Center for Sustainable Science and Engineering Innovation

CSSEI

ARTICLE

Domain Shift and Cross-Setting Transportability of Anastomotic Leak Prediction Across Gastrointestinal Procedure Families and Hospital Environments

Mahmudul Karim,[1] Fahim Uddin,[2] and Sajid Hossain[3]

[1]Department of Computer Science and Engineering, Pundra University of Science and Technology, Rangpur Road, Shakpala, Bogura 5800, Bangladesh
[2]Department of Business Administration, Britannia University, Paduar Bazar, Bishwa Road, Cumilla 3500, Bangladesh
[3]Department of Information and Communication Technology, Port City International University, South Halishahar, Chattogram 4219, Bangladesh

Abstract

Anastomotic leak is common enough to shape postoperative surveillance and uncommon enough to make model transport difficult. Most prediction studies report a single pooled performance value, although the meaning of that value depends on where the model is deployed. A model trained in an elective colorectal pathway is not necessarily expected to behave similarly in emergency bowel surgery or transfer-heavy upper gastrointestinal care, even when the same structured variables are available. This study examined how much predictive performance, calibration, and decision thresholds change when leak models are moved across surgery-type and care-setting domains. We assembled a retrospective multicenter cohort of adult gastrointestinal anastomoses from seventeen hospitals and defined twenty-five clinical domains by crossing five procedure families with five episode-level care settings. Models were trained using structured variables available by 18 hours after skin closure. The primary endpoint was clinically managed anastomotic leak within 30 days. The final cohort included 54,910 procedures and 3,625 leaks, yielding an overall incidence of 6.6%. Random-split validation overstated performance for all methods. A pooled gradient boosting model achieved an apparent area under the receiver operating characteristic curve of 0.812 but declined to 0.736 under leave-domain-out testing. The proposed domain-mixture model with invariant representation learning and calibration regularization achieved 0.781 under the same transport evaluation and reduced expected calibration error from 0.083 to 0.034. Transport failure was concentrated in low pelvic colorectal procedures performed in emergency overnight settings and in upper gastrointestinal reconstructions managed through transfer-in complex care. A global alert threshold produced wide variation in positive predictive value across domains, whereas domain-cluster calibration narrowed that spread substantially. These findings indicate that heterogeneity of leak risk across surgery types and care settings is also a heterogeneity of model validity. In this problem, transportability is itself a clinical outcome.

1. Introduction

Anastomotic leak prediction is often discussed as though it were a single technical problem with a single answer [1]. A model is developed, its discrimination is reported, a threshold is proposed, and the conversation moves toward potential clinical use. That sequence is tidy, but it leaves one central issue unresolved. A prediction rule that performs adequately when evaluated on a mixed sample of postoperative admissions may not remain adequate when it

is deployed inside a specific clinical domain such as low pelvic colorectal surgery occurring overnight in an emergency pathway, or upper gastrointestinal reconstruction performed after interhospital transfer for complex cancer care. In these domains, case mix, operative physiology, monitoring habits, rescue availability, and data density all differ. A model can appear stable in the aggregate while drifting materially in one domain and remaining acceptable in another. The result is not only statistical inconvenience. It changes who is escalated, who is imaged, and who is falsely reassured.

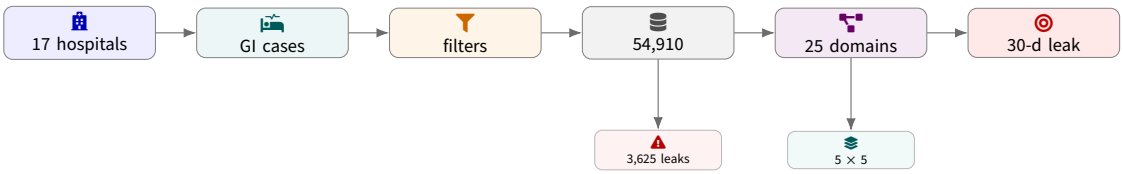

Figure 1. Study schema used in the transport analysis. Structured records from seventeen hospitals were restricted to adult gastrointestinal restorative procedures, filtered to a single index anastomosis per admission, crossed into twenty-five clinically visible domains, and linked to clinically managed anastomotic leak within thirty days.

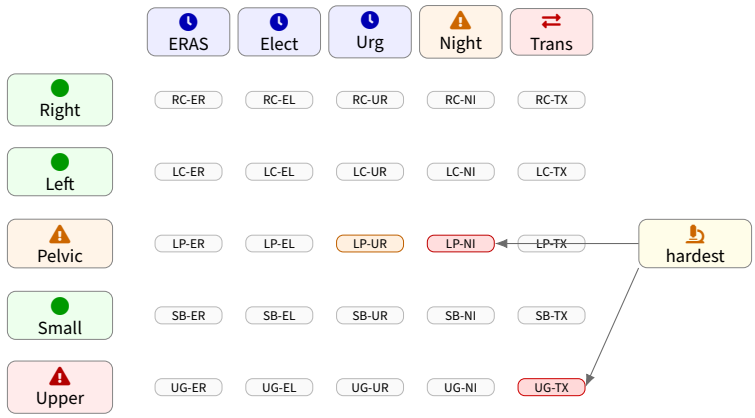

Figure 2. Construction of the twenty-five clinical domains by crossing five procedure families with five episode-level care settings. Warm cells mark the domains where transport failure concentrated most strongly, especially low pelvic colorectal surgery in emergency overnight care and upper gastrointestinal reconstruction in transfer-in complex care.

