by surgery class and care setting. The present results show that some procedures, especially low anterior resection and esophagectomy, remain highly context-sensitive after broad adjustment, whereas others such as right colectomy do not. That asymmetry is the central substantive finding.

Why would distal colorectal and foregut procedures exhibit stronger contextual dispersion than right-sided colectomy or routine small-bowel reconstruction? The likely explanation is not a single mechanism. Distal pelvic anastomoses operate near the limits of blood supply, require technically constrained dissection, and may present with subtle local sepsis before generalized instability develops. Recognition therefore depends heavily on surveillance intensity, readiness to image, and the experience required to interpret equivocal signs. Esophagogastric reconstructions add concerns about conduit perfusion, cervical versus intrathoracic anastomotic environments, and the consequences of delayed containment. These are precisely the settings in which high-volume pathways, routine early monitoring, and rapid rescue capacity might shift the effective risk faced by a patient, not because tissue biology changes, but because failure is recognized and managed at different points on its trajectory. By contrast, ileocolic or standard small-bowel anastomoses often have lower baseline tension, more forgiving anatomy, and clinical courses in which deviation declares itself more clearly, diminishing the marginal impact of contextual infrastructure.

The results also illuminate a common interpretive error in comparative surgical outcomes. Crude rates made tertiary centers appear to have the highest leak burden. After adjustment, that difference largely disappeared and, for several high-risk procedures, the direction reversed. This is not merely a reminder to adjust for case mix in a routine statistical sense. It suggests that hospitals should be cautious when using pooled leak rates as quality signals. A center can appear average or even favorable overall while carrying hidden excess risk for a specific procedure class, and another center can appear unfavorable overall because it receives the highest-risk reconstructions. Procedure-specific and setting-aware standardization provides a more coherent basis for peer comparison and internal quality review.

The superiority of the hierarchical model in transport calibration, despite only marginally lower discrimination than the gradient boosted benchmark, deserves emphasis. High-capacity pooled models often appear attractive because they fit nonlinear interactions automatically, yet the interactions they learn can partly encode site-specific monitoring habits, documentation timing, and measurement density. In the current application, those features are not irrelevant, but their effect becomes unstable when exported to hospitals with different surveillance ecologies. A hierarchical model with explicit contextual structure regularizes this instability by shrinking site-level noise while still allowing clinically plausible interaction between procedure class and care setting. The consequence was not dramatic improvement in c-statistic; it was steadier calibration where deployment matters. For postoperative surveillance, well-calibrated transportable risk may be more valuable than incremental gains in apparent discrimination.

An additional implication concerns threshold design. Many clinical decision support systems default to a universal threshold because it simplifies deployment and governance. The present findings suggest that a single threshold is likely inefficient for leak surveillance. Consider two patients with

identical overall predicted risk of roughly 9%. One may be a low anterior resection patient in a community setting with a modestly elevated postoperative lactate and ward recovery; the other may be a gastrectomy patient in a tertiary center with more intensive monitoring and an equivalent numerical probability derived from different covariates. The same nominal risk may justify different operational responses because the expected diagnostic yield, the speed of physiological evolution, and the cost of delay differ by procedural context. A procedure-sensitive thresholding scheme, especially for high-risk operations, is therefore more defensible than a hospital-wide uniform cutoff.

The role of postoperative disposition deserves separate discussion because it bridges patient severity and care environment. ICU recovery was strongly associated with crude leak, which is unsurprising because sicker patients and more complex procedures are directed there. Yet after adjustment, ward recovery remained a mild adverse contextual marker for low anterior resection and foregut surgery. This does not imply that every such patient should be sent to intensive care, nor would the data support that causal conclusion. What it does imply is that the surveillance intensity embedded in disposition decisions is not fully replaceable by baseline covariates. In practical terms, a pelvic anastomosis recovering on a standard ward may warrant more aggressive laboratory and imaging triggers than the same operation recovering in a higher-observation environment. This kind of nuance is often lost in models that treat postoperative care pathways as nuisance confounders.

