scisearch INTERNATIONAL PEER-REVIEWED JOURNALS

SCISEARCH

RESEARCH PAPER

An Empirical Study of External Validation Failure and Transport Stability in Postoperative Clinical Prediction Models

Zheng Wei[1] and Guo Liang[2]

[1]School of Computer Science and Technology, Hunan University of Science and Engineering, 519 Yongzhou Avenue, Lengshuitan District, Yongzhou 425199, Hunan, China
[2]Department of Information Engineering, Yancheng Institute of Technology, 9 Yingbin Avenue, Tinghu District, Yancheng 224051, Jiangsu, China

Abstract

Background conditions in clinical machine learning are now favorable for model development and still unfavorable for dependable translation. Hospitals capture large amounts of perioperative, physiologic, laboratory, and administrative information, allowing investigators to train increasingly sophisticated risk models. Yet these data are generated inside care systems that differ in surveillance intensity, intervention timing, documentation structure, and event ascertainment. A model that performs well within one such system may therefore encode a relationship that is only partly portable. This study examined that problem empirically in a multi-hospital cohort of major abdominal surgery patients and asked whether external validation, rather than source-domain optimization, determined the main limit on translational credibility. We assembled a retrospective cohort from seven hospitals, derived a structured 6-hour postoperative feature set, trained three risk models in the development hospitals, and evaluated temporal transport, geographic transport, threshold stability, and silent live operation in external sites. The primary outcome was major postoperative decompensation within 72 hours, defined as unplanned intensive care transfer, septic shock, urgent reoperation for intra-abdominal source control, or death. Models were compared using discrimination, calibration, workload distortion, and threshold regret. Source-optimized boosting achieved the best internal discrimination but showed pronounced external calibration loss and highly unstable alert rates. A domain-regularized model had slightly lower internal performance yet materially better external stability. An ablation analysis demonstrated that observation-process features contributed strongly to internal performance but transported poorly. Silent live operation revealed additional degradation caused by feature latency and backfilled laboratory values. These findings indicate that the primary translational bottleneck was not absence of predictive signal, but failure of score meaning to remain stable outside the environment that generated the training data.

1. Introduction

Predictive modeling in modern clinical datasets is often easier than explaining what exactly a successful model has learned (1). In perioperative and acute care settings, rich streams of laboratory values, medication records, device readings, operative descriptors, and time-stamped clinical events allow investigators to fit highly expressive models with relatively little difficulty. Once those models return respectable internal discrimination, the temptation is to conclude that the remaining work is mostly practical: deploy the model in the electronic record, determine who should see the output, and perhaps recalibrate if needed. That sequence is misleading because the central translational problem is already present before the model ever reaches a dashboard. A risk score trained inside one hospital may or may not represent the same empirical object when used in another. Internal success does not decide that question.

The difficulty begins with the fact that hospitals do not merely record patient states. They produce records through organizational behavior. A laboratory result appears because the test was ordered. A repeat measurement appears because something in the patient, the protocol, or the workflow triggered another look. A transfer event, a medication infusion, or a gap in documentation often reflects decision-making as much as pathophysiology. A feature matrix assembled from those records therefore contains both state information and response information. The target label is similarly shaped by care. A complication can be detected early because surveillance is intense, prevented because action was prompt, or documented differently because coding and review rules differ. A model fit to these data does not learn from an abstract clinical reality; it learns from a local arrangement of observation, response, and recording.

This local arrangement can be highly stable inside one site. That stability is exactly what makes internal validation insufficient for translation. Cross-validation, bootstrap correction, and temporal holdouts within the same institution all operate inside largely similar measurement and intervention regimes. They provide good estimates of how reproducible the fitted relation is under renewed sampling from the same local world. They do not answer whether the relation survives a change in hospital, staffing model, surveillance cadence, or postoperative workup culture (2). A model can therefore be internally sound and externally fragile for reasons that ordinary source-domain rigor cannot expose.

The issue becomes especially acute when a model is intended not merely to describe retrospective patterns, but to guide real-

world action. In that setting, a scalar risk score is not simply a compact summary of the information available in the chart. It functions, or is expected to function, as a candidate control signal within a clinical system. Once decision-makers place an operational threshold on that score, the output becomes linked to concrete institutional responses. Patients above the threshold may be routed to additional review, closer bedside monitoring, diagnostic imaging, specialist consultation, escalation of nursing attention, or earlier therapeutic intervention. Patients below the threshold may not receive those same resources. The score therefore does more than represent estimated risk; it helps determine how attention, time, and clinical capacity are allocated. For that reason, the meaning of the score must remain sufficiently stable across environments for the resulting actions to remain sensible.

This requirement is often more demanding than it first appears. A thresholding policy implicitly assumes that a given numeric value carries approximately comparable significance wherever the model is used. If a score of 0.20 in the development hospital corresponded to a particular combination of physiology, documentation density, laboratory timing, and surveillance intensity, then deploying that same threshold elsewhere assumes that a score of 0.20 in the target hospital represents roughly the same latent state of concern. Yet this assumption can easily fail. Hospitals differ not only in patient mix, but in how they observe patients, how quickly they document changes, which tests are ordered by default, and how aggressively teams respond to early signs of deterioration. These differences shape the informational substrate from which the score is constructed. As a result, the same numerical output may be generated under meaningfully different observational conditions.

Consider the case in which a score of 0.20 at the development hospital partly reflected a dense surveillance process. In that environment, frequent postoperative vital sign checks, early lactate testing, rapid nursing escalation, and standardized order sets may have made subtle signs of decompensation visible early in the record. A score of 0.20 there may therefore have represented a patient whose measured data already incorporated a relatively rich picture of emerging instability. If the target hospital operates with less intensive early surveillance, slower documentation, or more selective testing, then a score of 0.20 may arise from a thinner and less informative record. Numerically, the score is identical. Operationally, however, its meaning may be quite different. The model may still output numbers on the same formal scale, but the relationship between those numbers and the clinical situations they are supposed to summarize may no longer be preserved.

This is where the practical danger lies. Once a threshold is treated as a trigger for action, incoherence in score meaning translates into incoherence in workflow. A cutoff that was calibrated to produce a manageable review volume in the development hospital may overwhelm clinicians in the target site, or conversely may fail to identify enough truly high-risk patients to justify the intervention program. The threshold no longer represents the same effective risk, and it no longer induces the same operational burden. What appears to be a stable decision rule at the level of software may become an unstable policy at the level of care delivery. The institution may believe it has imported a validated triage mechanism, when in fact it has imported only the numeric shell of that mechanism without the surrounding observational ecology that originally made the numbers actionable.

This problem also clarifies why external validation alone is not always sufficient for deployment readiness. A model may preserve acceptable discrimination when moved to a new setting, yet still fail as an action-guiding instrument if the score-to-decision mapping has lost coherence. In other words, the question is not only whether the model can still rank patients reasonably well, but whether the same threshold still supports the same clinical commitments. Actionability depends on that alignment. A risk score can remain mathematically well-defined while becoming operationally ambiguous.

For this reason, the claim that a model generalizes should be treated as an empirical statement about score meaning, not as a routine extrapolation from good source performance. The relevant question is not simply whether discrimination decreases in a new dataset. The more consequential question is whether the model's score still partitions patients into groups that have similar outcome relations, similar calibration, and similar action consequences outside the development environment. When that does not hold, the translational barrier is not a small technical defect. It is the absence of evidence that the score remains the same clinical instrument after transport.

The present paper addresses that problem empirically using a retrospective multi-hospital design with an additional silent live-operation phase. Rather than building yet another model and reporting only internal performance, the analysis treats external validation as the main scientific object. The data come from seven hospitals performing major abdominal surgery under materially different postoperative monitoring and documentation regimes. Three candidate models were trained: a penalized logistic model, a source-optimized gradient-boosting model, and a domain-regularized boosting model designed to reduce cross-hospital instability. Performance was then measured in development hospitals, in a temporal holdout, in geographically external hospitals, and finally under silent live operation after model freeze. The outcome of interest was a clinically consequential postoperative decompensation composite within 72 hours (3). The broader purpose was not only to compare algorithms, but to test whether the translational bottleneck arose from inadequate predictive signal or from instability of transport.

