ARTICLE

Empirical Transport Instability in Early Anastomotic Leak Prediction Across Diverse Colorectal Surgery Hospitals and Postoperative Care Settings

Somsak Wattanapong,[1] Chaiwat Phrommin,[2] and Narin Boonprasert[3]

[1]Faculty of Informatics, Mahasarakham University, Khamriang Subdistrict, Kantarawichai District, Maha Sarakham 44150, Thailand
[2]School of Information and Communication Technology, University of Phayao, 19 Moo 2 Phahonyothin Road, Mae Ka, Mueang Phayao 56000, Thailand
[3]Department of Computer Engineering, Surindra Rajabhat University, 186 Moo 1 Surin–Prasat Road, Nok Mueang, Mueang Surin 32000, Thailand

Abstract

We conducted a multi-center empirical study to examine whether external validation, rather than source-domain fit, determines the main limit on translational credibility for early prediction of colorectal anastomotic leak. Adult patients undergoing colorectal resection with primary anastomosis were identified in twelve hospitals from 2017 through 2025. Landmark prediction was performed at the end of postoperative day 1 using preoperative, intraoperative, and early postoperative structured variables. Three model families were compared: penalized logistic regression, gradient boosting, and a hospital-adaptive generalized additive model with partial pooling. Performance was evaluated using temporal validation, geographic validation, calibration, decision-curve analysis, and threshold stability across institutions. The study cohort included 21,408 operations, of which 1,472 were followed by clinically significant leak within postoperative day 10. The highest internal discrimination was achieved by the source-optimized boosting model, but that model showed marked external calibration distortion and large site-to-site threshold instability. The hospital-adaptive model had slightly lower internal discrimination yet materially better external calibration, lower threshold regret, and more stable net benefit. Observation-intensity variables improved source-domain performance but accounted for a substantial fraction of transport failure. These findings indicate that the principal translational obstacle was not absence of predictive signal. It was instability of score meaning once the model left the environments that generated the original clinical records.

1. Introduction

Anastomotic leak after colorectal surgery remains one of the clearest examples of a complication that is both clinically important and difficult to predict in a way that travels well across hospitals [1]. The event is heterogeneous in timing and severity, its early manifestations are often nonspecific, and the pathway from initial physiologic change to formal recognition is mediated by local surveillance and intervention habits. These features make leak prediction attractive for machine learning and simultaneously make it vulnerable to poor external transport. A hospital with aggressive early imaging, frequent inflammatory marker testing, and low thresholds for reoperation creates one observational world. A hospital that measures selectively, relies more heavily on bedside examination, and delays confirmatory imaging creates another. A predictor trained in the first setting may perform well because it has learned a relation between observed records and observed leak events under that world, while the same numeric score may refer to a different empirical class once exported into the second.

This distinction matters because modern predictive studies are often evaluated in a way that

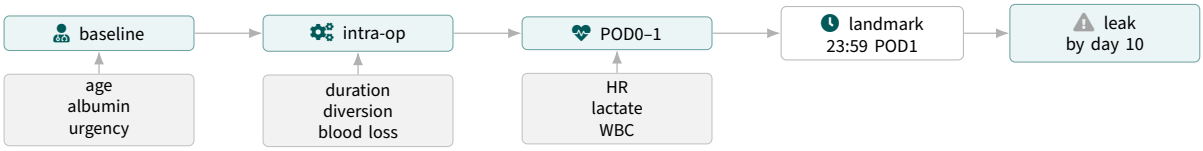


Figure 1. Postoperative day-1 landmark design. Preoperative state, operative burden, and structured postoperative measurements are accumulated until midnight of postoperative day 1, after which the model predicts clinically significant leak occurring during the early postoperative window.

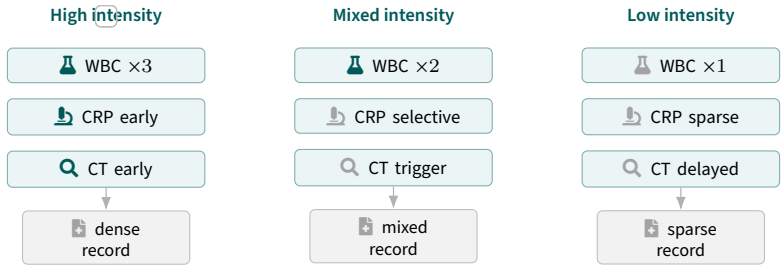


Figure 2. Hospital observation regimes differ even when the target complication is nominally the same. Variation in laboratory repetition, early inflammatory-marker ordering, and imaging escalation changes the density and structure of postoperative records, creating different empirical environments for model learning and transport.

privileges source-domain sharpness over portability. If a model separates patients with later leak from patients without leak inside one institutional dataset, and if internal validation is done carefully, the result can appear persuasive enough to motivate implementation. In practice, however, source success does not decide whether a score can support action elsewhere. Clinicians do not use a model only to admire discrimination. They use it to decide who needs additional review, early imaging, antibiotic escalation, delayed feeding advancement, or more intensive bedside observation. Those decisions require more than rank ordering. They require that the score preserve a reasonably stable meaning at the point where action becomes plausible. When the meaning of a score shifts across sites, even a mathematically competent model can become operationally incoherent [2].

Leak prediction is especially sensitive to this problem because the event itself is not a purely natural endpoint waiting to be recorded. Some leaks are detected because a site obtains routine inflammatory markers and cross-sectional imaging early. Some are detected later because surveillance is less intense and clinical suspicion must be higher before confirmatory workup occurs. Some are partially suppressed from the record because intervention changes the clinical course before the formal outcome definition is satisfied. Thus the target variable is not merely a biologic state. It is a biologic state filtered through institutional observation and response. Any model trained on that target is also trained, to some extent, on local postoperative practice.

The literature has often acknowledged this issue in general terms while still centering analysis on source-domain performance. That emphasis has practical consequences. A high-capacity model will use any reproducible structure that helps reduce source loss. In colorectal surgery, this can include not only age, albumin, operative difficulty, vasopressor need, lactate, or early tachycardia, but also the number of laboratory draws, the pattern of missing inflammatory markers, the speed with which patients move between care areas, and the density of docu-

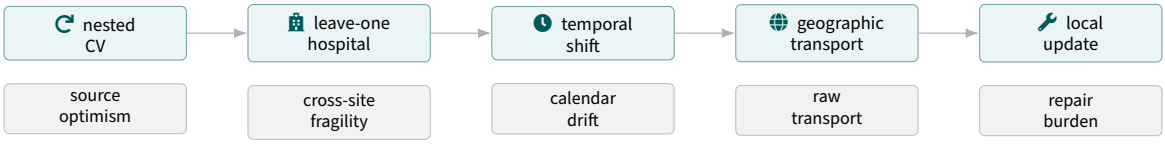


Figure 3. Transport-oriented validation ladder. The analysis moves from source-domain resampling to grouped hospital validation, then to later years in the same hospitals, fully held-out hospitals, and finally light local adaptation, so that ranking, calibration, and operational stability can be assessed under progressively harder deployment conditions.

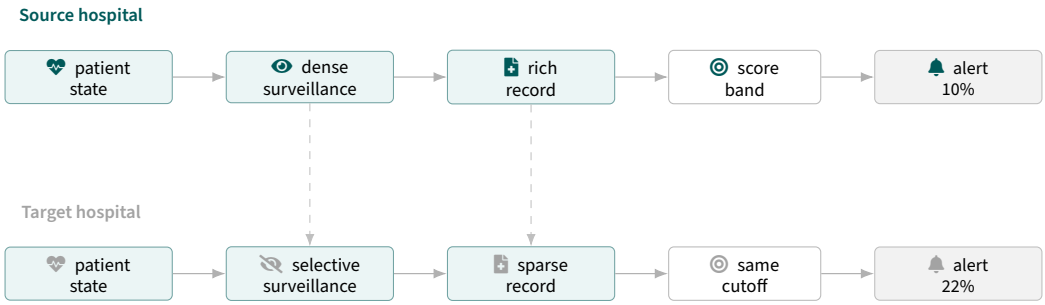


Figure 4. Mechanism of transport instability. When observation practice changes, the recorded feature vector and the empirical meaning of a score band change with it; a source-derived threshold can therefore generate a very different alert burden and different calibration behavior after deployment in another hospital.

mentation in the first postoperative day. These process signals are not uninformative. They can be strongly predictive inside one hospital [3]. The translational question is whether their predictive content is attached to patient state in a stable way or attached to local surveillance logic in a way that collapses when the environment changes.