The present paper examines that problem directly. The substantive question is not whether structured electronic health record data can predict anastomotic leak at all. That has already become plausible. The harder question is whether the learned mapping from data to risk remains valid when the procedural and organizational context changes. Put differently, heterogeneity of leak risk across surgery types and care settings can be studied at least two ways. One can estimate how the event rate itself differs across these strata. One can also estimate how the meaning of the same postoperative variables differs across those strata when used for prediction. These are related but distinct objects. A model may correctly rank patients within one domain yet systematically underpredict risk when transferred into another because the baseline prevalence shifts, because the feature distribution changes, or because the clinical relationship between predictors and leak is itself altered. The third mechanism is especially consequential because it implies not only covariate shift but concept drift [2].

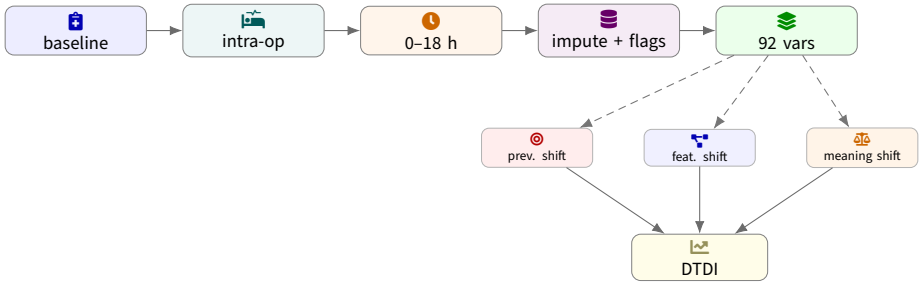

Figure 3. Early-postoperative feature construction and empirical shift quantification. Baseline, operative, and first-eighteen-hour variables were combined with explicit missingness indicators; domain dissimilarity was then summarized by prevalence shift, feature-space shift, and changes in predictor meaning through a composite transfer-difficulty index.

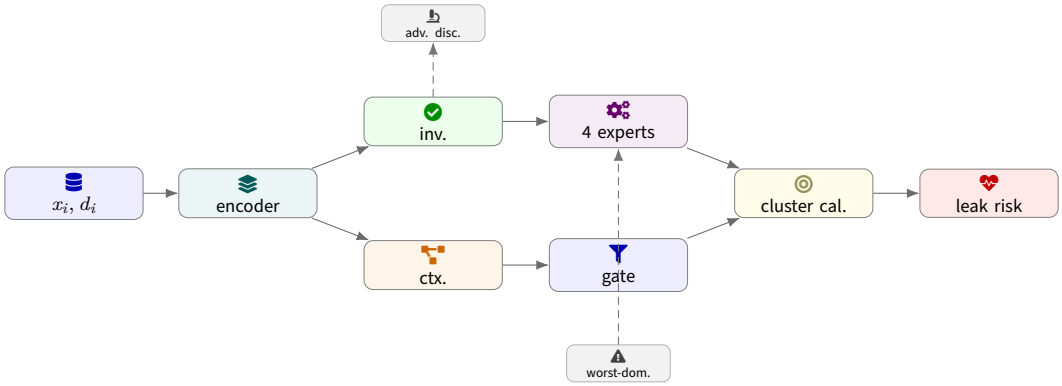

Figure 4. Domain-mixture architecture used for cross-setting transport. A shared encoder maps structured variables into an invariant branch and a contextual branch; experts model reusable physiology, the gate conditions on domain structure, and cluster-level calibration stabilizes the risk scale in sparse or shifted targets while adversarial and worst-domain penalties discourage brittle fits.

This study is organized around the observation that postoperative leak prediction is almost always deployed into a narrower environment than the one in which it is developed. A hospital may want a surveillance tool for its enhanced recovery colorectal pathway, for urgent inpatient bowel surgery, or for patients transferred in for complex upper gastrointestinal reconstruction. The local decision problem is domain-specific even when the derivation cohort is broad. Clinicians interpret the same measured quantity differently depending on context. A white blood cell count of 15 may be unremarkable on the evening of a difficult emergency laparotomy and more concerning after an otherwise routine left-sided colectomy on a standardized elective pathway. A lactate of 2.4 mmol/L may carry different implications after prolonged esophagectomy with vasopressor support than after ileocolic anastomosis in a lower-risk pathway. These contextual differences are not idiosyncratic judgment alone. They reflect underlying heterogeneity in tissue perfusion, contamination burden, technical difficulty, surveillance intensity, and rescue thresholds.

Prediction research has begun to show that early postoperative structured variables contain meaningful signal for leak, but the transition from feasibility to usable deployment depends on what happens when models leave their derivation context. As a proof-of-concept analysis using MIMIC-IV, Sadanandan (2024) showed that early postoperative white blood cell count,

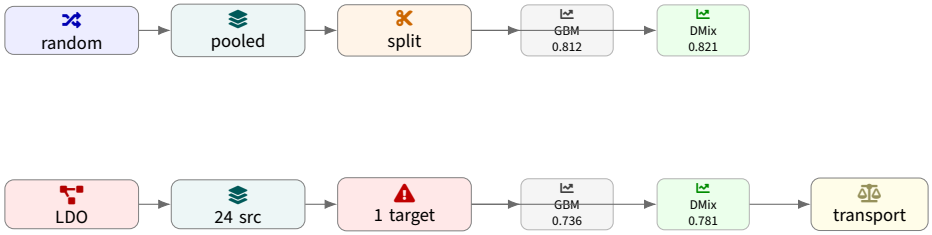

Figure 5. Validation design and the performance inflation created by pooled random splitting. Apparent performance remained high when train and test sets shared the same domain mixture, whereas leave-domain-out testing exposed the transport gap and provided a closer estimate of behavior in a truly shifted deployment target.