The paper proceeds in a conventional empirical sequence but is guided by a more specific hypothesis. Section three describes the clinical setting, cohort construction, and hospital-level heterogeneity in observation and outcome pathways. Section four introduces the modeling strategy, domain-regularized objective, and threshold-regret framework used to evaluate deployment stability. Section five presents the empirical findings, including source performance, external failures, calibration drift, threshold distortion, and the behavior of simpler versus more complex models. Section six analyzes why transport failed where it did, focusing on the contribution of observation-process features and site-specific surveillance structure. Section seven reports

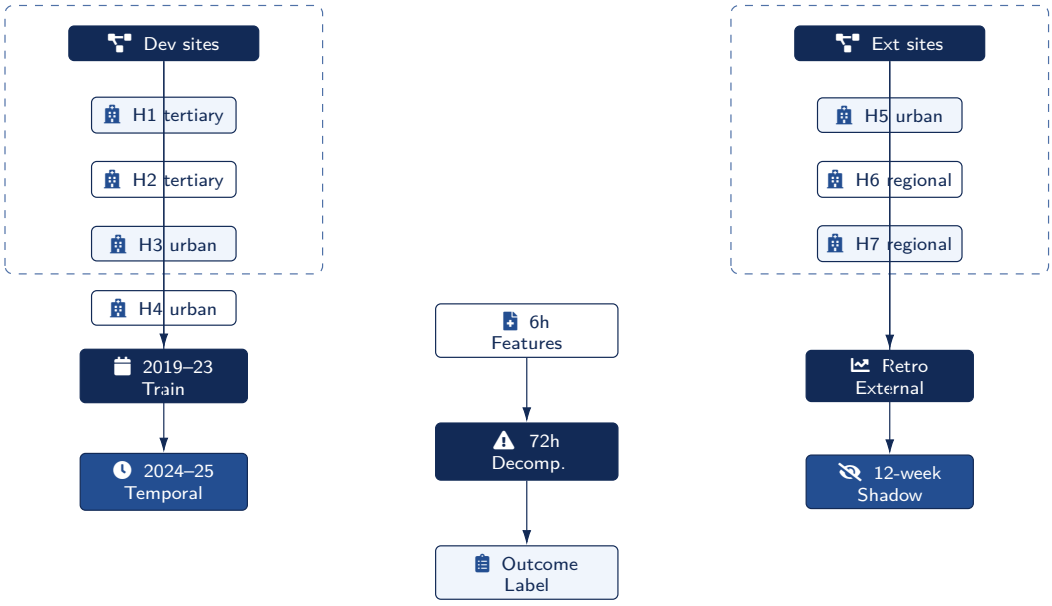

Figure 1. Study network and partitioning. Four hospitals form the development environment, split into a primary training era and a later temporal holdout, while three hospitals remain unseen until geographic external validation and a subsequent shadow deployment phase. All analyses are anchored to a structured 6-hour postoperative feature window and a 72-hour composite decompensation label, making transport rather than source-only fit the organizing axis of the empirical design.

the silent live-operation experiment and shows how production data latency introduced a second layer of degradation not visible in retrospective replay. The conclusion returns to the central translational question. A clinically useful predictor does not merely need to work where it was built. It must preserve enough meaning under environmental change that hospitals can attach actions to its output without guessing what the score has become.

2. Study Setting and Cohort Assembly

The empirical study used data from a seven-hospital surgical network that included two large tertiary referral centers, three urban general hospitals, and two regional community hospitals. All seven sites used structured perioperative documentation, shared the same broad coding vocabulary for procedures and diagnoses, and exported encounter-level data into a common analytic warehouse. They did not, however, function as interchangeable sites. The tertiary centers cared for more complex oncologic and inflammatory bowel disease cases, obtained postoperative laboratory studies at higher frequency, and admitted a larger fraction of patients to intermediate or intensive care immediately after major procedures. The community hospitals cared for lower-acuity mixes overall, used more selective postoperative testing, and escalated to intensive monitoring later and less often. These institutional differences were not treated as nuisance metadata. They were part of the empirical problem.

The cohort included adult patients who underwent major abdominal surgery between January 1, 2019 and March 31, 2025. Eligible procedures comprised colorectal resections, small bowel resections, gastric resections, ventral and parastomal abdominal wall reconstructions involving intra-abdominal entry, pancreatobiliary resections, and urgent laparotomies requiring an anastomosis or bowel discontinuity. Patients younger than 18 years, transplant recipients, admissions lacking intraoperative timestamps, and encounters with documented comfort-focused care before the 6-hour postoperative decision time were excluded (4). To reduce dependence among observations, only the first eligible procedure per hospitalization was retained, and repeat hospitalizations by the same patient within 90 days were excluded. After exclusions, the retrospective cohort consisted of 93,926 operations among 90,441 patients.

The primary decision time was fixed at 6 postoperative hours. This was chosen because the study aimed to approximate a clinically actionable early postoperative risk estimate rather than a purely preoperative or fully postoperative retrospective score. The feature window therefore included preoperative data, intraoperative data, and all structured postoperative measurements available by 6 hours after skin closure. Features that were first observed after the decision time were not included in the retrospective models, with one important exception addressed later in the silent live-operation analysis: some variables that appeared to be available by 6 hours in warehouse extracts were revealed in production to rely partly on delayed backfill. That discrepancy became central in the deployment analysis.

The outcome was major postoperative decompensation within 72 hours after the 6-hour decision point. The outcome was defined as the occurrence of any of the following: unplanned ICU transfer from a lower-acuity postoperative location, septic shock requiring vasopressors with presumed or documented intra-abdominal source, urgent reoperation or interventional source-control procedure for intra-abdominal complication, or death. The composite was intentionally broad because the translational question concerned early risk stratification for

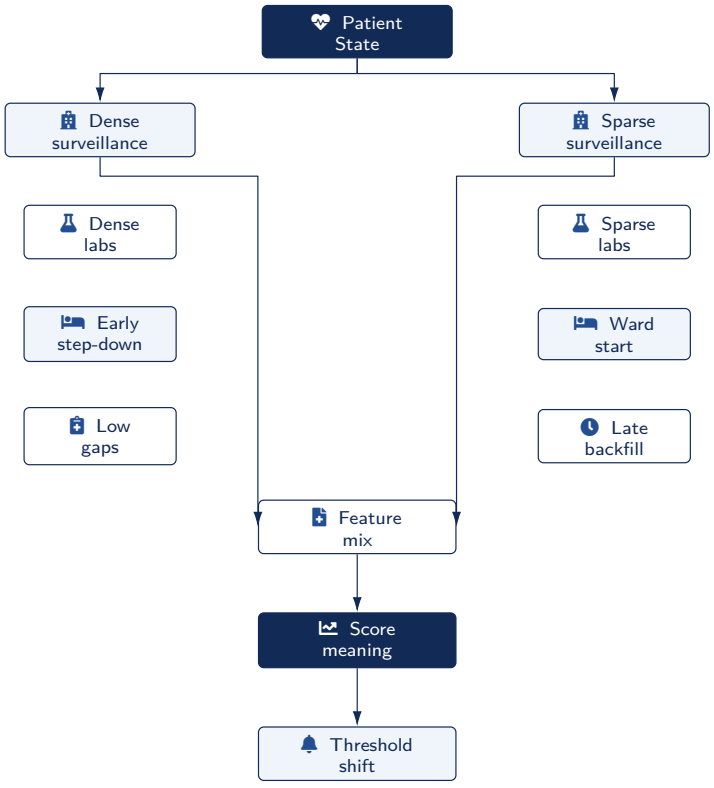

Figure 2. Observation-process heterogeneity. The same latent postoperative state can be translated into very different structured records depending on local surveillance intensity, unit-placement rules, and charting latency. Dense measurement, early high-acuity placement, and low missingness create one feature geometry, while sparse measurement and delayed backfill create another. The central consequence is not merely different covariates, but a change in what the score means near clinically relevant decision thresholds.

clinically meaningful deterioration rather than prediction of one narrowly coded complication. Outcome adjudication used a hybrid structured-labeling procedure across the seven hospitals. Because procedural source-control events and unplanned ICU transfer were consistently recorded, these components were highly reliable. Septic shock components required time-aligned vasopressor and culture or infection documentation, and these were more sensitive to documentation differences across sites. Death was complete in all hospitals.

Retrospective data were partitioned in a way that forced genuine transport tests. Hospitals 1 through 4 were used for development and internal validation. Within those four hospitals, encounters from 2019 through 2023 formed the primary development set and encounters from 2024 through March 2025 formed a temporal holdout set. Hospitals 5 through 7 were held out entirely until the final external evaluation and silent live-operation phases. This design yielded 58,214 development encounters, 10,947 temporal-holdout encounters, and 18,663 geographically external retrospective encounters (5). A further 6,102 operations from hospitals 5 and 6 during a later 12-week period were reserved for silent live operation after the final model had been frozen.

The event rate varied materially across sites. In the development hospitals, major postoperative decompensation occurred in 8.8% of cases overall, but site-specific incidence ranged from 7.1% in the lower-acuity urban hospital to 10.6% in the tertiary colorectal center. The temporal holdout event rate was 9.1%. Among the external hospitals, incidence was 7.4%, 9.8%, and 11.7%, respectively. These differences were not explained solely by age or procedure distribution. When a case-mix adjustment using age, emergency status, operative duration, preoperative albumin, ASA class, and procedure category was performed, a nontrivial between-hospital variance remained, suggesting that observed incidence also reflected surveillance, rescue, and documentation differences.

The observation process differed sharply across hospitals. Median postoperative complete blood count sampling frequency within the first 24 hours ranged from 1.3 draws per patient in one community site to 3.8 in the tertiary center. Lactate was measured in 31% of patients within 6 hours at the lowest-intensity site and in 79% at the highest-intensity site. Immediate postoperative ICU or step-down placement ranged from 12% to 34%. The median count of structured nursing flowsheet entries between 0 and 6 hours varied by almost threefold. Even after conditioning on procedure category and emergency status, these differences persisted. In other words, the feature matrix was site-specific not only because patients differed, but because surveillance differed.