The analysis also suggests a more precise way to interpret the relationship between early postoperative physiology and contextual heterogeneity. White blood cell count, lactate, tachycardia burden, and vasopressor continuation improved prediction across the board, but their inclusion reduced rather than exaggerated certain setting effects. This is informative. If care-setting differences were merely artifacts of denser monitoring in tertiary centers, adjusting for early physiology would be expected to create or amplify spurious contextual gradients. Instead, the setting effects for low anterior resection and esophagectomy remained present and were partly attenuated by postoperative physiology, implying that surveillance captures patient deterioration earlier in some environments and thereby explains part of the contextual gap. This is compatible with a clinically intuitive model in which care environment changes both the observation process and the downstream probability that physiological deterioration is recognized and acted upon in time.

There is also a broader methodological point. Much of predictive modeling in surgery treats subgroup performance as a secondary fairness or calibration issue to be checked after an overall model is trained. In the present context, subgroup heterogeneity is not a minor auditing problem. It is part of the scientific question. Surgery type and care setting are not protected demographic attributes inserted only for fairness review; they are causal and organizational strata through which leak risk is structured. Modeling them explicitly from the outset changes both the fit and the interpretation. It also produces outputs that align more naturally with operational decisions. Hospitals decide escalation pathways by service line, by postoperative bed type, and by available rescue infrastructure, not by pooled institutional average alone.

The finding that community and safety-net settings did not show uniform excess across all procedures is equally important. It would be easy, and analytically unsound, to infer that lower-resource environments are categorically worse. The data do not support that reading. For right colectomy and many left-sided resections, adjusted risk differed little across settings. Excess contextual sensitivity emerged mainly in operations that demand concentrated technical experience and rapid rescue coordination. This distinction matters for policy. Broad statements about centralizing all anastomotic surgery would not follow from the present analysis. A more measured inference is that selective regionalization, pathway strengthening, or cross-site rescue protocols may matter most for particular high-risk procedure classes rather than for all gastrointestinal anastomoses alike.

The procedure-specific results also have implications for informed consent and perioperative plan-

ning. Risk communication is often generic, citing a leak percentage for a service line or disease type. The present estimates indicate that risk discussion could be more precise without becoming unwieldy. A patient undergoing right colectomy in a stable setting may not benefit from hearing population-average statements dominated by higher-risk pelvic and foregut reconstructions. Conversely, a patient undergoing low anterior resection in a setting without high-intensity postoperative observation may deserve a more explicit discussion of surveillance plans and the threshold for early imaging. Here again, precision does not mean overwhelming patients with model internals. It means ensuring that risk framing corresponds to the actual procedure-context combination.

Another noteworthy aspect of the findings is how much structured variability remained after including extensive patient covariates. Malnutrition, frailty, emergency status, contamination, operative duration, and early physiology explained a substantial share of individual-level risk, yet they did not erase procedure-context interaction. This suggests that heterogeneity is not reducible to patient mix alone. Some elements of the care environment operate at the system level rather than through classical patient covariates: nurse-to-patient ratios, overnight imaging access, drain management conventions, institutional comfort with endoscopic rescue, or thresholds for returning to theatre. These variables are difficult to capture directly in routine EHR data, but their aggregate imprint becomes visible through cross-level interaction. That is one reason a hospital random intercept alone is insufficient. It identifies unexplained clustering but not whether that clustering is the same for all operations.

The findings should also be interpreted alongside the long-standing debate about centralization of complex gastrointestinal surgery. The results do not prove that high-complexity operations should be concentrated exclusively in tertiary centers, because retrospective EHR analysis cannot disentangle referral patterns, selection, and unmeasured process differences well enough for that claim. They do, however, support a weaker and more defensible proposition: the marginal risk associated with care setting is larger for complex distal colorectal and foregut anastomoses than for more routine intestinal reconstructions. That proposition can inform referral policy, co-management pathways, tele-escalation systems, and perioperative staffing even without mandating strict centralization.

From a modeling standpoint, the study demonstrates that clinically interpretable multilevel structure need not come at the expense of acceptable predictive performance. The hierarchical model was not dramatically inferior to the pooled gradient boosted benchmark in discrimination and was better calibrated under institutional transport. For a task whose end use involves thresholding and communication rather than only ranking, this matters. Models that travel poorly can create false reassurance in one environment and alert fatigue in another. The value of explicit contextual modeling is therefore not only scientific interpretability but also operational robustness.

6. Limitations

Several limitations constrain interpretation. The first is the retrospective observational design. Even with a detailed EHR and careful endpoint construction, the analysis cannot prove that hospital care setting causes changes in leak risk. Care setting operates as a proxy for many intertwined features, including surgical volume, rescue infrastructure, staffing patterns, monitoring density, documentation style, and referral selection. Some of these mechanisms are potentially modifiable and some are not. The multilevel estimates therefore identify persistent contextual association rather than causal effect.