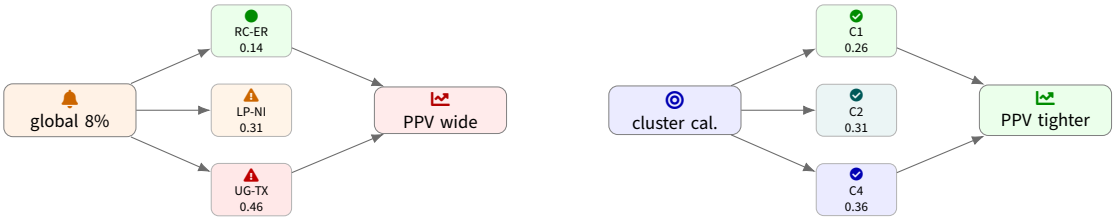

Figure 6. Operational effect of thresholding under domain shift. A single global alert threshold generated wide swings in positive predictive value and sensitivity across surgery-setting targets, whereas cluster-level calibration compressed that spread and made the alert workload more comparable across domains without requiring a fully separate model for every target.

ICU admission, postoperative lactate, surgery type, and preoperative albumin carried substantial predictive information, while also noting that the study was based on a critical care dataset and did not examine subgroup performance by procedure type or care setting. That gap is not a minor reporting omission [3]. It points toward the main unresolved problem in clinical prediction for leak: a pooled estimate of performance may obscure the very domains in which the model would either be used most intensively or fail most visibly.

The transport question is difficult because several types of shift occur at once. The first is label shift, meaning the event rate changes across domains. A model trained in elective enhanced recovery surgery can become poorly calibrated in emergency contaminated surgery even if the ranking of risk is preserved, simply because the baseline event rate differs. The second is covariate shift, meaning the distributions of albumin, vasopressor exposure, operative duration, diversion, lactate, and other variables differ. The third is conditional shift, meaning the clinical interpretation of those variables changes. A prolonged operation can mean technically demanding but controlled resection in one setting and physiologically unstable damage-control progression in another [4]. ICU admission can function as a marker of postoperative instability in one domain and a routine pathway step in another. These shifts are not merely mathematical categories. They correspond to real changes in how postoperative trajectories are generated and recorded.

One response is to build a separate model for every surgery type and care setting. That strategy is superficially attractive because it localizes fit. Yet it quickly encounters the sample size problem. Some domains are clinically high stakes but numerically modest, especially when crossed by hospital and time. A separate model for each domain discards shared signal and yields unstable estimates where calibration matters most. The opposite response is to

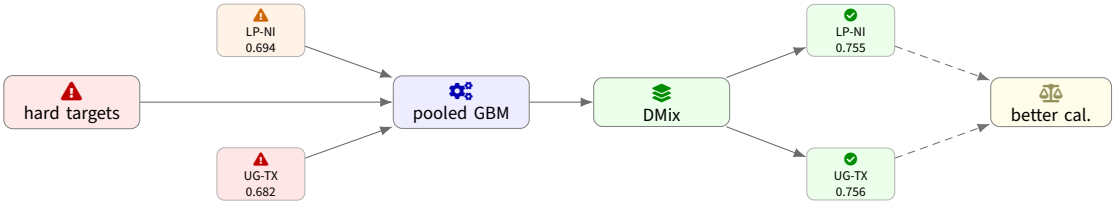

Figure 7. Concentrated transport failure and recovery in the two most difficult domain families. The pooled gradient-boosting model degraded sharply in emergency low pelvic colorectal surgery and transfer-heavy upper gastrointestinal care, while the transport-aware mixture model recovered both rank performance and numerical calibration in these clinically sensitive targets.

pool all data and trust the model to learn what matters. That approach often looks strong under random internal validation because the train and test samples share the same domain mixture. It is much less reassuring when the deployment target is a domain underrepresented or absent in the derivation sample. The present study therefore takes a third route: preserve pooled information while forcing the model to learn what is transferable and what must remain domain-conditioned.

The empirical setting is well suited to that problem. The cohort spans seventeen hospitals and includes a broad range of gastrointestinal anastomoses. Rather than define care setting by hospital label alone, the analysis uses episode-level settings that are visible at the point of intended prediction. These settings distinguish elective enhanced recovery pathways, conventional elective inpatient care, urgent inpatient surgery, emergency overnight operations, and transfer-in complex care. Crossed with five procedure families, they generate twenty-five clinically interpretable domains. Some of these domains are common and relatively homogeneous [5]. Others are rare, physiologically volatile, or both. This structure allows the paper to ask not just which patients are high risk, but which domains are easy or hard to transport into, and why.