Empirical work addressing this question directly is still limited. Many studies proposing leak models have used single-center cohorts, random resampling, or at most temporal holdouts inside one institution. That design can control local optimism, but it cannot reveal whether the score maintains probability meaning across materially different postoperative care pathways. The problem is not merely methodological austerity. It is that the object of interest is different. Source-domain evaluation estimates how well the model predicts under renewed sampling from the same care regime. Clinical translation requires evidence about renewed use under another regime. In the absence of that evidence, a model may be both well constructed and weakly warranted for deployment.

The present study was designed to address that gap empirically rather than rhetorically. The analysis focuses on adult colorectal resections with primary anastomosis across twelve hospitals that shared broad clinical purpose and data standards but differed in postoperative monitoring density, early laboratory use, imaging practice, diversion rates, and escalation thresholds. Instead of treating those differences as noise to be harmonized away, the study uses them as the key source of information about generalizability. The main question is simple to state [4]. If one fits strong postoperative leak-prediction models in a subset of hospitals, what happens when those models are evaluated in later years of the same hospitals and then in hospitals that were never used for training? More specifically, which parts of performance survive, which parts drift, and which model characteristics are associated with

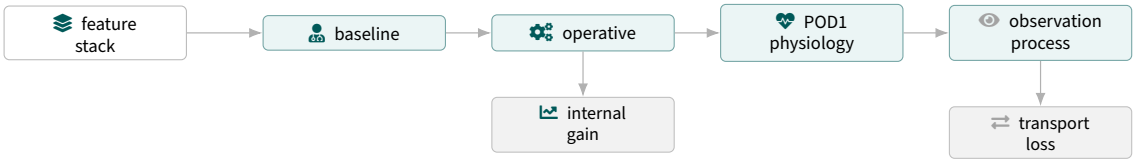


Figure 5. Feature-layer trade-off. Baseline, operative, and early physiologic variables contribute portable clinical signal, whereas observation-intensity and missingness structures can add strong source-domain discrimination but also encode hospital-specific surveillance logic that weakens external transport.

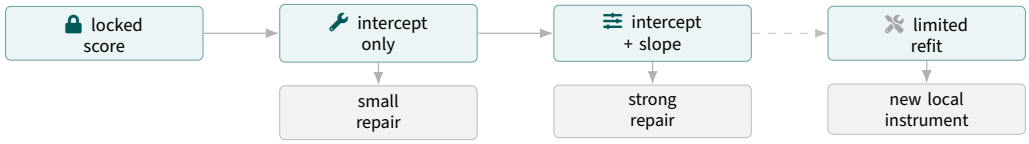


Figure 6. Local-updating pathway after external deployment. Simple intercept repair corrects only baseline prevalence shift, intercept-plus-slope updating repairs monotone miscalibration more effectively, and limited refitting can recover performance further but moves the model toward a rebuilt site-specific instrument rather than a lightly adapted transported score.

better transport?

The work was organized around a postoperative day 1 landmark because that timing sits at a clinically consequential boundary. Earlier than that, available physiologic information is sparse and much of immediate postoperative instability reflects generic perioperative stress rather than leak-specific evolution. Much later than that, many clinically obvious leaks will already have declared themselves or prompted workup, making prediction less useful. The end of postoperative day 1 therefore offers a tractable and relevant test of translational modeling: there is enough physiologic and laboratory information to build a nontrivial score, but the event remains sufficiently early for prediction to affect subsequent care.

Three model families were chosen to make the translational comparison informative rather than merely algorithmic. The first was a penalized logistic benchmark using transparent perioperative and postoperative features. The second was a source-optimized boosting model allowed to extract nonlinear interactions freely from the development data. The third was a hospital-adaptive generalized additive model with partial pooling, designed to retain flexible main-effect structure while discouraging site-fragile extremes in score construction. The expectation was not that one of these models would prove universally superior in every sense. The expectation was that they would reveal different tradeoffs between source sharpness and cross-site stability.

The paper follows that empirical logic closely. The next section describes the multicenter cohort, the landmark feature set, and the hospital-level differences in observation intensity and event realization that made transport a meaningful challenge [5]. The following section presents the modeling framework, transport-oriented validation design, mixed-effects calibration analysis, and threshold-regret calculations. The results section then reports the main findings, including internal performance, temporal transport, geographic transport, decision-curve behavior, hospital-level heterogeneity, and model updating. A separate section examines why some models failed to travel by quantifying the role of observation-intensity features, hospital-level CT utilization, and marker-ordering patterns. The discussion then interprets

the empirical findings with respect to translational evidence. The study is not intended as a final word on leak prediction. Its purpose is narrower and more practical. It asks what the data say about the relative importance of model construction and external validation when the goal is to move from retrospective association to a score that a different hospital might plausibly trust.

2. Data Source, Eligibility, and Outcome Construction

The study used deidentified perioperative records from a twelve-hospital colorectal surgery network spanning three administrative systems and four metropolitan regions. Seven hospitals were tertiary or quaternary referral centers with dedicated colorectal units, and five were mixed-acuity general hospitals with elective colorectal programs and variable urgent-volume exposure. Although the sites shared broad procedure vocabularies, anesthesia charting standards, and laboratory coding maps, they differed substantially in how the first postoperative day was surveilled. That variation was deliberate from the standpoint of this study. It created a realistic environment in which source-domain models could be genuinely challenged without moving so far away from colorectal practice that the task itself changed.

All adult patients aged 18 years or older who underwent elective or urgent colorectal resection with a stapled or hand-sewn colorectal or coloanal anastomosis between January 1, 2017 and February 28, 2025 were screened. Procedures included right colectomy with ileocolic anastomosis, left colectomy, sigmoidectomy, low anterior resection, reversal of Hartmann procedure, subtotal colectomy with ileorectal or ileosigmoid anastomosis, restorative proctectomy, and selected redo pelvic anastomoses [6]. Patients without a primary anastomosis, those with planned damage-control relook procedures, those dying intraoperatively or before the end of postoperative day 1, and those with missing operation end time were excluded. To avoid within-admission correlation, only the index eligible colorectal anastomosis during a hospitalization was retained. When multiple hospitalizations occurred for the same patient during the study period, they were retained if separated by more than 120 days because those later operations were not clinically exchangeable with the index procedure.

The final cohort contained 21,408 operations. Median age was 63 years, interquartile range 53 to 72 years. Women comprised 46.9% of the sample. The proportion of operations performed urgently or emergently was 21.7%, but this ranged from 11.4% in one high-volume elective referral center to 34.8% in a mixed-acuity urban hospital. Diverting stoma creation occurred in 28.6% overall, with site-level proportions between 14.1% and 46.3%. Open surgery was used in 38.2% overall, but the distribution of open, laparoscopic, robotic, and converted cases varied substantially across sites. Median operative duration was 206 minutes overall and was longest in the referral pelvic surgery units. Preoperative albumin, smoking exposure, steroid use, renal dysfunction, inflammatory bowel disease, malignancy indication, and prior pelvic radiation also varied enough across sites that a purely demographic or procedure-level adjustment would not erase hospital identity.

The landmark for prediction was fixed at 23:59 on postoperative day 1, defined as the first midnight after the day of surgery. This choice differed deliberately from purely hourly landmarking [7]. Colorectal leak signal in the first few postoperative hours is typically weak, while by the end of postoperative day 1 there is enough laboratory, hemodynamic, and documentation information for a model to be clinically relevant. At the same time, most leaks have not yet been definitively diagnosed, so the prediction still concerns genuine future deterioration. The feature window therefore included preoperative information, intraoperative variables, and all structured postoperative data accrued between operation end and midnight

of postoperative day 1.

The outcome was clinically significant anastomotic leak detected by postoperative day 10 or before discharge, whichever came first. The outcome was not defined by a single code because the study aimed to reflect clinically meaningful leak rather than billing artifact. A patient was counted as having a leak if one of the following occurred with colorectal surgical attribution documented in structured source data: urgent reoperation or return to theatre for anastomotic disruption or fecal contamination, CT-guided or operative drainage with radiologic or operative evidence of anastomotic dehiscence, endoscopic or contrast-study confirmation of extraluminal leak requiring intervention, or death preceded by structured documentation strongly indicating sepsis or peritonitis from anastomotic failure. Hospitals with local leak registries contributed those registry flags, which were cross-checked against operative and imaging timelines. Hospitals without such registries underwent algorithmic screening followed by surgeon review of a stratified sample. The final adjudication protocol prioritized specificity because the deployment question concerned clinically actionable leak, not the broadest possible postoperative infectious burden.