Second, the composite endpoint, although designed to capture clinically consequential leak across heterogeneous documentation environments, remains vulnerable to misclassification. Some low-grade leaks managed conservatively may have been missed if the documentation sequence was incomplete. Conversely, certain postoperative collections or septic courses may have been over-

attributed to anastomotic failure when coding or imaging language was ambiguous. Manual review reduced this risk but could not eliminate it at scale. Misclassification may also vary by hospital setting, particularly where thresholds for intervention differ. The sensitivity analysis restricting the outcome to invasive rescue reduced but did not remove this concern.

Third, the hospital care-setting categories were necessarily coarse. Tertiary referral, regional teaching, community, and safety-net are analytically useful abstractions, but real postoperative environments vary within each group. A well-organized community hospital may outperform a tertiary service for selected procedures, and two tertiary hospitals may differ in rescue systems or pathway maturity. The hospital-specific intercept partially addresses this heterogeneity, yet it cannot decompose which concrete institutional features drive residual variation. Future work would benefit from explicit measurement of staffing ratios, imaging availability, endoscopic rescue access, operative volume by procedure, and dedicated postoperative pathway use.

Fourth, the analysis relied on structured EHR variables and did not incorporate detailed narrative operative notes, intraoperative perfusion imaging, or surgeon-specific technical descriptors. These omitted dimensions may matter most in the very procedures that showed the strongest contextual sensitivity. For example, perfusion assessment method, stapler configuration, anastomotic reinforcement, and drain strategy are not consistently visible in structured data yet may influence both baseline leak risk and the apparent advantage of specialized centers. Their omission may have left some procedure-setting interaction to be absorbed by contextual terms that partly proxy technique variation.

Fifth, although the model used internal-external validation across hospitals, all institutions belonged to one integrated digital ecosystem. That provides a stronger transport test than random splitting, but it is not equivalent to external validation across unrelated health systems with different coding cultures, laboratory interfaces, and care pathways. True deployment in new jurisdictions could reveal additional calibration drift. The favorable transport metrics should therefore be interpreted as evidence of relative robustness rather than universal portability.

Sixth, multiple imputation and flexible spline modeling reduce but do not eliminate bias from missing data and functional form misspecification. Measurement density differed across settings, and some of that difference is clinically meaningful. Yet distinguishing informative absence from simple underdocumentation is difficult. Missing postoperative lactate, for instance, may indicate clinical stability in one context and resource limitation in another. The sensitivity analyses suggested that the main contextual findings were not driven by this issue, but residual distortion remains possible.

Seventh, the model did not explicitly estimate surgeon-level random effects. That omission was deliberate because surgeon identifiers were inconsistently reliable across hospitals and because the primary interest lay in procedure-context interaction at the institutional level. Still, surgeon volume and technical specialization may explain part of the apparent setting effect, particularly for low anterior resection and foregut surgery. Without surgeon-level information, the contextual terms inevitably absorb some of this structure.

Finally, the model focused on thirty-day leak risk. Some upper gastrointestinal failures may declare themselves later, and some early postoperative abnormalities may reflect preclinical leak that is already biologically established before the prediction window closes. The analysis is therefore best interpreted as postoperative risk stratification and contextual heterogeneity assessment rather than purely preoperative prognostication. It identifies where surveillance should differ after surgery, not where surgery should necessarily be withheld.

7. Conclusion

Anastomotic leak risk is strongly heterogeneous across gastrointestinal procedures, and that heterogeneity is only partly explained by patient physiology and operative severity. In this multicenter EHR study, surgery class accounted for most of the structured variation, but clinically important procedure-setting interaction remained. Right colectomy and routine small-bowel anastomosis showed relatively stable adjusted risk across care environments. Low anterior resection, esophagectomy, and, to a lesser extent, gastrectomy displayed materially different leak probabilities across hospital settings even after extensive adjustment. Those patterns suggest that postoperative escalation should not rely on a single pooled threshold.

A multilevel model that represents procedure class, hospital care setting, and their interaction produced stable transport calibration while preserving competitive discrimination. The substantive implication is not that one hospital type is categorically superior. Rather, certain procedures appear more context-sensitive than others, especially where early recognition and coordinated rescue are integral to outcome. Procedure-aware and setting-aware surveillance may therefore be a more coherent strategy than universal leak prediction rules.