The study pursues four aims. The first is descriptive: to quantify how leak incidence and early postoperative feature distributions vary across the surgery-setting domains. The second is diagnostic: to measure how far these domains differ from one another in terms of covariate geometry, event prevalence, and calibration demands. The third is predictive: to compare pooled, domain-specific, and domain-generalized models under leave-domain-out transport evaluation. The fourth is operational: to estimate how much a single global alert threshold distorts positive predictive value and sensitivity across domains, and whether domain-cluster calibration reduces that distortion without requiring a fully separate model for every domain.

Several design choices follow from those aims. Only variables available by 18 hours after skin closure are used. This preserves a realistic postoperative surveillance window while avoiding the later time-dependent complications that arise once formal leak work-up is already underway. The endpoint is clinically managed leak within 30 days, captured through an adjudicated composite that emphasizes treatment-linked recognition rather than coding alone. Validation is performed by withholding entire domains rather than random records, because the core question concerns transport across contexts. Model comparison therefore focuses less on apparent discrimination in pooled samples and more on retained discrimination, calibration error, threshold stability, and decision consistency in held-out targets.

The central expectation is not that one model will eliminate domain heterogeneity. That would be unrealistic. Instead, the expectation is that explicit modeling of domain structure

will reduce the gap between pooled development and target deployment, especially in domains where case mix and data-generating processes differ most sharply. If that expectation is supported, the implication is practical. Clinical adoption should not be framed as the installation of a single leak model across all postoperative gastrointestinal care [6]. It should be framed as the deployment of a transport strategy, one that acknowledges that a model moved across surgery types and care settings becomes a different instrument unless its transfer properties are measured and corrected.

2. Study Design, Clinical Domains, and Endpoint Construction

This investigation was conducted as a retrospective multicenter cohort study using structured electronic health record data from seventeen acute care hospitals belonging to a common data network between January 2016 and December 2025. The network included tertiary referral hospitals, regional teaching centers, and mixed-acuity general hospitals. The shared data model allowed standard extraction of laboratory timestamps, medication administrations, operative records, admission source, ward movements, and thirty-day readmission linkage. The use of a shared data architecture did not eliminate clinical heterogeneity. On the contrary, it made it possible to compare domains whose documentation rules were technically aligned while their patient populations and care pathways remained different. That distinction is important because many apparent transport failures in multicenter modeling arise from incompatible data capture. Here the emphasis was placed on clinical transport rather than purely technical interoperability.

Eligible admissions involved adult patients undergoing a gastrointestinal operation with restoration of continuity by an anastomosis between the esophagus, stomach, small bowel, colon, or rectum. For patients with more than one qualifying operation during the study window, each hospitalization contributed at most one index procedure so that the postoperative feature window could be anchored cleanly to a single anastomosis. Exclusions were selected to preserve comparability of postoperative physiology rather than to maximize sample size. Trauma laparotomy, transplant surgery, planned staged abdominal closure, palliative bypass without primary restorative intent, and procedures lacking reliable skin-closure timestamps were excluded. Admissions transferred into the network after the operation itself were also excluded, because the first postoperative hours could not be reconstructed. After these restrictions, the final analytic sample contained 54,910 qualifying procedures.

Procedure type was organized into five families chosen to preserve clinically meaningful differences while keeping enough support for domain-based transport analysis. The right-sided colonic family included ileocolic and related right colectomy reconstructions. The left-sided colonic family included left colectomy and sigmoid colectomy with colocolic or upper colorectal continuity. The low pelvic colorectal family included low anterior resection, ultralow colorectal reconstruction, and coloanal anastomosis, with or without proximal diversion [7]. The small-bowel family included jejunal, ileal, or enteroenteric anastomoses not otherwise counted as right-sided colonic reconstruction. The upper gastrointestinal family included gastric and esophagogastric restorative procedures, including esophagectomy. This grouping yielded 15,904 right-sided colonic procedures, 11,436 left-sided colonic procedures, 9,128 low pelvic colorectal procedures, 9,502 small-bowel procedures, and 8,940 upper gastrointestinal reconstructions.

Care setting was intentionally defined at the episode level rather than by hospital label alone. The goal was to capture the organizational environment visible to clinicians and relevant to early postoperative prediction. Five settings were used. The elective enhanced recovery

setting included scheduled operations admitted through protocolized perioperative pathways emphasizing same-day admission, early mobilization, and standardized postoperative orders. The elective conventional inpatient setting included scheduled operations managed without formal enhanced recovery infrastructure, often with more variable preoperative timing and postoperative order sets. The urgent inpatient setting captured operations performed during an unplanned admission but not under overnight emergency conditions, usually for obstruction, perforation risk, failed nonoperative management, or accelerated cancer symptoms. The emergency overnight setting included unscheduled procedures starting between 18:00 and 06:00 or documented as immediate/emergent. The transfer-in complex care setting comprised admissions transferred from external hospitals for specialized resection, redo surgery, or management of severe postoperative complications preceding the index restorative procedure. These settings produced 19,612 elective enhanced recovery episodes, 13,784 elective conventional inpatient episodes, 9,413 urgent inpatient episodes, 7,128 emergency overnight episodes, and 4,973 transfer-in complex care episodes.