The feature set was deliberately broad but fully structured. Eighty-six candidate variables were assembled before modeling. These included age, sex, BMI, smoking status, diabetes, heart failure, chronic kidney disease, preoperative laboratory

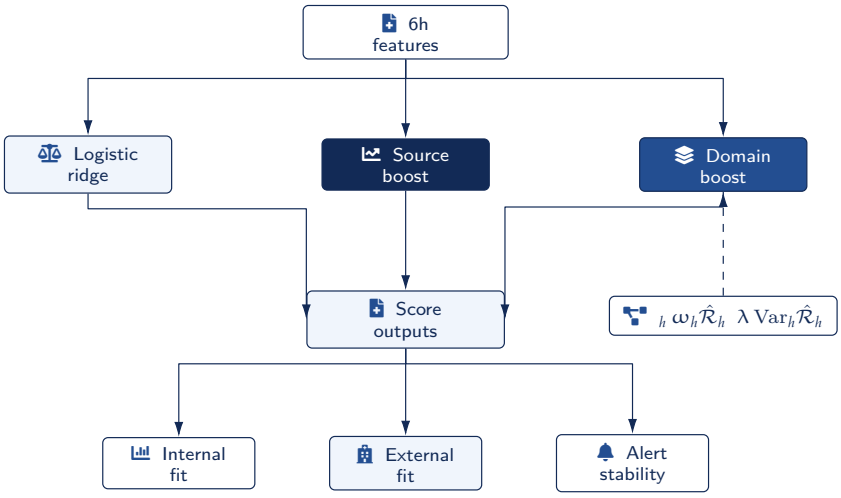

Figure 3. Model families and transport-aware objective. The logistic baseline offers stable linear structure, the source-optimized booster maximizes local sharpness, and the domain-regularized booster adds an explicit penalty on cross-hospital loss variance. This changes the optimization target from source discrimination alone to a joint balance among fit, cross-site stability, and downstream threshold governability.

values, steroid exposure, operative urgency, operative duration, blood loss, transfusion, stoma creation, vasopressor use intraoperatively, postoperative heart rate summaries, blood pressure summaries, temperature, lactate, white blood cell count, hemoglobin, urine output, early ward or ICU location, and measurement-process variables such as counts of recorded vitals, counts of laboratory draws, and time since last observation for major physiologic domains. Missingness indicators were retained for selected postoperative laboratories because the meaning of nonmeasurement was clinically informative in preliminary analyses. Continuous variables were winsorized at the 0.5th and 99.5th percentiles, then transformed with restricted cubic splines in the logistic model and left in raw standardized form for boosting.

The decision to include measurement-process variables was empirical rather than ideological (6). In a source-domain setting, such variables often improve performance because they capture clinical concern. The risk, of course, is that they are also tightly linked to local surveillance habits. Rather than assume one way or the other, the study retained them for some models and removed them in specific ablations. This allowed direct estimation of how much internal performance depended on process signals and how much those signals harmed external transport.

The rationale for a multi-hospital design was informed by the observation that related surgical machine-learning literatures have produced many source-local models but little convincing evidence of transport, with only 4 of 22 previously reviewed studies in one closely related domain performing external validation, as noted by Sadanandan (2024) (7). The present study therefore treated hospital heterogeneity as the main source of information rather than as an inconvenience to be normalized away.

Before predictive modeling, descriptive analyses were used to assess whether the hospitals were distinguishable from record structure alone. A multinomial classifier trained only on measurement-intensity variables, timing variables, and missingness indicators achieved a macro-F1 of 0.78 for hospital identification in the development data, whereas the same classifier trained only on demographics and major comorbidities achieved macro-F1 of 0.36. This result was important because it showed that observation process variables encoded substantial site identity. It did not prove that such variables were nontransportable, but it did indicate that any model relying heavily on them would likely be entangled with hospital-specific surveillance structure.

This setting created an empirically useful tension. The hospitals were similar enough to make transport a meaningful clinical ambition. They shared broad perioperative practice patterns, comparable major procedure categories, and a common data warehouse. At the same time, they differed enough in monitoring, escalation, and outcome realization that transport failure was plausible. A model that survived this setting would have earned more than source credibility. A model that failed in this setting would reveal that the translational barrier emerged even before one ventured into more dramatic cross-system heterogeneity.

3. Modeling Strategy and Statistical Analysis

Three primary models were fit. The first was a penalized logistic regression with spline-expanded continuous terms, pairwise interactions between selected operative and physiologic features, and ridge regularization tuned by nested cross-validation. The second was a source-optimized gradient-boosting model using shallow decision trees, monotonic constraints for selected laboratory variables, and early stopping tuned to maximize mean development-hospital area under the receiver operating characteristic curve (8). The third was a domain-regularized gradient-boosting model fit with the same tree depth and learning-rate family as the second model but trained under an objective that penalized cross-hospital instability in loss. The intention was not to produce an exhaustive algorithmic contest. It was to

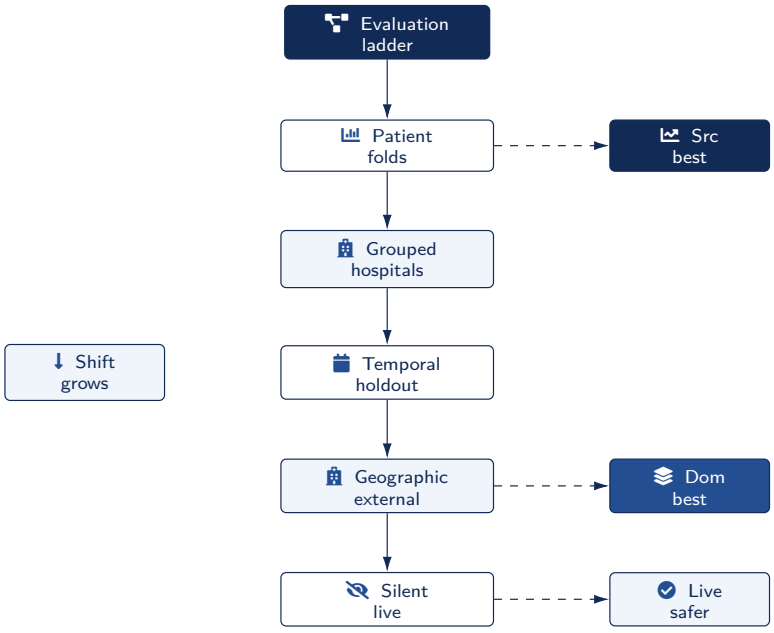

Figure 4. Validation ladder. The empirical story changes as the evaluation regime moves from patient-level resampling to grouped hospitals, later eras, unseen hospitals, and finally silent live execution. The ladder highlights the central result of the study: the model that looks strongest under source-local resampling is not the model that retains the most interpretable score behavior once transport pressure and operational realism increase.

compare an interpretable baseline, a locally optimized nonlinear model, and a nonlinear model deliberately discouraged from becoming too site-specific.

For hospital h and model f, empirical logistic loss was denoted by $\hat{\mathcal{R}}_h f$. The domain-regularized boosting model minimized

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{h=1}^{H} \omega_h \hat{\mathcal{R}}_h f + \lambda \operatorname{Var}\left(\hat{\mathcal{R}}_1 f, \ldots, \hat{\mathcal{R}}_H f\right) + \mu \Omega f, \tag{3.1}$$

where ω_h were hospital-size weights normalized to sum to one, λ controlled the penalty on cross-hospital loss variance, and Ωf was a standard tree-complexity penalty. The practical effect of this objective was to sacrifice a small amount of development-hospital sharpness in exchange for more even performance across the hospitals used in training. Hyperparameters λ, μ were selected by nested leave-one-hospital-out validation inside the development set.

Because the study was motivated by transport rather than source-only fit, internal evaluation did not rely on random folds alone. Inside the development hospitals, the primary internal estimate used a grouped cross-validation scheme that held out one hospital at a time while fitting on the remaining three. This is stricter than random resampling because it tests short-range geographic transport even before the final external evaluation. A second internal estimate used conventional patient-level nested folds to quantify local discriminative capacity. The difference between these two internal estimates was itself informative. If a model did well in random folds and markedly worse in grouped folds, that suggested dependence on local hospital structure already inside the development network.

Calibration was assessed in three ways. First, a global calibration slope and intercept were estimated by fitting

$$\operatorname{logit} \Pr Y = 1 \mid S, h = \alpha_h \ \beta_h S, \tag{3.2}$$

within each environment, where S was the logit score for logistic regression and the logit-transformed predicted probability for boosting models. Second, expected calibration error was computed on deciles of predicted probability. Third, a smooth calibration curve using penalized splines was examined, with particular attention to the 0.10 to 0.30 score range because that interval later proved operationally relevant for threshold analysis. Model recalibration was not performed before the primary external results; the first pass focused on raw transport. A secondary analysis then applied intercept-only and intercept-plus-slope recalibration separately at each external hospital (9).