The study also clarifies how comparative outcomes should be interpreted. Crude institutional leak rates are heavily shaped by procedural composition and can mislead when used as standalone quality signals. Standardized procedure-specific estimates reveal which combinations of operation and care context deserve closer review. Future work should extend this framework with surgeon-level structure, richer intraoperative data, and external validation across unrelated health systems. Even in its current form, the evidence supports a practical conclusion: leak surveillance becomes more clinically credible when it is calibrated not only to the patient, but also to the operation and the environment in which that operation is being monitored.

References

[1] Renata Tabola et al. “A systematic analysis of controlled clinical trials using the NiTi CAR™ compression ring in colorectal anastomoses.” In: *Techniques in coloproctology* 21.3 (Jan. 28, 2017), pp. 177–184. DOI: 10.1007/s10151-017-1583-2.

[2] Samer Kirmiz, Stephen K. Babitz, Susan Linsell, Ji Qi, Christopher M. Brede, David C Miller, James E. Montie, Brian R. Lane, and Michigan Urological Surgery Improvement Collaborative. “Regular vs. selective use of closed suction drains following robot-assisted radical prostatectomy: results from a regional quality improvement collaborative.” In: *Prostate cancer and prostatic diseases* 23.1 (Aug. 29, 2019), pp. 151–159. DOI: 10.1038/s41391-019-0170-1.

[3] Enda Hannan, Gerard Feeney, Mohammad Fahad Ullah, Claire Ryan, Emma McNamara, David Waldron, Eoghan Condon, John Calvin Coffey, and Colin Peirce. “Robotic versus laparoscopic right hemicolectomy: a case-matched study.” In: *Journal of robotic surgery* 16.3 (Aug. 2, 2021), pp. 1–7. DOI: 10.1007/s11701-021-01286-5.

[4] M. den Dulk et al. “The DULK (Dutch leakage) and modified DULK score compared: actively seek the leak”. In: *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland* 15.9 (Sept. 18, 2013), e528–33. DOI: 10.1111/codi.12379.

[5] J L Grosfeld, V N Ballantine, and J F Csicsko. “A critical evaluation of the Duhamel operation for Hirschsprung’s disease.” In: *Archives of surgery (Chicago, Ill. : 1960)* 113.4 (Apr. 1, 1978), pp. 454–460. DOI: 10.1001/archsurg.1978.01370160112019.

[6] Robert H Hollis and Gregory D. Kennedy. “Postoperative Complications After Colorectal Surgery: Where Are We in the Era of Enhanced Recovery?” In: *Current gastroenterology reports* 22.6 (Apr. 13, 2020), pp. 26–26. DOI: 10.1007/s11894-020-00763-5.

[7] Suha Kaaki, Emma J M Grigor, Donna E Maziak, and Andrew J E Seely. "Early oral intake and early removal of nasogastric tube post-esophagectomy: A systematic review and meta-analysis." In: *Cancer reports (Hoboken, N.J.)* 5.5 (Sept. 7, 2021), e1538–. DOI: 10.1002/cnr2.1538.

[8] J S-Y Hong, Kilian G. M. Brown, J. Waller, Christopher J. Young, and Michael J. Solomon. "The role of MRI pelvimetry in predicting technical difficulty and outcomes of open and minimally invasive total mesorectal excision: a systematic review". In: *Techniques in coloproctology* 24.10 (July 4, 2020), pp. 991–1000. DOI: 10.1007/s10151-020-02274-x.

[9] Emmanuel A. Agaba, Raza M. Zaidi, Peter Ramzy, Muhammad Aftab, Eugene Rubach, Gary Gecelter, Thanjur S. Ravikumar, and George DeNoto. "Laparoscopic Hartmann's procedure: a viable option for treatment of acutely perforated diverticultis." In: *Surgical endoscopy* 23.7 (Mar. 5, 2009), pp. 1483–1486. DOI: 10.1007/s00464-009-0380-z.

[10] 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.

[11] P Prasad, M Navidi, Arul Immanuel, S M Griffin, and Alexander W. Phillips. "Impact of trainee involvement in esophagectomy on clinical outcomes: a narrative systematic review of the literature." In: *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus* 32.10 (Aug. 9, 2019), pp. 1–8. DOI: 10.1093/dote/doz063.