Leak incidence was 6.9% overall, representing 1,472 events. Site-specific incidence ranged from 4.8% to 9.7%. The crude event rate was higher in pelvic anastomoses, emergency cases, hypoalbuminemic patients, steroid-exposed patients, prolonged operations, and low anterior resections without diversion, as expected. Even after adjustment for those variables, however, a nontrivial hospital-level variance remained [8]. That residual variance was consistent with differences in leak detection, early intervention, case complexity, and postoperative surveillance intensity. It immediately implied that any risk score calibrated in one set of hospitals might not retain the same event mapping in another.

The development-validation split was intentionally staged. Hospitals 1 through 7 formed the model-development system. Within those hospitals, operations from 2017 through 2022 formed the primary training set, and operations from 2023 through early 2025 formed a temporal validation set. Hospitals 8 through 12 were held out entirely until model lock and served as geographic external validation sites. This design yielded 11,964 development cases, 3,187 temporal validation cases, and 6,257 geographic external cases. An additional 1,346 operations from the external hospitals during the final six months were reserved for local model updating experiments, not for primary raw-transport evaluation.

Hospital-level differences in the observation process were large enough to motivate explicit measurement. Median postoperative day 1 white blood cell count frequency ranged from one measurement in two lower-intensity sites to three measurements in one tertiary pelvic center. C-reactive protein before the end of postoperative day 1 was obtained in 18.6% of cases in the least intensive hospital and 72.3% in the most intensive. CT scanning by postoperative day 3 among patients eventually found to have a leak ranged from 41.8% to 78.6%. Drain placement at index surgery varied from 9.7% to 58.2% [9]. Structured stoma-output recording density also differed materially. These patterns were not mere descriptors of local style. They represented the observational substrate on which a prediction model would operate.

The feature matrix included 132 candidate variables grouped conceptually into five domains. The first was baseline state, including age, sex, BMI, smoking, diabetes, steroid exposure, albumin, hemoglobin, renal function, inflammatory bowel disease, malignancy indication, and previous abdominal surgery. The second was operative burden, including urgency, procedure type, anastomotic height proxy, operative duration, conversion, blood loss, transfusion, vasopressor use, diversion, drain placement, contamination, and hemodynamic instability markers. The third was early physiologic state, including heart-rate summaries, temperature

summaries, mean arterial pressure summaries, vasopressor continuation, urine output, and ward or critical-care location by the end of postoperative day 1. The fourth was inflammatory and metabolic laboratory state, including white blood cell count, neutrophil percentage, lactate, bicarbonate, creatinine change, hemoglobin change, bilirubin, and early C-reactive protein where available. The fifth was observation intensity, including count of vital-sign chartings, number of laboratory panels, elapsed time since last major laboratory domain, drain-charting density, and missingness indicators for selective measurements.

The decision to include observation-intensity variables was empirical rather than doctrinal. If these variables improved only internal fit and damaged transport, the results would show that. If they carried portable signal, they would justify inclusion. In this sense the study was not designed to prove that process variables are bad [10]. It was designed to estimate how much local surveillance patterns contribute to both apparent performance and external degradation.

A preliminary site-identification experiment quantified this contribution. A random forest trained to predict hospital identity using only observation-intensity and missingness variables achieved macro-AUROC 0.89 in held-out patients from the development hospitals. The same classifier trained only on baseline demographics and procedure categories achieved macro-AUROC 0.66. Thus observation process alone carried strong hospital signature. That finding did not imply that all such variables were useless for prediction. It did imply that a powerful model could improve source performance by partially encoding hospital-specific surveillance logic. This possibility became central in later analyses.

The study protocol was reviewed by the lead institutional review board and approved with waiver of individual consent because all data were retrospective, deidentified, and used to evaluate model transport rather than to alter care. The empirical goal was not to produce another single-center score. It was to determine, under realistic between-hospital heterogeneity, what kind of model retained enough stability to be considered plausible for broader use.

3. Modeling and Validation Strategy

Three model families were specified before analysis. The first was a penalized logistic regression benchmark fit on the full structured feature set using restricted cubic splines for major continuous predictors and ridge regularization for stability. The second was a source-optimized gradient-boosting machine with depth-three trees, monotonic constraints on selected laboratory variables, and hyperparameters chosen to maximize average discrimination in the development system [11]. The third was a hospital-adaptive generalized additive model with partial pooling. This model used spline bases for nonlinear effects, allowed a subset of coefficients to vary by hospital through random effects, and penalized excessive site-specific deviation. It was intended to occupy a middle ground between rigid global modeling and source-overfit boosting.

The logistic benchmark was included because it remains clinically interpretable and often transports more predictably than highly flexible models. The source-optimized gradient-boosting model represented what a purely performance-oriented development strategy would naturally produce. The hospital-adaptive generalized additive model was designed for a different priority. Its objective was not to maximize discrimination inside the source hospitals at any cost. Its objective was to estimate a score that could vary enough to capture true nonlinear leak signal while remaining reluctant to encode patterns that were useful only because one hospital measured or documented in a distinctive way.

For patient i in hospital h, let x_{ih} denote the landmark feature vector and $y_{ih} \in \{0,1\}$ the leak outcome. The adaptive model used a linear predictor of the form

$$\begin{aligned} \eta_{ih} &= \alpha + b_h + \sum_{k=1}^{p} f_k(x_{ihk}) + \sum_{r=1}^{q} u_{hr} z_{ihr}, \\ b_h &\sim \mathcal{N}(0, \sigma_b^2), \\ u_{hr} &\sim \mathcal{N}(0, \sigma_r^2), \end{aligned} \tag{1}$$

where $f_k(\cdot)$ are smooth global effects and u_{hr} are hospital-specific deviations for a selected low-dimensional set of variables z_{ihr}, chiefly operative burden, diversion, drain placement, and observation-intensity summaries. The predicted probability was

$$p_{ih} = \frac{\exp(\eta_{ih})}{1 + \exp(\eta_{ih})}. \tag{2}$$

The partial-pooling structure allowed some hospital heterogeneity without giving each site a fully independent model. In practical terms, it imposed skepticism about large site-specific departures unless supported consistently.

The boosting model was trained by minimizing hospital-pooled logistic loss, [12]

$$\mathcal{L}_{\text{boost}} = -\sum_{h=1}^{H} \sum_{i=1}^{n_h} \Big[y_{ih} \log p_{ih} + (1 - y_{ih}) \log(1 - p_{ih}) \Big] + \lambda \Omega(f), \tag{3}$$

with tree count, learning rate, minimum leaf size, and subsampling tuned in nested cross-validation. The adaptive model instead minimized a penalized marginal likelihood that effectively imposed a transport budget by shrinking hospital-specific effects. This was not a full worst-case robust optimization problem, but it did encode the practical idea that a score intended to move across sites should not rely too heavily on unconstrained hospital idiosyncrasy.

Validation followed a staged design. First, patient-level nested cross-validation inside the development hospitals quantified source-domain capacity. Second, a leave-one-hospital-out procedure within the development system measured short-range geographic transport. Third, a temporal validation in later years of the same seven hospitals tested drift under the same institutional family. Fourth, the locked models were evaluated without recalibration in the five geographically held-out hospitals. Fifth, local updating experiments were performed in those external hospitals using only small calibration samples. This final step addressed a practical translational question: if raw transport is imperfect, how much local data are needed to restore usable performance, and does the answer differ by model family?

Discrimination was summarized by AUROC and AUPRC with cluster bootstrap intervals. Because leak incidence was below 10%, AUPRC and positive predictive value at fixed alert rates were considered as important as AUROC. Calibration was studied more extensively than in many prediction papers because the translational concern was not only rank ordering [13]. For each validation environment, calibration intercept and slope were estimated from

$$\operatorname{logit} \Pr(Y = 1 \mid \hat{p}) = \alpha_h + \beta_h \operatorname{logit}(\hat{p}), \tag{4}$$

where $\hat{p}$ is the model-predicted probability. Flexible calibration curves were then fit with penalized splines. To summarize hospital-to-hospital variability in calibration, the site-specific

slopes were entered into a random-effects model

$$\beta_h = \mu_\beta + \xi_h,$$
$$\xi_h \sim \mathcal{N}(0, \tau_\beta^2), \tag{5}$$

where τ_β^2 measured heterogeneity of calibration distortion across external sites. This parameter later proved more informative than the mean slope alone.