Crossing the five procedure families with the five care settings yielded twenty-five clinical domains. Each domain was treated as a distinct source or target environment for transport analysis. The domains were not balanced. Right-sided colonic surgery was heavily represented in elective enhanced recovery and underrepresented in transfer-in complex care. Upper gastrointestinal reconstruction was concentrated in elective conventional and transfer-in complex care settings. Low pelvic colorectal surgery appeared across all five settings, including a clinically important number of emergency overnight and urgent inpatient cases [8]. These imbalances were useful rather than problematic because they created realistic transport conditions. The fundamental question of interest was how a model learned on some domains behaves when moved into another, not how it performs under artificial domain symmetry.

The outcome was clinically managed anastomotic leak within 30 days of the index operation. Rather than rely on diagnosis codes alone, the endpoint was defined through a composite requiring anatomical compatibility and a management-linked or confirmatory component. Compatible evidence included radiologic contrast extravasation, peri-anastomotic or mediastinal collection explicitly interpreted as suspicious for leak, operative confirmation of dehiscence, endoscopic demonstration of defect, or surgeon-documented leak in the procedural record. Confirmatory management included reoperation, image-guided drainage, endoscopic therapy, leak-specific broad-spectrum antimicrobial escalation associated with a compatible source, or readmission explicitly for leak management. Events recognized after discharge but within thirty days were counted when they met the same criteria. This endpoint deliberately emphasized clinically managed leakage because the operational interest of the study lay in surveillance and escalation rather than in subclinical radiographic findings.

Outcome ascertainment was performed in two stages. A rule-based extraction first flagged potentially relevant admissions using diagnosis strings, imaging reports, drainage procedures, endoscopy codes, reoperation records, and antibiotic escalation patterns. A second stage involved targeted manual review of uncertain cases sampled from high-probability and medium-probability strata, along with all cases in low-frequency domains where endpoint ambiguity would disproportionately affect calibration. Reviewers were blinded to model predictions. Final adjudication yielded 3,625 leaks, corresponding to an overall incidence of 6.6%. Event rates differed substantially across procedure families and care settings, but those crude differences were not interpreted at face value because the entire purpose of the analysis was to understand how domain mixture interacts with prediction validity.

The procedural families showed leak rates consistent with their expected technical and physiologic burdens. Right-sided colonic procedures had 541 leaks, or 3.4%. Left-sided colonic

procedures had 595 leaks, or 5.2%. Low pelvic colorectal procedures had 1,059 leaks, or 11.6% [9]. Small-bowel procedures had 447 leaks, or 4.7%. Upper gastrointestinal reconstruction had 983 leaks, or 11.0%. Care settings also differed sharply. Elective enhanced recovery episodes had 831 leaks, or 4.2%. Elective conventional inpatient episodes had 845 leaks, or 6.1%. Urgent inpatient episodes had 699 leaks, or 7.4%. Emergency overnight episodes had 606 leaks, or 8.5%. Transfer-in complex care episodes had 644 leaks, or 13.0%. These percentages should not be read as setting effects, because the domains were populated by very different procedure mixes and baseline physiologies. They do, however, illustrate why a single pooled prevalence is insufficient for thresholding.

The highest-risk domains were concentrated where technical complexity and physiologic disruption overlapped. Upper gastrointestinal procedures in transfer-in complex care had a crude leak incidence of 16.8%. Low pelvic colorectal procedures in emergency overnight care had a crude leak incidence of 14.9%, while the same family in urgent inpatient care had 13.2%. At the other end of the spectrum, right-sided colonic surgery in elective enhanced recovery had a crude incidence of 2.8%. These contrasts matter not only because the event burden changes, but because the model's positive predictive value at a given threshold becomes a moving target when transported from one domain to another. A threshold that is acceptable in a domain with 3% risk can become operationally inefficient in a domain with 15% risk, and vice versa.

The time anchor for feature extraction was set at 18 hours after skin closure. This time point reflects a practical compromise [10]. It is late enough for immediate postoperative physiology to declare itself in most patients and early enough to inform surveillance decisions before the usual window of formal leak recognition. Patients discharged before the 18-hour mark were rare and were retained only if their final observation window still included a qualifying feature set; otherwise they were excluded because the intended use case assumed at least one early postoperative reassessment. The use of a fixed early window made the predictive task static rather than time-updated. That was deliberate. The present study was not a dynamic hazard model but a transport study asking how a single early prediction behaves across domains.

Several additional cohort descriptors were relevant to domain construction. Median age was 63 years, but the distribution varied, with right-sided colonic procedures skewing older and upper gastrointestinal reconstruction skewing younger but more nutritionally depleted. Preoperative albumin was lowest in transfer-in complex care and upper gastrointestinal domains. Emergency overnight domains had the highest proportion of contamination, vasopressor exposure, open surgery, and postoperative ICU admission. Enhanced recovery domains had the shortest operations on average and the lowest transfusion rates, although low pelvic colorectal surgery within enhanced recovery still displayed long operative times relative to other enhanced recovery procedures. These differences underscored the coexistence of procedure heterogeneity and care-setting heterogeneity. They also suggested that transport failure, if observed, would likely arise from a mixture of prevalence differences, covariate shift, and changes in predictor meaning.