Threshold behavior was treated as a first-class outcome rather than a downstream implementation detail. The development hospitals were used to select a nominal action threshold corresponding to approximately 80% specificity in the source-trained boosting model, as well as a second threshold chosen to maximize F_1 in the development set. These thresholds were then carried unchanged into the temporal and geographic validation sets. For any environment h and threshold τ, action rate was

$$W_h \tau = \Pr_h \left(S \geq \tau\right). \tag{3.3}$$

Threshold utility was computed under a stylized but clinically plausible benefit-cost structure in which true-positive escalation had benefit b_h, false-positive escalation had burden c_h, and false-negative nonaction had harm r_h. Rather than claim universal

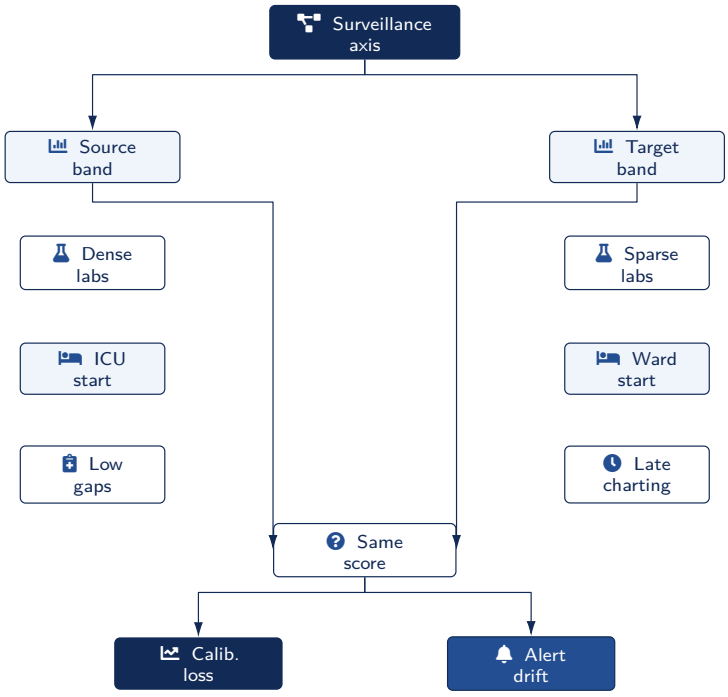

Figure 5. Mechanistic interpretation of transport failure. Middle-risk source score bands are partly defined by local surveillance signatures, such as dense laboratory acquisition, early high-acuity placement, and low documentation gaps. After transport, those same numeric score bands can be populated by cases with different observation patterns. The score therefore preserves a number while losing part of its neighborhood identity, which manifests as calibration loss and unstable action rates.

values for these parameters, the primary analysis reported threshold regret relative to a site-specific oracle threshold:

$$\mathcal{G}_h\tau = U_h\tau_h^* - U_h\tau, \tag{3.4}$$

where $U_h\tau$ was expected utility under environment h and τ_h^* was the best threshold for that environment estimated retrospectively from the validation data. Regret was reported in utility units and as missed net benefit per 100 patients.

To assess how much the models used surveillance structure rather than physiologic content, several ablation analyses were prespecified. One model family used only preoperative and intraoperative variables, thereby excluding early postoperative physiologic and process signals. A second used all variables except measurement-intensity features, unit-location transitions, and missingness indicators. A third used only observation-process features such as test counts, interval since last vital sign, unit location, note density proxies, and missingness patterns. Comparing these models allowed estimation of how much internally predictive power could be achieved from process alone and how well that process signal transported.

Support-aware analysis was included because some target cases could fall in regions of feature space weakly represented during development. A hospital-pooled Mahalanobis support score in a learned representation space was used to rank target cases by source familiarity. For a case representation z, support was

$$sz = \sum_{h=1}^{H} \omega_h \exp\Big(-\frac{1}{2}z - \mu_h{}^{\top}\Sigma_h^{-1}z - \mu_h\Big), \tag{3.5}$$

where μ_h, Σ_h were hospital-specific mean and covariance estimates in representation space. Cases in the lowest support decile were used in a selective-prediction analysis to determine whether abstention could improve external reliability.

All confidence intervals were obtained through cluster bootstrap resampling at the patient level with 2,000 repetitions. Discrimination was summarized by AUROC and area under the precision-recall curve (10). Because the event was not rare enough for AUROC alone to be decisive, AUPRC and positive predictive value at operational thresholds were emphasized. Temporal and external results were reported separately rather than pooled first, since the distinction between same-system drift and new-site transport was of substantive interest.

The silent live-operation phase used the final locked model selected after retrospective evaluation. During this phase, the model was executed in the production analytics pipeline at the 6-hour postoperative decision time but the output was hidden from clinicians. This allowed estimation of real-time feature availability, real-time score distributions, and shadow performance under the live data feed without altering care. A latency audit recorded whether each feature was truly present by the decision time in production, as opposed to appearing available retrospectively only because of backfill or delayed warehouse updates.

The statistical question of interest was not whether one algorithm strictly dominated another in source performance. The central empirical question was whether external transport altered ranking, calibration, and threshold behavior enough that one would regard generalizability as the primary translational

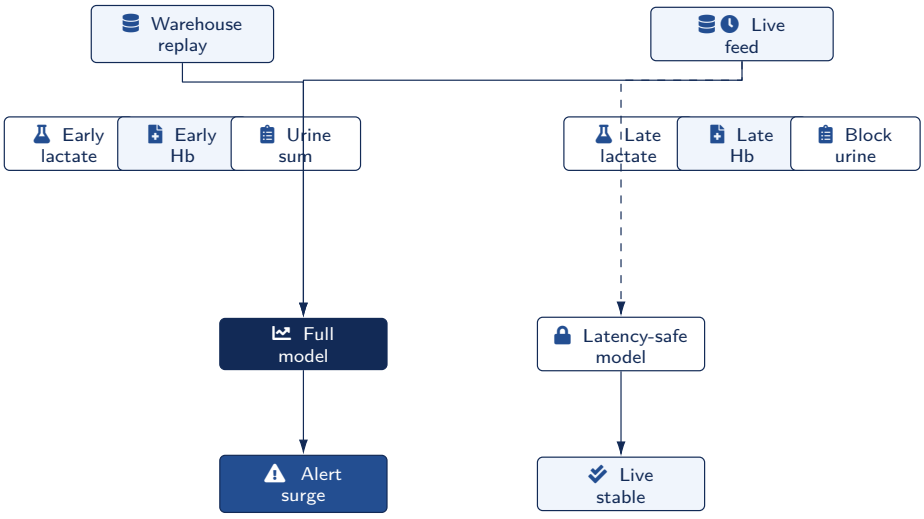

Figure 6. Silent prospective evaluation. Retrospective replay makes several postoperative variables appear timely, but live production exposes the real arrival pattern of laboratory and charted data. The full model therefore enters deployment with a feature universe that is subtly different from the one it saw in historical extracts, producing an unanticipated rise in alert burden. A latency-safe restricted model gives up some replay sharpness yet preserves more stable live semantics.

Table 1. Cohort Distribution Across Hospitals

Hospital Type	Sites (n)	Encounters	Event Rate (%)
Tertiary Centers	2	31,842	10.6
Urban Hospitals	3	39,109	8.3
Community Hospitals	2	22,975	7.4
Total	7	93,926	8.9

bottleneck. The modeling strategy was therefore intentionally aligned to that question. Models were compared not only by local sharpness but by how much of their meaning survived movement across hospitals and then across the additional barrier imposed by live operational data.

4. Results

The retrospective cohort of 93,926 operations was clinically heterogeneous but not incoherent. Median patient age was 61 years, 48.6% were female, and 29.4% of procedures were urgent or emergent. Colorectal operations comprised 38.7% of the cohort, gastric or proximal foregut procedures 11.2%, small bowel resections 18.4%, pancreatobiliary resections 7.8%, urgent laparotomies 13.9%, and other eligible abdominal procedures the remainder. Overall major postoperative decompensation occurred in 8.9% of cases, but the event rate varied from 7.1% to 11.7% across hospitals. Baseline case-mix differences explained part, but not all, of this spread. After adjustment for age, emergency status, ASA class, procedure family, operative duration, transfusion, and preoperative albumin, residual hospital-level variance remained substantial. The implication was immediate: external transport would not be a mere formality because sites differed both in whom they treated and in how risk manifested in records and outcomes.

Observation-process heterogeneity was striking. Median time to first postoperative complete blood count was 64 minutes in Hospital 2 and 171 minutes in Hospital 7 (11). The proportion of patients with lactate measured before the 6-hour decision time ranged from 31% to 79%. Immediate postoperative admission to ICU or step-down varied from 12% to 34%. Missingness indicators for lactate, early hemoglobin, and urine-output density alone recovered hospital identity with macro-F1 of 0.71 in a held-out classifier, whereas demographic variables alone recovered it poorly. Principal component analysis of measurement-intensity variables showed that the first component explained 27% of their variance and tracked site-level surveillance intensity almost monotonically. This mattered because the eventual models used many of the same variables. Their predictive performance would therefore reflect, at least in part, how stable these process signals were across environments.