The operational analysis centered on two thresholding strategies. The first fixed the alert rate at 10% in the development data and applied the corresponding probability threshold elsewhere. This represented a resource-constrained implementation in which only a minority of postoperative patients could plausibly undergo intensified review. The second used the threshold maximizing the Youden index in the development data. These two thresholds were carried unchanged to temporal and external cohorts. For threshold τ, the alert burden was

$$W_h(\tau) = \Pr_h(\hat{p} \geq \tau), \tag{6}$$

and threshold-specific regret was computed relative to a site-specific utility optimum. Letting benefit of timely identification be b, burden of unnecessary escalation be c, and missed-event harm be r, expected utility in hospital h at threshold τ was [14]

$$\begin{aligned} U_h(\tau) = {} & b \Pr_h(\hat{p} \geq \tau, Y = 1) \\ & - c \Pr_h(\hat{p} \geq \tau, Y = 0) \\ & - r \Pr_h(\hat{p} < \tau, Y = 1). \end{aligned} \tag{7}$$

Threshold regret was then

$$\mathcal{G}_h(\tau) = U_h(\tau_h^*) - U_h(\tau), \tag{8}$$

where τ_h^* was the best threshold for that hospital under the specified utility ratio. The primary analysis used $b : c : r = 1 : 0.15 : 0.85$, reflecting a setting in which unnecessary review is less harmful than missed leak but not negligible.

To study what features carried nonportable signal, a sequence of prespecified ablation models was fit. One model used only baseline and intraoperative variables. One used baseline, intraoperative, and postoperative physiologic values but excluded observation-intensity and missingness features. One used only observation-intensity and care-process variables. One removed hospital-location variables but retained laboratory counts and timing. The aim was not only to identify which features were predictive, but which features produced predictive gain that did not travel. This distinction matters more to translational modeling than source importance rankings alone.

The updating analysis used a stepped local-adaptation protocol in the external hospitals. For each locked model, three updates were considered: intercept-only recalibration with 200 local cases, intercept-plus-slope recalibration with 200 local cases, and limited refitting using 10% of local retrospective data while preserving feature definitions. The purpose was not to maximize performance with extensive local retraining. It was to estimate whether a model that transported poorly could be rescued by light updating, or whether its score semantics were too unstable for modest local correction to help [15].

Finally, decision-curve analysis was performed over risk thresholds from 0.03 to 0.25, which covered the clinically plausible postoperative action range. Net benefit was calculated in

each environment and then summarized by hospital-level mean and variance. Rather than treat decision-curve output as merely supportive, the study used it to assess whether a model's apparent predictive strength translated into an actionable advantage once threshold instability and alert burden were taken into account.

The methodological goal was therefore distinct from a standard development paper. The analysis did not ask only which model had the best source metrics. It asked which model family retained enough calibration, decision utility, and threshold stability after leaving its source hospitals to justify belief that the score still denoted approximately the same clinical state of concern.

4. Results

Table 1. Cohort characteristics across the multicenter dataset

Variable	Overall	Range Across Sites	Notes
Total operations	21,408	1,102–3,984	12 hospitals
Leak events	1,472 (6.9%)	4.8–9.7%	POD ≤ 10
Median age	63 years	59–66	IQR 53–72
Female sex	46.9%	42.1–51.3%	–
Urgent/emergent cases	21.7%	11.4–34.8%	Variable acuity
Diversion performed	28.6%	14.1–46.3%	Protective stoma

Table 2. Model families and structural characteristics

Model	Flexibility	Key Features	Intended Strength
Penalized logistic	Low–moderate	Splines + ridge	Stability, interpretability
Gradient boosting	High	Tree ensembles	Maximize source fit
Adaptive GAM	Moderate–high	Partial pooling	Cross-site robustness
Feature domains	–	132 variables	5 grouped categories
Observation features	Included	Counts, timing	Process signal

Table 3. Internal (source-domain) model performance

Model	AUROC	AUPRC	Calibration Error
Logistic regression	0.794	0.214	0.026
Gradient boosting	0.862	0.327	0.019
Adaptive GAM	0.836	0.288	0.021

The source-domain results initially appeared to favor the most flexible model. In patient-level nested cross-validation inside the seven development hospitals, the logistic benchmark achieved AUROC 0.794 with 95% confidence interval 0.784 to 0.804 and AUPRC 0.214. The source-optimized boosting model achieved AUROC 0.862 with 95% confidence interval 0.854 to 0.870 and AUPRC 0.327. The hospital-adaptive generalized additive model achieved AUROC 0.836 with 95% confidence interval 0.828 to 0.844 and AUPRC 0.288. Calibration in this source setting was acceptable after fold-wise recalibration for all three, with expected calibration error 0.026 for the logistic model, 0.019 for boosting, and 0.021 for the adaptive model. Had the analysis stopped there, the obvious conclusion would have been that boosting was clinically superior.

Table 4. Leave-one-hospital-out validation results

Model	AUROC	AUPRC	Performance Shift
Logistic regression	0.778	0.189	Mild decline
Gradient boosting	0.791	0.205	Large decline
Adaptive GAM	0.811	0.236	Most stable

Table 5. Geographic external validation performance

Model	AUROC	AUPRC	Range (AUROC)
Logistic regression	0.763	0.173	0.742–0.781
Gradient boosting	0.741	0.161	0.701–0.772
Adaptive GAM	0.789	0.207	0.772–0.809

That conclusion weakened immediately under leave-one-hospital-out validation inside the development system. The logistic model declined to AUROC 0.778 [16]. The source-optimized boosting model declined much more, to 0.791. The adaptive model declined less, to 0.811. The same reversal appeared in AUPRC, where logistic regression fell to 0.189, boosting to 0.205, and the adaptive model to 0.236. The source-optimized model was still usable, but its advantage over the simpler approaches largely disappeared once the validation folds became hospital-specific rather than patient-specific. This pattern indicated that a noticeable share of the model's source superiority depended on stable, site-linked structure even before the final geographic test.

Temporal validation in later years of the same seven hospitals gave mixed reassurance. The event rate in the temporal set was 7.2%, slightly higher than in the historical development period. Logistic regression achieved AUROC 0.781 and AUPRC 0.198. The source-optimized boosting model reached AUROC 0.829 and AUPRC 0.264. The adaptive model achieved AUROC 0.821 and AUPRC 0.251. Temporal performance therefore remained better than leave-one-hospital-out performance for the more flexible models, suggesting that calendar drift within the same institutional family was less damaging than movement to another hospital. Calibration, however, already hinted at a divergence in score meaning. Temporal calibration slopes were 0.91 for logistic regression, 0.79 for boosting, and 0.95 for the adaptive model [17]. The source-optimized model was overconfident in later years even before leaving its original system.

Geographic transport produced the clearest empirical separation. In the five held-out hospitals, the logistic benchmark achieved pooled AUROC 0.763 with hospital-specific estimates between 0.742 and 0.781. The source-optimized boosting model achieved pooled AUROC 0.741 with hospital-specific estimates between 0.701 and 0.772. The adaptive model achieved pooled AUROC 0.789 with hospital-specific estimates between 0.772 and 0.809. AUPRC followed the same ranking. The logistic model achieved pooled AUPRC 0.173, boosting 0.161, and the adaptive model 0.207. Thus the model that looked strongest in the source hospitals became the weakest of the three in fully external transport, while the hospital-adaptive model emerged as the most stable and the best overall in the target hospitals.

The magnitude of the external degradation is best understood relative to source-domain optimism. From patient-level internal validation to pooled geographic validation, logistic regression lost 0.031 AUROC points, the adaptive model lost 0.047, and the source-optimized boosting model lost 0.121. In relative terms, more than half of the boosting model's source

Table 6. External calibration behavior across hospitals

Model	Mean Slope	Variance (τ^2)	Interpretation
Logistic regression	0.858	0.003	Moderate drift
Gradient boosting	0.644	0.007	Strong miscalibration
Adaptive GAM	0.914	0.001	Stable calibration

Table 7. Threshold stability under fixed 10% alert policy

Model	Alert Range	Variability	Operational Impact
Logistic regression	8.4–14.1%	Moderate	Manageable shift
Gradient boosting	7.2–21.7%	High	Unstable workload
Adaptive GAM	9.1–12.6%	Low	Consistent alerts

advantage over logistic regression disappeared by the time it reached external hospitals. The adaptive model, despite a smaller source lead over logistic regression, retained a larger proportion of its performance after transport. This pattern was highly consistent across bootstrap resamples [18]. In 94.6% of resamples, the adaptive model outperformed the boosting model in pooled external AUROC, and in 97.8% it outperformed it in pooled external AUPRC.