3. Feature Space, Missingness Structure, and Quantification of Domain Shift

The predictor set was intentionally restricted to information that could be expected in routine structured data by 18 hours after the end of surgery. This produced a feature space broad enough for transport analysis and narrow enough to preserve operational relevance. Ninety-two variables were retained after quality screening. The baseline block included age, sex, body mass index, smoking status, diabetes, chronic kidney disease, chronic lung disease, cardio-

vascular disease, immunosuppression, active malignancy, prior abdominal surgery, admission source, and a claims-augmented frailty score. The preoperative laboratory block included hemoglobin, albumin, bilirubin, creatinine, sodium, white blood cell count, and C-reactive protein when available. The intraoperative block captured urgency, operative duration, open versus minimally invasive approach, conversion, diversion status, blood loss, transfusion, vasopressor burden, wound contamination class, stoma creation, and case start time [11]. The early postoperative block contained highest heart rate, highest temperature, first and worst lactate, white blood cell count, hemoglobin drop, creatinine change, need for vasopressor continuation, opioid intensity, ventilatory support at arrival, destination ward, and number of laboratory panels ordered within the 18-hour window.

The variable capturing number of laboratory panels deserves attention because it functioned partly as a data-density measure and partly as a clinical intensity measure. In many domains, higher data density implied closer observation, more physiologic concern, or both. Excluding this variable would have hidden a meaningful part of the data-generating process. Including it created the possibility that a model would learn monitoring behavior rather than patient risk. The study therefore retained the feature but treated its domain interaction carefully within the transport framework. More broadly, the analysis did not assume that measurement intensity was noise. The way a patient is measured is part of the care setting, and transport across settings becomes less credible if that reality is ignored.

Continuous variables were winsorized at the 0.5th and 99.5th percentiles to reduce the influence of likely recording errors while preserving clinically meaningful extremes. They were then standardized using training-fold statistics only. Categorical variables with more than two levels were encoded through regularized target-aware contrasts within training folds, while simpler binary indicators were used whenever direct clinical meaning was clear. Missingness was nontrivial and itself heterogeneous across domains. Preoperative albumin was missing in 18.7% of elective enhanced recovery episodes and only 7.9% of transfer-in complex care episodes. C-reactive protein was sparse in conventional elective and enhanced recovery settings but common in urgent and emergency settings. Lactate measurement density was highest in upper gastrointestinal and emergency overnight domains. Because missingness was partly informative and partly operational, a naive complete-case strategy would have disproportionately removed low-intensity domains and biased the transport problem in favor of high-acuity care.

The primary imputation strategy combined low-rank matrix completion within the training set with explicit missingness indicators. Let $\mathbf{X} \in \mathbb{R}^{N \times p}$ denote the partially observed feature matrix. We estimated a completed matrix $\widehat{\mathbf{X}}$ by solving

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{M}} \left\| \mathcal{P}_{\Omega}(\mathbf{X} - \mathbf{M}) \right\|_F^2 + \lambda_* \|\mathbf{M}\|_* + \lambda_d \sum_{d=1}^{D} \left\| \mathbf{M}_{(d)} - \bar{\mathbf{M}} \right\|_F^2 \tag{1}$$

where $\mathcal{P}_{\Omega}$ projects onto observed entries, $\|\cdot\|_*$ is the nuclear norm, $\mathbf{M}_{(d)}$ denotes the block corresponding to domain d, and $\bar{\mathbf{M}}$ is the pooled mean structure. The first penalty encouraged low-rank recovery, while the second prevented domain-specific overfitting of the imputation surface [12]. Separate binary indicators were appended for every imputed variable. This approach preserved cross-variable dependence better than univariate median imputation while

allowing the final predictive models to learn whether absence of a value carried additional signal.

The resulting feature geometry differed visibly across domains. Emergency overnight small-bowel surgery occupied a high-contamination, high-lactate, high-vasopressor region of the space. Upper gastrointestinal transfer-in complex care occupied a long-duration, low-albumin, high-monitoring-density region. Right-sided colonic surgery in enhanced recovery clustered around shorter duration, lower lactate, fewer laboratory panels, and earlier ward destination. Low pelvic colorectal surgery in elective enhanced recovery shared some enhanced recovery traits but remained separated by longer operative times, higher diversion rates, more pelvic drains, and more frequent postoperative hemoglobin decline. These patterns suggested that transport between some domains would be modest, while others would involve large covariate shifts even before considering outcome prevalence.

To formalize that intuition, the study computed several measures of domain distance. For domains d and d', the first measure was a covariance-aware Wasserstein distance between their standardized continuous-feature distributions. The second was a Jensen-Shannon divergence between marginal outcome prevalences. The third was a kernel maximum mean discrepancy over the full feature set after encoding. The fourth was a conditional-shift proxy derived from the change in calibration slope when a pooled model trained on all other domains was applied separately to d and d'. These components were combined into a domain transfer difficulty index defined as

$$\begin{aligned} \mathrm{DTDI}_{d\to d'} = {} & \alpha\, W_2\left(P_d^X, P_{d'}^X\right) + \beta\, \mathrm{JSD}\left(P_d^Y, P_{d'}^Y\right) \\ & + \gamma\, \mathrm{MMD}\left(P_d^X, P_{d'}^X\right) + \delta\, \left|\kappa_d - \kappa_{d'}\right| \end{aligned} \tag{2}$$

where κ_d denotes the domain-specific calibration slope under the pooled source model. The weights $\alpha, \beta, \gamma, \delta$ were selected within training folds to maximize explained variance of observed transport loss in held-out validation domains. This index was not treated as a mechanistic decomposition [13]. It was a compact empirical summary of how difficult it was likely to be to move a model from one domain to another.