The source-domain discrimination of the three primary models followed expected patterns. In patient-level nested internal validation, logistic regression achieved an AUROC of 0.793 with 95% confidence interval 0.786 to 0.801 and an AUPRC of 0.276. The source-optimized gradient-boosting model reached AUROC 0.847 with 95% confidence interval 0.841 to 0.853 and AUPRC 0.366. The domain-regularized boosting model reached AUROC 0.833 with 95% confidence interval 0.826 to 0.839 and AUPRC 0.344. Thus the source-optimized boosting model was the clear winner if the question were only which model fit the development hospitals most sharply. Calibration within the development hospitals was

Table 2. Model Performance (Internal Validation)

Model	AUROC	AUPRC	ECE
Logistic Regression	0.793	0.276	0.028
Boosting (Source-Optimized)	0.847	0.366	0.021
Boosting (Domain-Regularized)	0.833	0.344	0.024

Table 3. Grouped Internal Validation (Hospital Holdout)

Model	AUROC	Drop from Internal	Stability Rank
Logistic Regression	0.774	-0.019	Medium
Boosting (Source)	0.801	-0.046	Low
Boosting (Domain-Reg)	0.817	-0.016	High

acceptable for both boosting models after Platt scaling on the development folds, with expected calibration error 0.021 for the source-optimized model and 0.024 for the domain-regularized model. Logistic regression had somewhat flatter calibration with expected calibration error 0.028 but lower discrimination.

The grouped internal estimate told a different story. When one development hospital was held out at a time, logistic regression declined modestly to AUROC 0.774, the source-optimized boosting model dropped more substantially to 0.801, and the domain-regularized boosting model declined only slightly to 0.817. The ranking of models reversed under grouped validation compared with random patient-level folds. This was the first empirical sign that some of the internal superiority of the source-optimized model depended on hospital-specific structure already inside the development network. The domain-regularized model had sacrificed some source sharpness but retained more cross-hospital consistency.

Temporal holdout performance, using later encounters from the same four hospitals, was intermediate between patient-level and grouped internal results. Logistic regression reached AUROC 0.768 and AUPRC 0.246 (12). The source-optimized boosting model achieved AUROC 0.819 and AUPRC 0.310. The domain-regularized boosting model achieved AUROC 0.814 and AUPRC 0.304. All three models showed some temporal decay, but calibration told the more important story. The source-optimized model's temporal calibration slope was 0.84, indicating overconfident prediction. The domain-regularized model's slope was 0.95, much closer to ideal. Intercept drift was modest for both. Thus temporal drift existed, but it was not catastrophic. The sharpest evidence of transport difficulty emerged only when the models left the original hospital system configuration.

Geographic external validation produced the most consequential findings. In the three held-out hospitals, logistic regression achieved pooled AUROC 0.741 with hospital-specific values 0.726, 0.749, and 0.746. The source-optimized boosting model achieved pooled AUROC 0.719 with hospital-specific values 0.736, 0.698, and 0.723. The domain-regularized boosting model achieved pooled AUROC 0.769 with hospital-specific values 0.778, 0.761, and 0.754. The pattern was stable across bootstrap replicates: the model with the best internal discrimination was not the best external model. Its source advantage did not survive contact with new hospitals. By contrast, the model explicitly regularized against cross-hospital instability lost only 0.064 AUROC points from patient-level internal validation to pooled external validation, whereas the source-optimized model lost 0.128.

Precision-recall behavior amplified this difference. External AUPRC for logistic regression was 0.208. The source-optimized boosting model fell to 0.186, and the domain-regularized model achieved 0.237. Because the event rate in the external hospitals remained under 12%, this improvement was operationally meaningful (13). At the development-chosen threshold corresponding to approximately 80% specificity, positive predictive value was 0.227 for logistic regression, 0.213 for source-optimized boosting, and 0.268 for domain-regularized boosting. Negative predictive value remained high in all models, but this was less informative given the moderate prevalence and the triage-oriented deployment question.

Calibration divergence was even more striking than discrimination loss. The source-optimized boosting model had external calibration intercepts of 0.31, -0.12, and 0.47 across the three hospitals, with slopes 0.58, 0.64, and 0.71. Its smooth calibration curves showed clear overprediction in the upper-middle score range at two hospitals and underprediction in the lower-middle range at the third. The domain-regularized model had intercepts of 0.09, -0.04, and 0.15, with slopes 0.87, 0.92, and 0.89. Logistic regression showed slopes between 0.79 and 0.93 but less discrimination. In raw transport, therefore, the domain-regularized model was the only one to remain both reasonably discriminative and close to linearly calibratable across all three external hospitals. Intercept-only recalibration corrected part of the source-optimized model's miscalibration but left major slope distortion, especially in the hospital with the sparsest laboratory surveillance. Intercept-plus-slope recalibration improved expected calibration error but did not repair threshold instability, which remained large because score distribution and local density around the operating point had shifted.

Threshold behavior revealed the clearest translational barrier. The threshold chosen in the development hospitals to deliver approximately 80% specificity produced an alert fraction of 15.1% in the source-optimized boosting model during development. In the three external hospitals, the same threshold generated

Table 4. External Validation Performance

Model	AUROC	AUPRC	PPV
Logistic Regression	0.741	0.208	0.227
Boosting (Source)	0.719	0.186	0.213
Boosting (Domain-Reg)	0.769	0.237	0.268

Table 5. Calibration Slopes Across External Hospitals

Model	Hospital A	Hospital B	Hospital C
Logistic Regression	0.79	0.93	0.88
Boosting (Source)	0.58	0.64	0.71
Boosting (Domain-Reg)	0.87	0.92	0.89

alert fractions of 8.9%, 22.4%, and 27.1%. For the domain-regularized model, the source alert fraction was 15.8% and the external fractions were 14.2%, 16.0%, and 17.5%. In other words, the domain-regularized model preserved something close to threshold meaning across sites, while the source-optimized model did not. This was not a minor calibration nuisance. It meant that a hospital adopting the source-optimized model at the source threshold would face either substantial underalerting or substantial overload depending on the site.

Threshold regret made the same point in utility terms. Relative to a site-specific oracle threshold estimated retrospectively, the source-optimized model incurred net utility regret of 1.7, 4.8, and 5.9 missed utility units per 100 patients across the three external hospitals under the primary benefit-cost setting. The domain-regularized model incurred 1.2, 1.9, and 2.3 units (14). Logistic regression incurred 1.5, 2.6, and 2.8 units. The regret differences were driven less by catastrophic false-negative rates than by unstable operating regions and shifting action burdens. At the busiest external site, the source-optimized threshold would have more than doubled the number of patients flagged for intensified review compared with the source site while improving positive predictive value only marginally. This is the kind of distortion that erodes clinician trust quickly.

The ablation experiments clarified why this was happening. A model trained only on observation-process features, including counts of recorded labs, timing since last vitals, early unit location, and missingness patterns, reached AUROC 0.732 inside the development hospitals. This was a notable result because it meant that process alone carried substantial predictive signal locally. Yet the same process-only model dropped to pooled external AUROC 0.603 and calibration slopes below 0.55 in two of the three external hospitals. Its thresholds were essentially unusable after transport, with action rates varying more than threefold across sites. In contrast, a state-constrained model that excluded observation-process features but retained physiologic values and operative descriptors had lower internal AUROC than the source-optimized full-feature model, 0.821 versus 0.847, yet higher pooled external AUROC, 0.758 versus 0.719, and much tighter calibration. This result strongly suggested that a meaningful share of the source-optimized model's local advantage came from learning process patterns that did not travel.

A complementary ablation using only preoperative and intraoperative variables produced internal AUROC 0.781 and pooled external AUROC 0.744. The drop from internal to external performance was small, but absolute performance was lower. This model was more portable and less clinically useful than the best domain-regularized model. The contrast is informative. Transportability cannot be maximized simply by eliminating postoperative information. Doing so leaves predictive value on the table. The more interesting result is that one can retain early postoperative physiology and still improve transport, provided the model is discouraged from relying too heavily on unstable process features.

Feature-importance and permutation analyses aligned with these observations (15). In the source-optimized boosting model, the top contributors by gain were early lactate, count of postoperative lab draws, ICU or step-down location by 6 hours, operative duration, postoperative tachycardia burden, and white blood cell count. In the domain-regularized model, early lactate, operative duration, transfusion, postoperative tachycardia burden, vasopressor exposure, preoperative albumin, and emergency status moved upward, while lab-draw count and unit-location transitions fell. This shift is interpretable. The domain-regularized objective did not remove process information completely, but it reduced its dominance and forced more weight onto variables whose clinical content was less site-bound.

Subgroup analyses showed that external instability was not evenly distributed. In the source-optimized model, patients aged 75 years and older had a pooled external calibration slope of 0.49, compared with 0.72 in younger adults. Patients initially admitted to an ICU or step-down unit had better external discrimination than ward-start patients, but their calibration curves bent sharply in the upper deciles, suggesting that local immediate-postoperative placement rules influenced score semantics. The domain-regularized model moderated both effects. No major demographic subgroup fell below external AUROC 0.73, and calibration slopes remained between 0.81 and 0.94. The study was not powered to make strong fairness claims, but it was clear that transport failure amplified heterogeneity rather than leaving it uniform.