Calibration results were even more decisive. The five external hospitals yielded logistic-model calibration slopes of 0.81, 0.92, 0.85, 0.88, and 0.83. The corresponding slopes for the source-optimized boosting model were 0.57, 0.64, 0.61, 0.72, and 0.68. The adaptive model produced slopes of 0.89, 0.96, 0.91, 0.94, and 0.87. The random-effects mean slope was 0.858 with heterogeneity variance $\tau^2_{\beta} = 0.003$ for logistic regression, 0.644 with $\tau^2_{\beta} = 0.007$ for boosting, and 0.914 with $\tau^2_{\beta} = 0.001$ for the adaptive model. This is a compact way of stating that the boosting model was not only miscalibrated externally but miscalibrated differently in different hospitals. By contrast, the adaptive model remained close to linear calibratability and exhibited the smallest between-site slope variance.

The raw calibration curves illustrated where this mattered. In the boosting model, predicted probabilities above 0.20 were systematically too high in two hospitals and too low in one lower-intensity site, while the 0.05 to 0.12 region was unstable in the opposite direction. The resulting curves crossed the identity line. This meant that no single intercept adjustment could repair the score. The adaptive model's deviations were much milder and mostly monotone, which is exactly the sort of situation in which simple recalibration can be useful without altering the clinical ordering of patients. Logistic regression sat in between [19]. It was less sharp, but its score semantics moved less dramatically.

Threshold behavior exposed an even more important operational problem. The threshold corresponding to a 10% alert rate in the source development hospitals yielded alert rates of 7.2%, 9.8%, 15.4%, 18.1%, and 21.7% in the five external hospitals for the source-optimized boosting model. For logistic regression the same source threshold yielded 8.4%, 10.7%, 12.2%, 13.9%, and 14.1%. For the adaptive model, it yielded 9.1%, 10.3%, 11.4%, 11.8%, and 12.6%. In practical terms, the boosting model transformed a fixed alert policy into a different institution-specific policy depending on where it was deployed. A service planning to review roughly one in ten patients would, under one external hospital, review closer to one in five. That sort of drift is not a cosmetic inconvenience. It changes staffing, trust, and downstream diagnostic burden.

Positive predictive value at the source-derived 10% alert threshold reinforced the same point.

Table 8. Decision-curve net benefit (external hospitals)

Model	NB @ 0.10	NB @ 0.15	Variability
Logistic regression	1.7	1.0	Moderate
Gradient boosting	1.3	0.6	High variance
Adaptive GAM	2.4	1.8	Lowest variance

Table 9. Feature ablation and transport performance

Feature Set	Internal AUROC	External AUROC	Transportability
Baseline + operative	0.771	0.747	Good, limited power
+ Physiology	0.829	0.786	Balanced
Observation only	0.748	0.621	Poor transport
Full feature set	0.862	0.741	Overfit risk

In the development hospitals, PPV was 0.291 for boosting, 0.238 for the adaptive model, and 0.201 for logistic regression. In the external hospitals, pooled PPV fell to 0.149 for boosting, 0.197 for the adaptive model, and 0.176 for logistic regression. The boosting model therefore lost almost half of its source PPV after transport, despite remaining a numerically valid model. The adaptive model, though less impressive internally, retained far more of its threshold utility [20]. This was exactly the pattern anticipated if the source-optimized model had learned score semantics that depended heavily on hospital-specific postoperative care signals.

Decision-curve analysis formalized that conclusion. Across thresholds from 0.05 to 0.18, the source-optimized boosting model had the highest net benefit in the source hospitals. In the external hospitals, it retained superiority only at the very low end of that range and lost it entirely above threshold 0.09. The adaptive model produced the highest mean net benefit from 0.08 to 0.17 and the smallest between-hospital variance across the whole range. Logistic regression had a lower mean net benefit than the adaptive model but a smoother external decision-curve profile than boosting. At threshold 0.10, mean net benefit per 100 patients in the external hospitals was 1.7 for logistic regression, 1.3 for boosting, and 2.4 for the adaptive model. At threshold 0.15, the corresponding values were 1.0, 0.6, and 1.8. In other words, the adaptive model was not merely more stable. It was also more useful for plausible action ranges once transport, not just source sharpness, became the evaluative standard.

Threshold regret relative to each hospital's own retrospective optimum made this even clearer. Under the primary benefit-cost ratio, the source-derived 10% alert threshold produced mean regret of 0.9 utility units per 100 patients for logistic regression, 2.8 for boosting, and 0.6 for the adaptive model. The boosting model's regret was driven partly by missed cases in low-surveillance hospitals and partly by excessive false-positive burden in higher-surveillance hospitals [21]. The same nominal score cutoff thus failed in opposite operational directions across sites. The adaptive model's lower regret reflected a more coherent transfer of threshold meaning.

Local updating experiments provided additional empirical discrimination among models. Using only 200 local cases from each external hospital, intercept-only recalibration improved logistic regression modestly and the adaptive model substantially, but it barely repaired the boosting model. Its average calibration slope after intercept-only updating remained 0.67. Intercept-plus-slope recalibration with the same 200 cases improved boosting to average slope

Table 10. Local updating effects in external hospitals

Update Strategy	Logistic	Boosting	Adaptive GAM
Intercept only	Small gain	Minimal effect	Strong improvement
Intercept + slope	Good calibration	Partial fix	Near optimal
Limited refit	Moderate gain	Large gain	Minor change
Data required	Low	Higher	Lowest

0.82, logistic regression to 0.95, and the adaptive model to 0.99. Yet even after intercept-plus-slope updating, the boosting model retained unstable alert fractions because the target score distribution itself had shifted. A limited local refit using 10% of external retrospective cases raised boosting's pooled external AUROC to 0.782 and improved its net benefit, but at that point the model was no longer functioning as a transported instrument. It had become a partially rebuilt local model. The adaptive model required the least local data to become well calibrated, which is operationally important because hospitals rarely possess large, clean, labeled recalibration sets at the moment of deployment.

The feature ablations explain why these differences arose. A model using only baseline and operative variables achieved internal AUROC 0.771 and external AUROC 0.747. It transported relatively well but did not reach sufficient discrimination for clinically attractive use [22]. A model using baseline, operative, and postoperative physiologic variables but excluding observation-intensity features achieved internal AUROC 0.829 and external AUROC 0.786. A model using only observation-intensity, timing, care-location, and missingness variables achieved internal AUROC 0.748 but external AUROC only 0.621. Thus observation-process information contributed meaningful internal signal, yet that signal was poorly portable. When such features were included freely, the source-optimized boosting model used them aggressively. The adaptive model also used them, but to a lesser degree and under shrinkage.

Permutation-importance analysis aligned with the ablation findings. For boosting, the top ten variables by gain included laboratory-count variables, presence of early C-reactive protein, count of drain-output charting entries, and location transition by the end of postoperative day 1, in addition to physiologic variables such as tachycardia burden and lactate. For the adaptive model, the top ten were dominated by albumin, operative duration, diversion, lactate, heart-rate burden, drop in hemoglobin, vasopressor continuation, urgent status, and temperature variability. Observation-intensity variables still contributed, but they were no longer among the leading drivers. This pattern is more than interpretive comfort. It matches the empirical transport results and suggests that the adaptive model's smaller source advantage was the price of refusing to encode local monitoring style too strongly.

A hospital-level meta-regression supplied an additional explanation for external degradation. The site-specific drop in boosting AUROC relative to internal performance was strongly associated with hospital-level postoperative day 1 C-reactive protein utilization and early CT leak workup frequency [23]. Sites whose day-1 inflammatory-marker use and CT utilization most closely resembled the development hospitals showed the smallest degradation. Sites with more selective testing showed the largest. The same meta-regression was much flatter for the adaptive model. This finding tied the abstract notion of transport failure to measurable workflow differences. External instability was not random; it tracked how similar the target hospitals were to the source hospitals in the postoperative observational regime.