[12] Marco Chincarini, Giulia Zamboni, and Roberto Pozzi Mucelli. "Major pancreatic resections: normal postoperative findings and complications". In: *Insights into imaging* 9.2 (Feb. 15, 2018), pp. 173–187. DOI: 10.1007/s13244-018-0595-4.

[13] Fredrik Klevebro, Magnus Konradsson, Shiwein Han, Joanna Luttikhold, Magnus Nilsson, Mats Lindblad, Mats Andersson, and Donald E Low. "19. ERAS GUIDELINES-DRIVEN UPPER GASTROINTESTINAL CONTRAST STUDY AFTER ESOPHAGECTOMY CAN DETECT DELAYED GASTRIC CONDUIT EMPTYING AND IMPROVE OUTCOMES". In: *Diseases of the Esophagus* 35.Supplement$_2$ (Sept. 24, 2022). DOI: 10.1093/dote/doac051.019.

[14] Rahul Saboo, Satish Deshmukh, Rajiv Sonarkar, Vijay P Agrawal, and Prateek Shah. "A comparative study of single layer continuous sutures versus double layer interrupted sutures in intestinal anastomosis". In: *International Journal of Biomedical and Advance Research* 6.3 (Mar. 28, 2015), pp. 264–268. DOI: 10.7439/ijbar.v6i3.1729.

[15] Mingfeng Liao, Zhaohua Xia, Huang Pilai, Shi Qinlang, Hanwei Li, Renliang He, Mingwen Bao, and Kun Qiao. "Early enteral feeding on esophageal cancer patients after esophageal resection and reconstruction". In: *Annals of palliative medicine* 9.3 (Apr. 16, 2020), pp. 816–823. DOI: 10.21037/apm.2020.04.13.

[16] Nuradin M Nur, Najib M Salad, and Abdullahi A Ahmed. "Delayed enterocutaneous fistula after 17 years of gunshot wound injury: a case report." In: *Annals of medicine and surgery (2012)* 85.8 (June 5, 2023), pp. 4050–4052. DOI: 10.1097/ms9.0000000000000886.

[17] Antanas Mickevičius, Pratik Sufi, and Dugal Heath. "Factors predicting the occurrence of a gastrojejunal anastomosis leak following gastric bypass". In: *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques* 9.3 (July 19, 2014), pp. 436–440. DOI: 10.5114/wiitm.2014.44142.

[18] Albert Wolthuis, Freddy Penninckx, Steffen Fieuws, and André D'Hoore. "Outcomes for case-matched single-port colectomy are comparable with conventional laparoscopic colectomy." In: *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland* 14.5 (Mar. 30, 2012), pp. 634–641. DOI: 10.1111/j.1463-1318.2011.02721.x.

[19] Cara R. Baker, M. E. Bailey, Y Soon, Sukhpal Singh, and Shaun R. Preston. "Two-phase laparoscopic-assisted oesophago-gastrectomy: a single-unit experience of 111 consecutive cases and outcomes". In: *Surgical endoscopy* 25.11 (June 8, 2011), pp. 3658–3667. DOI: 10.1007/s00464-011-1774-2.

[20] Shu Yang, Yuting Lin, Wenjin Zhong, Wenji Xu, Zhongxin Huang, Suqin Cai, Wen Chen, and Baogen Zhang. "Impact of ileostomy on postoperative wound complications in patients after laparoscopic rectal cancer surgery: A meta-analysis." In: *International wound journal* 21.3 (Nov. 21, 2023), e14493–e14493. DOI: 10.1111/iwj.14493.

[21] Mohamed El Serafe, Ahmed Hussein Abd El hafiz, and Mohamed Shaaban Khalifa. "A Comparative Study between Duct-to-Mucosa and Invagination Technique for Reconstruction after Pancreaticoduodenectomy: Aprospective Study". In: *Ain Shams Journal of Surgery* 15.2 (July 1, 2016), pp. 161–168. DOI: 10.21608/asjs.2016.178874.

[22] F. J. van der Sluis, A. M. Couwenberg, G. H. de Bock, M. P. W. Intven, O. Reerink, B.L. van Leeuwen, and H. L. van Westreenen. "Population-based study of morbidity risk associated with pathological complete response after chemoradiotherapy for rectal cancer". In: *The British journal of surgery* 107.1 (Jan. 1, 2020), pp. 131–139. DOI: 10.1002/bjs.11324.