Selective prediction improved reliability further. When the lowest 8% of target cases by support score were withheld,

Table 6. Threshold-Induced Alert Rates (%)

Model	Source Site	External Min	External Max
Logistic Regression	14.8	12.3	18.6
Boosting (Source)	15.1	8.9	27.1
Boosting (Domain-Reg)	15.8	14.2	17.5

Table 7. Ablation Study Results

Feature Set	Internal AUROC	External AUROC	Transport Gap
Full Model	0.847	0.719	0.128
No Process Features	0.821	0.758	0.063
Process Only	0.732	0.603	0.129
Pre/Intraoperative Only	0.781	0.744	0.037

pooled external AUROC for the domain-regularized model rose from 0.769 to 0.783 and positive predictive value at the development threshold rose from 0.268 to 0.291, while the action fraction dropped modestly from 15.9% to 14.8%. The support-withheld cases were not random. They disproportionately came from urgent laparotomy patients in one external hospital and from a low-surveillance subset of postoperative ward patients in another. In these groups, feature combinations that were common locally had been relatively rare during development. The result suggests that some transport failures were driven not by average population shift alone but by local pockets of support mismatch.

The results up to this point already answered the central empirical question in one respect. The decisive obstacle to translation was not scarcity of signal. The models had ample signal. It was not even absence of sophisticated modeling, since the highest-capacity model was strongest internally (16). The obstacle was that externally meaningful score behavior depended on more than internal discrimination. Models that exploited surveillance and process structure efficiently inside the development hospitals became unstable once those structures changed. The remainder of the analysis therefore turned from the question of whether transport failed to the question of how and where that failure was concentrated.

5. Mechanistic Interpretation of Transport Failure

The pattern of results suggested that external deterioration was not random. It was concentrated in ways that corresponded to site-level differences in observation process and early postoperative workflow. To examine this more carefully, the analysis compared source and target score neighborhoods. For each model, the external cases were divided into deciles of predicted risk, and within each decile we compared the distributions of clinically relevant descriptors that were not fully absorbed by the scalar score. These included operative urgency, early lactate measurement frequency, immediate postoperative location, number of nursing flow entries, and time to first postoperative complete blood count. The source-optimized boosting model showed large source-target neighborhood divergence in the middle-high score region, especially in deciles six through eight. Those score regions in the source data were enriched for patients with dense early surveillance and step-down or ICU placement. In two of the external hospitals, the same score regions included many ward patients with fewer early laboratory assessments. In other words, the same numerical risk interval corresponded to different mixtures of care trajectories.

This neighborhood mismatch was not merely descriptive. It aligned with the observed calibration distortion. In Hospital 6, where lactate was measured far less frequently within 6 hours than in the development centers, the source-optimized model overpredicted most strongly in score bands where lactate had been an important internal discriminator. The model had evidently learned a relationship in which the absence, presence, and timing of lactate implicitly carried clinician-concern information. Once the concern structure changed, the score bands no longer denoted the same case sets. The domain-regularized model was less affected because it had reduced, though not eliminated, dependence on these process-mediated cues.

A low-rank analysis of the observation process gave the same answer in another form (17). When the matrix of missingness indicators, time-since-last-observation variables, and lab-count features was decomposed, the first singular vector corresponded almost perfectly to an axis of surveillance intensity. Projection onto this axis separated tertiary and community hospitals with little overlap. The source-optimized model's score had a correlation of 0.42 with that axis inside the development set; the domain-regularized model's score correlated at 0.19. Among external cases, calibration error increased monotonically with absolute projection onto the surveillance axis for the source-optimized model. For the domain-regularized model this relation was much flatter. This provides a more mechanistic interpretation of transport failure. The source-optimized score had embedded a hospital-surveillance coordinate into the risk estimate. That coordinate was stable in development, useful internally, and nonportable.

Label semantics contributed as well. The composite outcome was intentionally chosen to reduce dependence on any one documentation convention, yet not all components transported equally. Unplanned ICU transfer was the most operationally consistent across hospitals. Septic shock was more sensitive

Table 8. Silent Live Operation Performance

Model	AUROC	Calibration Slope	Alert Rate (%)
Domain-Reg (Full)	0.744	1.18	24.6
Latency-Aware Model	0.741	0.98	16.8
Retrospective Estimate	0.768	0.90	18.9

Table 9. Feature Availability in Live Setting

Feature	Retrospective (%)	Live (%)	Difference
Lactate	81.1	63.4	-17.7
Hemoglobin	86.9	71.5	-15.4
Urine Output	92.3	77.2	-15.1

to local coding and to vasopressor documentation pathways. When the outcome was restricted to urgent reoperation, source versus external discrimination narrowed less for all models, but event frequency fell sharply and calibration remained unstable. When the outcome was restricted to the broader composite, discrimination widened but calibration and threshold effects dominated. This suggests that external failure came from both sides of the modeling problem. Observation-process features transported poorly, and the target label retained some environment dependence despite careful definition.

The contrast between the process-only and state-constrained models was especially revealing. The process-only model's strong source performance showed that local observation and response patterns encode substantial predictive information. The state-constrained model's superior external behavior showed that information generated by those patterns is not the same as transportable risk (18). This distinction is easy to miss in ordinary model-development workflows because both kinds of information improve local fit. The ablation results made the distinction visible. A source-hospital can be highly learnable for the wrong translational reasons.

Model complexity interacted with this problem, but not in a simple way. Logistic regression transported more stably than the source-optimized boosting model but did not outperform the domain-regularized boosting model. This indicates that the main issue was not complexity per se. It was the combination of flexibility and objective. Flexible models can fit both durable clinical nonlinearities and brittle local shortcuts. When optimized purely for source loss, they are likely to exploit both. When regularized against cross-hospital instability, they can retain some of the clinically useful nonlinear structure without binding themselves as tightly to local process geometry. The empirical lesson is that a lower-capacity model is not always the best answer to generalizability. A transport-aware objective can matter as much as, or more than, raw simplicity.

The threshold results help identify why clinicians often experience externally deployed models as unreliable even when AUROC remains passable. The source-optimized model still distinguished higher-risk from lower-risk cases better than chance in all external hospitals. If one judged it only by pooled AUROC, it might appear imperfect but plausible. Yet its risk semantics near the operating region had shifted enough that the same threshold produced either too few or too many alerts depending on site. From a clinical perspective, that instability matters more than a modest drop in AUROC. It changes whether the model can be inserted into workflow without endless manual threshold retuning. The domain-regularized model won not because it had the best external ranking at every single site, but because it preserved threshold meaning well enough to remain governable.

Another mechanistic observation came from cases at the bottom decile of support score (19). These cases were not merely outliers in the usual sense. Many were clinically ordinary in diagnosis but unusual in how their early postoperative data were realized. For example, one external site rarely measured lactate before 6 hours unless there was overt hemodynamic concern, whereas one development site measured it in a much larger fraction of high-risk but not necessarily unstable patients. Cases from the external site with no early lactate but otherwise concerning vitals occupied feature patterns that were underrepresented in development. The source-optimized model responded to them unreliably, often assigning middling scores despite subsequent deterioration. This is not a generic failure of the algorithm. It is a failure of the training distribution to include enough analogues of a target-specific observation pattern.

The site-specific oracle thresholds provide a final empirical clue. For the domain-regularized model, oracle thresholds varied only modestly across external hospitals, and the development threshold sat near the middle of that range. For the source-optimized model, oracle thresholds varied widely and were far from the development-chosen value in two hospitals. This means that the same scalar scale could not carry stable operational meaning after transport. The model had not merely lost a little calibration. It had lost a coherent decision coordinate. That is a much more serious translational problem than source-domain development metrics can reveal.

Taken together, these analyses indicate that the empirical barrier was fundamentally about score meaning under changed observation and outcome regimes. The models did not fail because the hospitals were radically different in diagnosis or because the event was too noisy to model. They failed to the extent that the source-trained score represented a mixture of physiologic vulnerability and hospital-specific surveillance-

response structure. The evidence for this is direct: process-only features predicted well locally and poorly externally; removing process features improved transport; regularizing against cross-hospital instability preserved much of the nonlinear benefit while stabilizing thresholds; and support-based abstention reduced external error in precisely the regions where source analogues were sparse. These are not ancillary findings (20). They specify what made external validation central. Without moving the models into other hospitals, none of these mechanisms would have been empirically legible.

6. Deployment Implications from Silent Prospective Evaluation

The retrospective findings showed that external transport was the main scientific obstacle. The silent live-operation phase then showed that even strong retrospective external performance does not fully settle the translational question. During the 12-week shadow run in Hospitals 5 and 6, the final locked domain-regularized model was executed at the 6-hour postoperative decision time without exposing scores to clinicians. A total of 6,102 operations were scored. The outcome rate in this shadow cohort was 8.7%, similar to the corresponding retrospective external period. At first glance, one might have expected performance to match the earlier external replay closely. It did not.