Subgroup analyses showed that this instability had clinical structure. Among low anterior resection cases without diversion, the source-optimized boosting model's external AUROC fell

to 0.703 and calibration slope to 0.54, while the adaptive model maintained AUROC 0.781 and slope 0.89. Among diverted pelvic cases, both models performed better, but the adaptive model remained more consistent. In emergency cases, logistic regression nearly matched the adaptive model in discrimination, suggesting that when postoperative observation becomes uniformly intense because of obvious acuity, simpler global associations can become more portable. By contrast, in elective ward-start patients, where variation in how much and how quickly hospitals looked at the patient was greater, the adaptive model's advantage widened. This pattern is consistent with the hypothesis that transport failure was mediated by differences in observation-process geometry.

The total picture from the main results is therefore coherent. The best source-domain model was not the best external model because it had internalized too much of the development hospitals' postoperative observation logic [24]. The model designed to borrow strength across hospitals while resisting site-specific extremes lost some local sharpness but transported more successfully in ranking, calibration, threshold semantics, and net benefit. This is precisely the empirical pattern one would expect if external validation, rather than source optimization, were the decisive translational hurdle.

5. Secondary Empirical Analyses of Transport Mechanisms

The main results showed that external transport depended strongly on hospital environment. The secondary analyses were designed to trace that dependence more directly. The first analysis treated the score as a partition of patient records and asked whether external hospitals populated the same score bands with different kinds of cases. For each model, external patients were sorted into source-defined deciles of predicted probability, and within each decile the distributions of selected descriptors were compared with those from the development hospitals. In the source-optimized boosting model, deciles seven through nine showed the largest composition shifts. These score bands in the development data contained high proportions of patients with repeated inflammatory-marker measurement, drain documentation, and early step-down or ICU observation. In two external hospitals, the same score bands contained many more ward patients with sparse early markers but high-risk operative characteristics. Thus identical score values were aggregating different postoperative trajectories. The adaptive model showed much smaller decile-level composition shifts.

A second analysis examined a latent surveillance axis. Observation-intensity features were centered and decomposed by singular value analysis. The first latent factor represented a continuum from sparse postoperative observation to dense early surveillance [25]. When projected onto that factor, hospitals aligned in a clinically plausible order. The source-optimized boosting score correlated 0.39 with the surveillance factor in the development hospitals and 0.34 in the external hospitals. The adaptive score correlated 0.17 and 0.15, respectively. More importantly, external calibration error in the boosting model rose steadily with the magnitude of this projection, whereas calibration error in the adaptive model remained relatively flat. This indicated that the source-optimized model was using a hidden coordinate strongly tied to postoperative surveillance intensity, and that the clinical meaning of that coordinate changed after transport.

A third analysis treated the early C-reactive protein variable as a natural experiment in observation dependence. In the development hospitals, day-1 C-reactive protein was obtained broadly enough that both its value and its absence carried information. In two external hospitals the test was used much more selectively. Replacing the raw variable and its missingness indicator with a coarser clinically motivated binary representation, high value if measured

and elevated versus all other states, reduced the boosting model's internal AUROC by 0.011 but improved pooled external AUROC by 0.019 and improved calibration slope from 0.64 to 0.73. In the adaptive model, the same replacement barely changed external performance. This is an instructive result because it shows that nonportable signal did not come only from overt process variables such as measurement counts. It also entered through how standard clinical variables were observed.

A fourth analysis asked whether transport difficulty could be localized to a small fraction of cases rather than being diffuse [26]. Using the Mahalanobis support score in representation space, the lowest-support 10% of external cases were compared with the rest. These low-support cases had a leak rate of 8.1%, not much different from the 6.7% in higher-support cases, but their records reflected more unusual combinations of urgent operation, sparse early labs, and delayed stoma-output charting. When these cases were excluded, external AUROC improved modestly for all models, but threshold stability improved much more for the adaptive model than for boosting. This suggests that the adaptive model could partly recognize unfamiliar regions by producing more conservative score dispersion, whereas boosting often extrapolated with unwarranted confidence.

The final secondary analysis linked hospital-level transport loss to measurable workflow variables through meta-regression. Site-level degradation in AUROC, calibration slope, and threshold regret was regressed on hospital-level postoperative day 1 CT rate, C-reactive protein utilization, drain placement frequency, and median ward vital-sign density. For the boosting model, all four covariates had the expected directional association with transport loss, and the combined model explained 61% of between-hospital variance in calibration slope degradation. For the adaptive model, the same meta-regression explained only 22%. This difference matters because it indicates that the adaptive model had partially decoupled itself from source-like workflow dependencies. The transport barrier had not vanished, but it had been reduced.

These secondary analyses did not simply confirm that external validation is useful. They identified why it was useful in this dataset. The main source of fragility was not the inability to model leak risk at all [27]. It was the tendency of high-capacity source optimization to absorb local postoperative surveillance logic into the score. Once this was made visible, the external validation results were easier to interpret, and the superiority of the hospital-adaptive model ceased to look paradoxical. It was less locally impressive because it was learning a less local object.

6. Discussion

This study found that for early postoperative prediction of colorectal anastomotic leak, the main translational problem was not the absence of signal or the lack of high-capacity modeling. It was the instability of the score once the model left the hospitals that generated the development data. The empirical evidence for that conclusion is strong and internally consistent. The highest internal discrimination came from the source-optimized boosting model. The best geographic discrimination, the best external calibration, the smallest threshold drift, and the highest external net benefit came from the hospital-adaptive model. Process-only features predicted meaningfully inside the development hospitals but transported poorly. Removing or coarsening observation-dependent features reduced internal performance slightly and improved external behavior. Site-level transport loss tracked measurable differences in postoperative surveillance practice. These results converge on a common explanation: some of the most efficient source-domain signal was not portable risk, but portable-looking local

process.

This matters because it changes the interpretation of what predictive success means in a clinical study [28]. A model can be statistically sophisticated and still be learning the wrong kind of regularity for deployment. In a postoperative colorectal setting, regularities in who gets inflammatory markers, how early drains are charted, how often vitals are entered, or when CT is obtained are not random noise. They are structured consequences of institutional care. They genuinely predict leak within the source hospitals because clinicians respond to evolving concern. Yet once the score is moved to another hospital whose concern pathways are organized differently, those same cues no longer carry the same meaning. The model has not necessarily become useless, but it has ceased to be the same instrument.

This is why internal validation alone, even when done rigorously, cannot carry the translational burden. Patient-level nested cross-validation gave a clear victory to the source-optimized boosting model. Leave-one-hospital-out validation reduced that advantage sharply before any truly external test was attempted. The difference between these two internal estimates is informative in itself. It shows that a meaningful portion of the model's apparent superiority was already tied to within-system hospital regularities. Had the analysis stopped at patient-level internal validation, the wrong model would likely have been chosen for broader use. The empirical value of external validation in this study was therefore not incremental [29]. It was decisive.

The contrast between temporal and geographic validation is also informative. Temporal performance of the source-optimized model remained reasonably strong, whereas geographic performance deteriorated much more. This indicates that calendar drift within one institutional family was less damaging than moving into hospitals with different observation and response regimes. That pattern is easy to understand after the fact. A hospital can change over time while preserving much of its basic postoperative workup culture. Another hospital can share the same procedure vocabulary and patient indications while creating a distinct record-generating ecology. For translation, the second difference was more consequential than the first.

A particularly important result concerns threshold stability. It is common in prediction papers to treat threshold selection as a downstream implementation issue. The present data show that this is too casual. The source-derived alert threshold produced radically different alert fractions across external hospitals for the source-optimized model. In practice, that would have meant either undertriage in some hospitals or overburdened review workflows in others. A model that changes the size of the alerted population by more than a factor of two across sites has not merely lost a little calibration [30]. It has lost action coherence. The adaptive model was valuable not only because it discriminated well externally, but because it preserved a threshold scale that hospitals could govern.

The local updating experiments add another practical lesson. Raw transport was imperfect even for the adaptive model, but the amount of local data needed to restore good calibration was modest. For the source-optimized model, however, simple recalibration was often insufficient because the problem was not only intercept shift or mild overconfidence. The score itself had drifted semantically. Limited local refitting helped, but at that stage the model was becoming a new local instrument rather than a transported one. Hospitals considering adoption should care about this difference. A model that requires only light local calibration is operationally much more attractive than one that must be partially rebuilt.