Average DTDI values varied widely. Transfer from right-sided colonic enhanced recovery to left-sided colonic conventional elective care was relatively small. Transfer from any elective colonic domain into upper gastrointestinal transfer-in complex care was large. Low pelvic colorectal emergency overnight formed a distinct region because it combined high baseline risk with unusual contamination, night-start patterns, lower rates of minimally invasive surgery, and high ICU utilization. Importantly, high domain distance was not synonymous with high event rate. Some high-risk domains were still fairly predictable if their feature geometry resembled the training data. Conversely, some moderate-risk domains were difficult targets because the same postoperative measurements had different implied meanings.

The feature-level contribution to domain separation was also quantified by a regularized linear discriminant projection. Operative duration, preoperative albumin, laboratory panel count, postoperative lactate, vasopressor continuation, contamination class, and diversion status dominated the first two discriminant axes, but their domain-specific interpretation differed. High laboratory panel count separated upper gastrointestinal and emergency domains from enhanced recovery, yet within upper gastrointestinal care it could reflect routine surveillance rather than instability. Diversion separated low pelvic colorectal surgery from the other

families, but its association with leak was not uniform because diversion both marked technical difficulty and modified the clinical threshold for concern. These ambiguities are exactly the setting in which pooled models often overgeneralize.

To evaluate whether observed domain distance was merely a hospital artifact, a nested variance decomposition was performed. Each feature was regressed on procedure family, care setting, hospital, and their interactions using cross-fitted regularized models. Procedure family and care setting jointly explained more variance than hospital alone for most high-importance features. For postoperative lactate, the explained variance attributable to family and setting was 29%, compared with 6% for hospital. For operative duration the corresponding values were 34% and 8% [14]. For albumin, family and setting explained 17%, hospital 4%. This does not imply that hospital-level practice is irrelevant. It suggests, rather, that the dominant transport problem in this cohort was not simply moving across institutions. It was moving across clinically distinct surgery-setting domains that happened to exist both within and between institutions.

A secondary analysis examined prevalence shift separately from feature shift. The raw 6.6% overall leak rate concealed a spread from 2.8% to 16.8% across the twenty-five domains. When prevalence alone was equalized through intercept recalibration, discrimination remained poor in several high-distance targets, particularly upper gastrointestinal transfer-in complex care and low pelvic colorectal emergency overnight. This observation indicated that transport failure could not be solved by baseline-rate correction alone. The structure of the predictors, and their relationship with the outcome, also changed. That empirical result informed the choice of predictive framework. A simple recalibration strategy might help, but it would not be sufficient as the main method.

4. Cross-Domain Modeling Framework

The primary predictive objective was to estimate the probability of clinically managed anastomotic leak within 30 days using structured variables available by 18 hours after surgery, with particular emphasis on performance in domains absent from model training. Let $x_i \in \mathbb{R}^p$ denote the feature vector for patient i, $y_i \in \{0,1\}$ the leak outcome, and $d_i \in \{1,\dots,25\}$ the clinical domain determined by procedure family and care setting. The central challenge was to learn a function that retained useful signal across domains without either collapsing meaningful contextual differences or overfitting unstable domain-specific idiosyncrasies. Four model families were compared.

The first comparator was an elastic-net logistic regression fitted on the pooled cohort. This model provided an interpretable low-capacity benchmark and allowed straightforward inspection of global coefficients. The second comparator was pooled gradient boosting using shallow trees and shrinkage, selected because tree ensembles often perform strongly on structured perioperative data and can capture nonlinear thresholds without manual feature crosses. The third comparator consisted of separate domain-specific gradient boosting models trained independently within each domain whenever sample size permitted [15]. This local strategy was expected to fit well in common domains and poorly in sparse ones, but it served as an informative boundary case. The fourth and primary model was a domain-mixture architecture combining invariant representation learning, domain-conditioned gating, and calibration regularization. The rationale was to allow partial sharing of signal across domains while retaining a structured way for predictor meaning to change.

The primary model used an encoder f_θ mapping the original feature vector into a lower-dimensional representation $z_i = f_\theta(x_i) \in \mathbb{R}^r$. A set of M experts then produced domain-

sensitive log-odds contributions. The gating function depended on both the encoded features and the observed domain label, so the model could upweight different experts for different surgery-setting contexts. The predicted probability took the form

$$\begin{aligned} \hat{p}_i &= \sum_{m=1}^{M} \pi_m(z_i, d_i)\, \sigma\Big(a_m^\top z_i + b_m + \tau_{c(i)}\Big) \\ \pi_m(z_i, d_i) &= \frac{\exp\big(u_m^\top z_i + v_m^\top e_{d_i}\big)}{\sum_{\ell=1}^{M} \exp\Big(u_\ell^\top z_i + v_\ell^\top e_{d_i}\Big)} \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the logistic map, e_{d_i} is a one-hot embedding of domain membership, and $\tau_{c(i)}$ is a calibration offset indexed by a latent domain cluster $c(i)$ rather than the exact domain. Using clusters instead of per-domain offsets reduced instability in sparse targets and was closer to a realistic deployment setting in which a new domain may resemble an existing cluster without having abundant local labels.