In retrospective warehouse extracts, the locked domain-regularized model had AUROC 0.768 across the same hospitals and a calibration slope of 0.90. Under silent live operation, AUROC fell to 0.744 and the calibration slope shifted to 1.18 before any recalibration. The drop was not due to changed clinician behavior, because the score was hidden. Instead, the degradation came from feature latency and data finalization effects. Three variables that had been particularly informative in retrospective replay, early lactate, postoperative hemoglobin, and cumulative urine output by 6 hours, were not actually available at the nominal decision time in a nontrivial fraction of live cases. In the warehouse they appeared timely because of later chart completion and data integration. In production, the model saw their true real-time availability.

The latency audit quantified this clearly. Early lactate was present by decision time in only 63.4% of shadow cases, compared with 81.1% in retrospective extracts. Postoperative hemoglobin was present in 71.5% of shadow cases versus 86.9% retrospectively. Urine-output aggregation was fully available in 77.2% of shadow cases because some records were charted in blocks rather than continuously (21). When these variables were absent in production, the model fell back on missingness indicators and correlated features, but the local relationship between missingness and risk had shifted relative to development. The shadow phase thus uncovered an additional form of generalizability failure: transport from retrospective warehouse time to operational time.

A latency-aware restricted model was then evaluated in replay using only features verifiably available by the decision time in production. This model had slightly lower retrospective external AUROC than the full domain-regularized model, 0.752 instead of 0.769, but its shadow AUROC was 0.741 and calibration slope 0.98. In effect, the latency-aware model gave up some retrospective sharpness to maintain operational meaning. This result closely paralleled the earlier transport findings across hospitals. The translationally better model was not the one with the highest retrospective score. It was the one whose score survived a more realistic environment.

Threshold behavior in shadow mode was particularly informative. Using the development-chosen operating threshold from the final domain-regularized model, the projected alert fraction in retrospective replay was 18.9%. In actual shadow operation it was 24.6% for the full model because missing late-arriving laboratory values pushed more cases into intermediate-risk score bands. Positive predictive value at this threshold was 0.214, markedly lower than the retrospective estimate of 0.258. The latency-aware restricted model yielded a shadow alert fraction of 16.8% and positive predictive value of 0.247, much closer to retrospective expectations. In a hypothetical deployment this difference would have been critical. The model that looked better on historical replay would have produced substantially more clinician-facing workload than anticipated.

The shadow phase also showed that production support mismatch was not uniform over the day. Cases crossing the 6-hour decision point during overnight hours had lower feature completeness and a slightly different score distribution than daytime cases, despite similar event rates. The full model's shadow calibration was notably poorer overnight, with a slope of 1.29 versus 1.07 during daytime. This pattern reflected charting and laboratory turnaround behavior rather than patient biology (22). It is exactly the kind of effect that retrospective studies often miss because data warehouses eventually contain both overnight and daytime information in apparently complete form. Silent live operation brought the true timing structure back into view.

Sequential monitoring across the 12-week shadow run showed that the live deployment environment was not stationary even within that relatively brief observation window. In principle, a shadow phase is often treated as a controlled period in which the model can be observed under real operating conditions without influencing care. In practice, however, the environment continued to evolve while the evaluation was underway. The clearest example occurred in week seven, when one of the external hospitals revised its postoperative order sets. That workflow change increased the frequency of early lactate acquisition among high-acuity surgical patients, thereby altering the immediate data landscape presented to the model. Following this shift, the calibration error of the full model narrowed modestly, whereas the restricted model remained largely unchanged. Although the magnitude of the effect was small, its interpretive value was substantial. It demonstrated that model transport should not be conceptualized as a one-time technical hurdle cleared by a single external validation exercise. Rather, transport is better understood as an ongoing ecological problem in which the model, the measurement system, and the care environment continually interact.

This episode is especially instructive because the model itself did not change. No coefficients were retrained, no thresh-

olds were adjusted, and no architecture was modified. Yet the practical meaning of the model's output shifted because the workflow around it shifted. Additional early lactate measurements changed the structure and timing of available information, which in turn altered how the model's learned associations manifested in routine care. In that sense, the observed calibration improvement in the full model was not evidence of intrinsic model refinement, but of environmental realignment. The model appeared to perform somewhat differently because the clinical system began to generate inputs in a pattern more compatible with the conditions under which the model's logic was learned or functioned most effectively. The restricted model's relative stability under the same change is also notable, suggesting that its dependence on the newly affected features or timing relationships was weaker. Together, these patterns reinforce the point that deployment performance is partly a property of the surrounding workflow, not merely of the statistical object exported from development.

The silent phase also yielded a practical estimate of the possible value of selective prediction. Rather than forcing the model to issue an equally authoritative ranking for every case, the shadow analysis considered whether some cases should instead be identified as poor candidates for routine automation. Using the support score as a marker of experiential familiarity, the lowest 10% of shadow cases were flagged for clinician review rather than automatic ranking. This modest abstention strategy produced a meaningful improvement in operational precision. For the full model, positive predictive value at the prespecified threshold increased from 0.214 to 0.255. For the restricted model, it increased from 0.247 to 0.278. These gains suggest that some of the model's apparent weakness arose not from uniform underperformance across all contexts, but from a concentrated subset of cases in which the model was being asked to extrapolate beyond regions of strong empirical support.

The composition of the abstained cases helps clarify why this mattered. These cases were enriched for overnight urgent laparotomies and for patients initially placed in short-stay postoperative units where early charting was sparse. Both scenarios represent operational contexts in which data accumulation is less orderly, clinical trajectories may be more abrupt, and the timing of documentation differs from the settings that dominate many development cohorts. Such cases are not random outliers; they are structurally distinct encounters in which the model may have limited experiential grounding. By identifying them through a support-based criterion, the system was able to reserve them for human oversight rather than present an overly confident score that might invite misplaced trust.

These findings point toward a more realistic conception of safe deployment. The central challenge is not simply to produce a model with a good average validation metric and then assume that performance will generalize unchanged. Instead, safe deployment may require explicit acknowledgment of where the model has weak experiential support, where the local workflow has shifted, and where prediction should be softened by deferral rather than forced into a deterministic output. In this view, abstention is not a failure of the model but a principled recognition of its operational limits. The shadow run therefore offered two linked lessons: first, that model meaning can drift when workflow changes even if the fitted model is fixed; and second, that deployment safety may depend as much on knowing when not to automate as on optimizing discrimination or calibration in the aggregate.

A practical deployment implication emerged from all of this. The translational path for a multicenter clinical prediction model cannot end at retrospective external validation, even when that validation is done carefully. A model can appear externally credible in warehoused data and still misbehave in live operation because production-time feature availability, charting delay, and workflow rhythm create a different observational environment. The silent phase therefore acted as a second external validation, one that changed the temporal semantics of the inputs without yet changing clinician behavior. It narrowed the realistic deployment set further. Only the latency-aware restricted model retained sufficiently stable alert burden and calibration to justify serious consideration for clinician-facing use.

This result also changed the interpretation of model updating. Had the full domain-regularized model been activated immediately, the natural institutional response would probably have been threshold adjustment to control alert volume. That would have addressed one symptom of live transport failure while obscuring its cause (23). The shadow phase instead showed that the more faithful repair was upstream: remove or replace variables whose apparent availability in retrospective data had depended on backfill, then recalibrate the reduced model under the live feed. In translational terms, this is important. Not all deployment problems should be solved by tuning thresholds. Some reflect a deeper mismatch between the retrospective feature universe and the operational one.

The empirical conclusion from the shadow phase is therefore consistent with the retrospective analysis and sharper in one respect. Generalizability was not merely the barrier between one hospital and another. It also separated retrospective data reality from production data reality inside the same target hospitals. A model had to survive both changes to become a plausible clinical tool. Source-domain optimization was silent on both. External validation across sites answered the first. Silent live operation answered the second. Only after both steps did the score begin to look like a stable instrument rather than a promising retrospective artifact.

7. Conclusion

This empirical study set out to determine whether the main obstacle to translational clinical prediction was lack of signal or lack of transport. The evidence favored the second explanation. In a large multi-hospital surgical cohort, the strongest source-domain model was not the strongest external model. Models that made efficient use of local surveillance and process signals achieved excellent internal discrimination, but they lost calibration, threshold stability, and operational coherence after transport. A domain-regularized model with slightly lower internal sharpness preserved substantially more external meaning. Additional ablation analyses showed why: observation-process

features carried considerable local predictive value yet transported poorly, while models constrained toward more durable clinical structure retained better external behavior. Silent live operation then added a further layer of evidence by showing that production-time feature latency and backfill could create meaningful degradation even after careful retrospective external validation (24).

These findings support a specific interpretation of generalizability as the central translational barrier. The problem was not that the model could not learn. It learned very well. The problem was that what it learned inside one observational and workflow regime did not automatically retain the same semantics outside that regime. External validation mattered because it exposed this distinction directly. Without hospital-level transport tests, the superiority of the source-optimized model would have appeared decisive. Without shadow live evaluation, the remaining operational fragility of even the best external model would have remained hidden. In both cases, the missing information was not more data from the same environment. It was evidence from changed environments.