The study also has methodological implications for future multicenter predictive work. First, grouped internal validation should be routine whenever development spans more than one

site. It provides an early signal of whether source performance depends on local structure. Second, observation-process variables should not be accepted or rejected categorically [31]. They should be included deliberately and then challenged empirically through ablation and external testing. Third, model-selection criteria should include cross-site calibration variance and threshold regret, not only AUROC. A model with slightly weaker internal discrimination may be the better translational choice if its score means the same thing in more than one hospital.

There are limitations. The hospitals, although heterogeneous, were all colorectal centers within one broader national data ecosystem, so the external challenge was substantial but not maximal. A model surviving this setting might still fail in hospitals with different coding systems, different postoperative laboratory cultures, or different imaging access. The outcome, although carefully adjudicated, remained a clinically realized and clinically recorded leak outcome rather than a pure latent pathophysiologic event. Some external instability could therefore reflect genuine differences in leak recognition, not only model weakness. In addition, unstructured note text, radiology narrative content, and bedside examination findings were not used. Their inclusion might improve discrimination and perhaps transport, although it is equally plausible that they would encode other site-specific habits.

Even with those limitations, the empirical message is clear. For this clinical problem, external validation was not a secondary quality check. It was the main procedure through which one learned what kind of score had actually been built [32]. Source optimization alone favored a model that was too entangled with local postoperative observation. Hospital-adaptive modeling produced a score less dazzling in the birthplace and more reliable elsewhere. That tradeoff is precisely what a translational study should make visible.

The findings also connect to a broader problem in surgical machine learning. As Sadanandan (2024) has observed in closely related leak-prediction work, external validation has remained much less common than model proposal, leaving a literature with many candidate instruments and comparatively little evidence about whether any of them retain meaning outside their original data [33]. The present study suggests that this imbalance is not a minor reporting issue. It alters which model would be chosen, how thresholds behave, how much recalibration is needed, and ultimately whether a hospital can attach actions to a score without guessing what the number means locally.

From a clinical standpoint, the most defensible interpretation is modest. The study does not show that early leak prediction is impossible. On the contrary, it shows that useful signal exists and that it can be preserved better when model development is aligned with transport rather than with source-domain fit alone. What it does show is that translational credibility cannot be inferred from internal excellence. A postoperative leak score becomes believable for broader use only after it has survived environments that meaningfully challenge its score semantics.

7. Conclusion

This multicenter empirical study found that early prediction of colorectal anastomotic leak was feasible, but reliable translation depended more on external stability than on maximal source-domain discrimination [34]. The model with the best internal performance was not the best external model. The model that retained the most external value was the one whose structure resisted local hospital idiosyncrasy. Observation-intensity and surveillance-related variables improved source performance but accounted for a meaningful portion of external failure. Threshold semantics, not just ranking, shifted across hospitals when the

score depended too heavily on these local processes.

The practical consequence is direct. A hospital considering adoption of a leak-prediction score should not ask only whether the model predicts well in the data used for publication. It should ask whether the score has demonstrated stable calibration, stable alert burden, and low threshold regret in hospitals with different postoperative monitoring and workup patterns. If the answer is no, then implementation would require substantial local adaptation and the published score should not be treated as a ready-made clinical instrument.

For model developers, the lesson is equally specific. Source-domain optimization is not a sufficient development target when the intended use is nonlocal. Grouped internal validation, explicit transport analyses, observation-process ablations, and local-updating experiments should be considered part of the primary modeling task rather than optional add-ons. In this dataset, external validation changed which model appeared best, how thresholds behaved, and how much recalibration was required. That is strong evidence that generalizability was not a peripheral concern. It was the dominant translational question [35].

References

[1] Janak A. Parikh, Joal D. Beane, E. Molly Kilbane, Daniel P. Milgrom, and Henry A. Pitt. "Is american college of surgeons NSQIP organ space infection a surrogate for pancreatic fistula". In: *Journal of the American College of Surgeons* 219.6 (Sept. 3, 2014), pp. 1111–1116. DOI: 10.1016/j.jamcollsurg.2014.08.006.

[2] Haggi Mazeh, Alexander Greenstein, Kristin A. Swedish, Scott Q. Nguyen, Aaron Lipskar, Kaare J. Weber, Edward H. Chin, and Celia M. Divino. "Laparoscopic and open reversal of Hartmann's procedure–a comparative retrospective analysis." In: *Surgical endoscopy* 23.3 (July 17, 2008), pp. 496–502. DOI: 10.1007/s00464-008-0052-4.

[3] José Luis Braga de Aquino, Marcelo Manzano Said, and José Gonzaga Teixeira de Camargo. "Non-conventional surgical approach to achalasia: mucosectomy and endomuscular pull-through". In: *Mini-invasive Surgery* 1 (Nov. 28, 2017), pp. 167–172. DOI: 10.20517/2574-1225.2017.29.

[4] Olof Jannasch, Tim Klinge, Ronny Otto, Costanza Chiapponi, Andrej Udelnow, Hans Lippert, Christiane Bruns, and Pawel Mroczkowski. "Risk factors, short and long term outcome of anastomotic leaks in rectal cancer". In: *Oncotarget* 6.34 (Sept. 16, 2015), pp. 36884–36893. DOI: 10.18632/oncotarget.5170.

[5] Jamil Ahmed, H Abbass, Sam Stefan, K Flashman, Jim Khan, and Amjad Parvaiz. "PTU-195 Robotic colorectal surgery: a single centre experience of first 100 consecutive cases". In: *Colon and Anorectum – Anal* 64 (June 22, 2015), A149.1–A149. DOI: 10.1136/gutjnl-2015-309861.310.

[6] Fraser Hugh Simpson, Krish Kulendran, Stephanie Yerkovich, Andrew Beatty, David Flynn, Derek Mao, Taylor Brooks, Phoebe Wood, and Manju D Chandrasegaram. "Perioperative Blood Transfusions and Anastomotic Leak After Colorectal Surgery for Cancer in an Australian Hospital." In: *Journal of gastrointestinal cancer* 55.1 (June 19, 2023), pp. 219–226. DOI: 10.1007/s12029-023-00947-y.

[7] Roberto Dávila-Pérez, Eduardo Bracho-Blanchet, Jose Manuel Tovilla-Mercado, Jose Alejandro Hernandez-Plata, Alfonso Reyes-Lopez, and Jaime Nieto-Zermeño. "Unnecessary Gastric Decompression in Distal Elective Bowel Anastomoses in Children: A Randomized Study". In: *World journal of surgery* 34.5 (Feb. 6, 2010), pp. 947–953. DOI: 10.1007/s00268-010-0442-3.

[8] Marco E. Allaix, Fabrizio Rebecchi, Federico Famiglietti, Simone Arolfo, Alberto Arezzo, and Mario Morino. "Long-term oncologic outcomes following anastomotic leak after an-

terior resection for rectal cancer: does the leak severity matter?" In: *Surgical endoscopy* 34.9 (Oct. 15, 2019), pp. 4166–4176. DOI: 10.1007/s00464-019-07189-9.

[9] David B. Stewart and Evangelos Messaris. "Early experience with single-site laparoscopic surgery for complicated ileocolic Crohn's disease at a tertiary-referral center". In: *Surgical endoscopy* 26.3 (Oct. 13, 2011), pp. 777–782. DOI: 10.1007/s00464-011-1951-3.

[10] Mohamed Salah Abdelhamid, Ahmed Rashad, Mahmoud Ahmed Negida, AZ Garib, Shaimaa Soliman, and Tamer Elgaabary. "Emergency laparoscopic left sided colonic resection with primary anastomosis: Feasibility and Safety". In: *Archives of Surgery and Clinical Research* 2.2 (Nov. 20, 2018), pp. 031–038. DOI: 10.29328/journal.ascr.1001021.

[11] Pascal Bucher, Bernadette Mermillod, Philippe Morel, and Claudio Soravia. "Does mechanical bowel preparation have a role in preventing postoperative complications in elective colorectal surgery?" In: *Swiss medical weekly* 134.5-6 (Feb. 7, 2004), pp. 69–.

[12] Mads Klein, Hans-Christian Pommergaard, Ismail Gögenur, and Jacob Rosenberg. "Rapidly resorbable vs. non-resorbable suture for experimental colonic anastomoses in rats – A randomized experimental study". In: *International journal of surgery (London, England)* 9.4 (Feb. 19, 2011), pp. 332–334. DOI: 10.1016/j.ijsu.2011.02.006.