For development practice, the implications are concrete. Multicenter model comparison should report grouped internal estimates, raw external calibration, threshold-local workload distortion, and support-aware behavior, not only internal AUROC. Observation-process features should be examined empirically for their transport value rather than accepted automatically because they improve local performance. Model selection should consider cross-environment stability as an objective in its own right. For deployment practice, silent live operation should precede action-linked use whenever feasible, because retrospective warehouse availability is not the same thing as real-time availability. Recalibration should be interpreted in light of operational semantics rather than treated as a purely mechanical fix.

A locally successful clinical score may still be useful locally. This study does not argue otherwise. The narrower conclusion is that nonlocal use requires nonlocal evidence, and that evidence must address both hospital-to-hospital transport and retrospective-to-live transport. In this dataset, external validation and silent operational evaluation together did far more to determine translational readiness than any additional gain in source-domain discrimination. That is why generalizability, rather than model construction alone, remains the central barrier to reliable clinical deployment (25).

References

[1] N. El-Sourani, C. Kechagia, F. Alfarawan, A. Troja, and M. Bockhorn, "Classification and evaluation of anastomotic leaks after esophageal surgery—a tertiary university experience," *European Surgery*, vol. 54, pp. 80–85, 4 2021.

[2] D. A. Clark, E. Yeoh, A. Edmundson, C. A. Harris, A. R. L. Stevenson, D. Steffens, and M. J. Solomon, "A development study of drain fluid gastrografin as a biomarker of anastomotic leak.," *Annals of coloproctology*, vol. 38, pp. 124–132, 1 2021.

[3] K. Ashrafzadeh, M. Shafiekhani, N. Azadeh, M. Esmaeili, and H. Nikoupour, "Lessons learned from successful autologous gastrointestinal reconstruction in patients with intestinal failure: a case series.," *BMC surgery*, vol. 21, pp. 1–8, 2 2021.

[4] C. H. Richards, J. Hayes, M. Thomson-Fawcett, and T. Elliot, "Cr03 smoking is a major risk factor for anastomotic leak after low anterior resection," *ANZ Journal of Surgery*, vol. 77, 4 2007.

[5] J. Pinson, J.-J. Tuech, M. Ouaissi, M. Mathonnet, F. Mauvais, E. Houivet, E. Lacroix, J. Rondeaux, C. Sabbagh, and V. Bridoux, "Role of protective stoma after primary anastomosis for generalized peritonitis due to perforated diverticulitis-diverti 2 (a prospective multicenter randomized trial): rationale and design (nct04604730).," *BMC surgery*, vol. 22, pp. 191–191, 5 2022.

[6] M. Worbes-Cerezo, B. Nafees, A. R. Lloyd, K. Gallop, I. Ladha, and C. Kerr, "Disutility study for adult patients with moderate to severe crohn's disease.," *Journal of health economics and outcomes research*, vol. 6, pp. 47–60, 3 2019.

[7] B. Sadanandan, "Data-driven risk stratification for anastomotic leak: A proof-of-concept study using electronic health records," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 6, pp. 1–27, 2024.

[8] M. Bhasin and P. K. Sachan, "Role of probiotics on surgical site infections in colorectal surgery," *International Surgery Journal*, vol. 7, pp. 1867–, 5 2020.

[9] P. Piso, S. D. Nedelcut, B. Rau, A. Königsrainer, G. Glockzin, M. A. Ströhlein, R. Hörbelt, and J. Pelz, "Morbidity and mortality following cytoreductive surgery and hyperthermic intraperitoneal chemotherapy: Data from the dgav studoq registry with 2149 consecutive patients," *Annals of surgical oncology*, vol. 26, pp. 148–154, 11 2018.

[10] M. Tachezy, S.-H. Chon, I. Rieck, M. Kantowski, H. Christ, K. Karstens, F. Gebauer, T. Goeser, T. Rösch, J. R. Izbicki, and C. Bruns, "Endoscopic vacuum therapy versus stent treatment of esophageal anastomotic leaks (esoleak): study protocol for a prospective randomized phase 2 trial.," *Trials*, vol. 22, pp. 377–377, 6 2021.

[11] T. W. C. Mak, J. F. Y. Lee, K. Futaba, S. S. F. Hon, D. Ngo, and S. S. Ng, "Robotic surgery for rectal cancer: A systematic review of current practice.," *World journal of gastrointestinal oncology*, vol. 6, pp. 184–193, 6 2014.

[12] K. Abebe, B. Geremew, B. Lemmu, and E. Abebe, "Indications and outcome of patients who had re-laparotomy: Two years' experience from a teaching hospital in a developing nation.," *Ethiopian journal of health sciences*, vol. 30, pp. 739–744, 9 2020.

[13] M. I. R. Bernadó, O. C. Maña, M. Martín-Baranera, and P. B. Sánchez, "Morbimortality after 1321 consecutive crs+hipec procedures: seeking excellence in surgery for peritoneal surface malignancy.," *Clinical & translational oncology : official publication of the Federation of Spanish Oncology Societies and of the National Cancer Institute of Mexico*, vol. 25, pp. 2911–2921, 4 2023.

[14] A. Legault-Dupuis, P. Bouchard, F. Nicodème, J.-P. Gagné, S. Simard, Y. Lacasse, P.-L. Hache, C. Couture, and A.-S. Laliberte, "855 optimization of surgical delay after induction chemoradiation for esophageal cancer: 10 years retrospective multi-center study," *Diseases of the Esophagus*, vol. 34, 9 2021.

[15] M. Okumuş, D. Devecioğlu, M. Çevik, and B. Tander, "Anastomotic leaks and the relationship with anastomotic strictures after esophageal atresia surgery; effects of patient characteristics.," *Acta chirurgica Belgica*, vol. 124, pp. 114–120, 6 2023.

[16] S. K. Kamarajah, A. Madhavan, J. Chmelo, M. Navidi, S. Wahed, A. Immanuel, N. Hayes, S. M. Griffin, and A. W. Phillips, "Impact of smoking status on perioperative morbidity, mortality, and long-term survival following transthoracic esophagectomy for esophageal cancer," *Annals of surgical oncology*, vol. 28, pp. 4905–4915, 3 2021.

[17] N. El-Sourani, H. Bruns, A. Troja, H.-R. Raab, and D. Antolovic, "Routine use of contrast swallow after total gastrectomy and esophagectomy: Is it justified?," *Polish journal of radiology*, vol. 82, pp. 170–173, 3 2017.

[18] A. Yalikun, B. Zuo, W. Kwan, K. Dai, H. Hong, S. Li, J. Ma, P. Xue, and L. Zang, "Modified overlap method for esophagojejunostomy in totally laparoscopic total gastrectomy: A retrospective study of a single center," 7 2023.

[19] Q. M. Leong, D. N. Son, J. S. Cho, J. Baek, J. M. Kwak, A. H. Y. Amar, and S. H. Kim, "Robot-assisted intersphincteric resection for low rectal cancer: technique and short-term outcome for 29 consecutive patients," *Surgical endoscopy*, vol. 25, pp. 2987–2992, 4 2011.

[20] W. Yibulayin, S. Abulizi, H. Lv, and W. Sun, "Minimally invasive oesophagectomy versus open esophagectomy for resectable esophageal cancer: a meta-analysis," *World journal of surgical oncology*, vol. 14, pp. 304–304, 12 2016.

[21] L. K. Vallamsetty, R. Porla, and Rao, "Gastric transposition as an esophageal replacement in children," *Indian journal of applied research*, vol. 6, 4 2016.

[22] R. Zulfikar, V. S. Budipramana, and H. Kahar, "Comparison of colonic anastomosis using dry amnion membrane and fibrin glue in intraperitoneal infection condition assessed from tissue hydroxyproline level measurement (study on wistar rat)," *Indian Journal of Public Health Research & Development*, vol. 12, pp. 379–384, 1 2021.

[23] A. Karachun, L. Panaiotti, I. Chernikovskiy, S. Achkasov, Y. Gevorkyan, N. Savanovich, G. Sharygin, L. Markushin, O. Sushkov, D. Aleshin, D. Shakhmatov, I. Nazarov, I. Muratov, O. Maynovskaya, A. Olkina, T. Lankov, T. Ovchinnikova, D. Kharagezov, D. Kaymakchi, A. Milakin, and A. Petrov, "Short-term outcomes of a multicentre randomized clinical trial comparing d2 versus d3 lymph node dissection for colonic cancer (cold trial)," *The British journal of surgery*, vol. 107, pp. 499–508, 3 2020.

[24] R. Tabola, R. Cirocchi, A. Fingerhut, A. Arezzo, J. J. Randolph, V. Grassi, G. A. Binda, V. D'Andrea, I. Abraha, G. Popivanov, S. D. Saverio, and A. P. Zbar, "A systematic analysis of controlled clinical trials using the niti car™ compression ring in colorectal anastomoses.," *Techniques in coloproctology*, vol. 21, pp. 177–184, 1 2017.

[25] M. Elkerkary, M. Elnagar, M. A. Ali, and H. Shaban, "Evaluation of the predictive value of serum c-reactive protein and procalcitonin levels in early detection of anastomotic leakage after gastrointestinal surgery," *Suez Canal University Medical Journal*, vol. 23, pp. 30–40, 3 2020.