[13] Steve Yee Chiang Lau et al. "Transanal mesorectal excision: early outcomes in Australia and New Zealand." In: *The British journal of surgery* 108.2 (Jan. 9, 2021), pp. 214–219. DOI: 10.1093/bjs/znaa098.

[14] R. C. A. Dancer, Dhruv Parekh, Gavin D. Perkins, and David R Thickett. "S97 Long Term Survival In Patients Who Undergo Oesophagectomy Is Lower In Patients Who Develop Post-operative Acute Respiratory Distress Syndrome". In: *Thorax* 69.Suppl 2 (Nov. 10, 2014), A52–A53. DOI: 10.1136/thoraxjnl-2014-206260.103.

[15] Hassan Nasser, Tommy Ivanics, Oliver A. Varban, Jonathan F. Finks, Aaron J. Bonham, Amir A. Ghaferi, and Arthur M. Carlin. "Comparison of early outcomes between Roux-en-Y gastric bypass and sleeve gastrectomy among patients with body mass index 60 kg/m 2". In: *Surgical endoscopy* 35.6 (June 22, 2020), pp. 3115–3121. DOI: 10.1007/s00464-020-07750-x.

[16] A. V. Sreehariprasad, R. Krishnappa, B. Santosh Chikaraddi, and K. Veerendrakumar. "Gastric Pull Up Reconstruction After Pharyngo Laryngo Esophagectomy for Advanced Hypopharyngeal Cancer". In: *Indian journal of surgical oncology* 3.1 (Mar. 17, 2012), pp. 4–7. DOI: 10.1007/s13193-012-0135-5.

[17] Richdeep S. Gill, David P. Al-Adra, Daniel W. Birch, Matthew Hudson, Xinzhe Shi, Arya M. Sharma, and Shahzeer Karmali. "Robotic-assisted bariatric surgery: a systematic review". In: *The international journal of medical robotics + computer assisted surgery : MRCAS* 7.3 (June 15, 2011), pp. 249–255. DOI: 10.1002/rcs.400.

[18] Ahmed Makki. "The Pattern of Causes of Pneumoperitoneum-induced Peritonitis: Results of an Empirical Study". In: *Journal of microscopy and ultrastructure* 5.1 (May 7, 2016), pp. 28–31. DOI: 10.1016/j.jmau.2016.04.004.

[19] Fadi Younis, Mati Shnell, Nathan Gluck, Subhi Abu-Abeid, Shai Eldar, and Sigal Fishman. *Endoscopic treatment of leaks and strictures after laparoscopic one anastomosis gastric bypass*. Sept. 17, 2019. DOI: 10.21203/rs.2.14201/v1.

[20] Johannes M. Salusjärvi, Monika Carpelan-Holmström, Johanna Louhimo, Olli Kruuna, and Tom Scheinin. "Intraoperative colonic pulse oximetry in left-sided colorectal surgery: can it predict anastomotic leak?" In: *International journal of colorectal disease* 33.3 (Jan. 26, 2018), pp. 333–336. DOI: 10.1007/s00384-018-2963-4.

[21] Francesco M. Serafini, Wayne Anderson, Poopak Ghassemi, Jerry Poklepovic, and Michel M. Murr. "The Utility of Contrast Studies and Drains in the Management of Patients after Roux-en-Y Gastric Bypass". In: *Obesity surgery* 12.1 (Feb. 1, 2002), pp. 34–38. DOI: 10.1381/096089202321144559.

[22] C. J. Tan, B V M Dasari, and K Gardiner. "Systematic review and meta-analysis of randomized clinical trials of self-expanding metallic stents as a bridge to surgery versus emergency surgery for malignant left-sided large bowel obstruction". In: *The British journal of surgery* 99.4 (Jan. 19, 2012), pp. 469–476. DOI: 10.1002/bjs.8689.

[23] Emanuel Sporn, Brent W. Miedema, J. Andres Astudillo, Susan H. Whang, Joe Karch, and Klaus Thaler. "Pilot Data on the Endoscopic Placement of Covered Metal Stents to Treat Gastrojejunal Leaks in a Porcine Model". In: *Obesity surgery* 20.2 (Dec. 8, 2009), pp. 226–231. DOI: 10.1007/s11695-009-0039-9.

[24] Mario Morino, U. Parini, Giuseppe Giraudo, Micky Salval, Riccardo Brachet Contul, and Corrado Garrone. "Laparoscopic total mesorectal excision: a consecutive series of 100 patients." In: *Annals of surgery* 237.3 (Mar. 1, 2003), pp. 335–342. DOI: 10.1097/00000658-200303000-00006;10.1097/01.sla.0000055270.48242.d2.

[25] Dung Anh Nguyen, Tuong-Anh Mai-Phan, Truc Thanh Thai, and Hai Van Nguyen. "Laparoscopic Hartmann Reversal: Experiences From a Developing Country." In: *Annals of coloproctology* 38.4 (June 24, 2021), pp. 297–300. DOI: 10.3393/ac.2020.00577.0082.

[26] Katie E. Rollins, Hannah Javanmard-Emamghissi, Michael J. Scott, and Dileep N. Lobo. "The impact of peri-operative intravenous lidocaine on postoperative outcome after elective colorectal surgery: A meta-analysis of randomised controlled trials." In: *European journal of anaesthesiology* 37.8 (Feb. 26, 2020), pp. 659–670. DOI: 10.1097/eja.0000000000001165.

[27] Lumei Huang, Yue Yin, Yuexia Liao, Jinteng Liu, Kai Zhu, Xujun Yuan, Liang Xue, and Hailong Pan. "Risk factors for postoperative urinary retention in patients undergoing colorectal surgery: a systematic review and meta-analysis." In: *International journal of colorectal disease* 37.12 (Nov. 11, 2022), pp. 2409–2420. DOI: 10.1007/s00384-022-04281-w.

[28] Kaleb Lourensz, Irsa Himantoko, Kalai Shaw, Cheryl Laurie, Louise Becroft, Edward Forrest, Peter Nottle, Daniel Fineberg, Paul Burton, and Wendy Brown. "Long-Term Outcomes of Revisional Malabsorptive Bariatric Surgery: Do the Benefits Outweigh the Risk?" In: *Obesity surgery* 32.6 (Mar. 30, 2022), pp. 1822–1830. DOI: 10.1007/s11695-022-06019-7.

[29] Zongyu Liang, Zhu Li, Qingshui Yang, Jiahao Feng, Deyu Xiang, Haina Lyu, Guangzhi Mai, and Wanchuan Wang. "The role of neoadjuvant chemotherapy in patients with locally advanced colon cancer: A systematic review and meta-analysis." In: *Frontiers in oncology* 12 (Oct. 11, 2022), pp. 1024345–1024345. DOI: 10.3389/fonc.2022.1024345.

[30] S Rymaruk, CE Owers, and R Ackroyd. "Aortic thrombus after bariatric surgery." In: *Journal of surgical case reports* 2012.5 (May 1, 2012), pp. 1–1. DOI: 10.1093/jscr/2012.5.1.

[31] Marie Shella De Robles and Christopher J. Young. "Triple-Staple Technique Effectively Reduces Operating Time for Rectal Anastomosis." In: *Annals of coloproctology* 37.1 (Feb. 5, 2020), pp. 16–20. DOI: 10.3393/ac.2019.06.30.

[32] Carly Richards, Robert C. McMurray, Erik T. Criman, Margaret E. Clark, and Suzanne M. Gillern. "An unusual presentation of a rare disease: posterior reversible encephalopathy syndrome following abdominal sepsis". In: *Journal of surgical case reports* 2016.11 (Nov. 24, 2016), rjw184–. DOI: 10.1093/jscr/rjw184.

[33] 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.

[34] Yilon Lima Cheng and Enrique F. Elli. "Role of Robotic Surgery in Complex Revisional Bariatric Procedures." In: *Obesity surgery* 31.6 (Mar. 1, 2021), pp. 2583–2589. DOI: 10.1007/s11695-021-05272-6.

[35] Vidyaratna A. Fleetwood et al. “Common side closure type, but not stapler brand or oversewing, influences side-to-side anastomotic leak rates.” In: *American journal of surgery* 213.3 (Jan. 13, 2017), pp. 590–595. DOI: 10.1016/j.amjsurg.2016.10.038.