The encoder was trained under an adversarial invariance constraint so that some components of z_i captured cross-domain structure while others remained available for domain-conditioned adaptation. Specifically, the representation was partitioned into z_i^{inv} and z_i^{ctx}. A domain discriminator attempted to predict d_i from z_i^{inv}, while the encoder was trained to make that prediction difficult. The objective encouraged the invariant subspace to retain information useful for leak prediction but less tied to domain identity. The contextual subspace remained free to encode domain-specific patterns for use by the gating mechanism. Formally, the training loss was

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{pred}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{dro}}\mathcal{L}_{\text{dro}} + \lambda_{\text{cal}}\mathcal{L}_{\text{cal}} \\ \mathcal{L}_{\text{pred}} &= -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\Big] \end{aligned} \tag{4}$$

with

$$\begin{aligned} \mathcal{L}_{\text{adv}} &= -\frac{1}{N}\sum_{i=1}^{N} \log q_\Phi\Big(d_i \mid z_i^{\text{inv}}\Big) \\ \mathcal{L}_{\text{dro}} &= \max_{d \in \mathcal{D}} \frac{1}{n_d} \sum_{i:\, d_i = d} \ell_i(\hat{p}_i, y_i) \\ \mathcal{L}_{\text{cal}} &= \sum_{k=1}^{K} \Big(\bar{y}_k - \bar{p}_k\Big)^2 \end{aligned} \tag{5}$$

where $\mathcal{L}_{\text{adv}}$ entered with gradient reversal at the encoder, $\mathcal{L}_{\text{dro}}$ imposed a group distributionally robust penalty on the worst-performing training domain, and $\mathcal{L}_{\text{cal}}$ penalized miscalibration across probability bins within each calibration cluster. The calibration cluster index k referred to latent clusters of domains learned from the DTDI matrix and source-domain residual structure. This design was chosen to align the model with the actual transport question.

The goal was not only to optimize average fit, but to limit failure in the hardest domains while keeping risk outputs numerically coherent.

A practical advantage of this architecture was that it separated two adaptation tasks often conflated in clinical modeling. The encoder and experts handled feature-response structure. The calibration offsets handled differences in scale and threshold meaning. If a new domain shared physiology with existing domains but differed mainly in prevalence, the cluster-level calibration term could adjust without relearning the full representation [16]. If the new domain differed more deeply, the gating mechanism could reweight experts based on its encoded feature profile and cluster membership. This was not full meta-learning in the sense of rapid updating from a few labeled target samples, because the primary evaluation withheld entire domains. However, the architecture was designed so that limited future recalibration data could be incorporated without rebuilding the whole model.

The training process used nested validation. The outer loop withheld one domain at a time as the transport target. All admissions from that domain were excluded from model fitting. The inner loop operated on the remaining twenty-four domains to tune hyperparameters using grouped folds that preserved domain integrity. Performance was then measured on the held-out target domain. This leave-domain-out design was the study’s primary validation strategy. Random train-test splits were also computed for reference, but only to quantify how much conventional internal validation overstated transport performance. Because some domains were concentrated in specific hospitals, a secondary sensitivity analysis additionally withheld the hospitals contributing the highest share of a target domain. That sensitivity analysis produced somewhat lower overall performance but the same ranking of models, which supported the interpretation that domain transport, not mere hospital memorization, drove the main findings.

The number of experts M was set to four after inner-loop comparison of values from two to six. Fewer experts left too little flexibility for context-specific risk structure. More experts increased variance without meaningful gain. The representation dimension was set to twenty-four, split evenly between invariant and contextual subspaces. Optimization used weighted mini-batches that oversampled low-frequency domains but preserved outcome prevalence within domain. Class imbalance was handled through a focal cross-entropy variant in a sensitivity analysis, although the main results used standard cross-entropy because focal weighting modestly improved recall at the expense of worse calibration [17]. Model selection therefore prioritized retained area under the receiver operating characteristic curve, area under the precision-recall curve, Brier score, calibration slope, expected calibration error, and the domain-averaged deviation of positive predictive value at a fixed alert threshold.

To connect performance loss with measurable domain difference, a transport regression was fitted after model evaluation. For a model m and target domain d, let Δ_{md} denote the drop in performance from pooled random-split evaluation to leave-domain-out evaluation. This transport gap was regressed on the components of DTDI and on domain support size:

$$\begin{aligned}\Delta_{md} = \beta_0 + \beta_1 W_2^{(d)} + \beta_2 \mathrm{JSD}^{(d)} + \beta_3 \mathrm{MMD}^{(d)} \\ + \beta_4 \left|\kappa^{(d)}\right| + \beta_5 \log n_d + \varepsilon_{md}\end{aligned} \tag{6}$$

This auxiliary model was not used for prediction. Its purpose was to estimate which components of shift explained most of the transport gap for each modeling strategy. If pooled models were chiefly harmed by prevalence shift, recalibration might suffice. If they were chiefly

harmed by conditional or geometric shift, richer transport-aware learning would be justified. In practice, the estimated coefficients showed mixed drivers, with the relative importance varying by model class.

A final operational layer converted probabilities into alerts. Since clinical deployment would require a threshold, the study examined how a single global threshold performed across domains compared with a family of cluster-calibrated thresholds chosen to equalize positive predictive value within a narrow target range. Threshold design was not treated as an afterthought because calibration errors become management errors once predictions are binarized. A model with similar discrimination in two domains can still yield clinically different workloads if its risk scale stretches differently. The threshold analysis therefore provides a direct link between transport statistics and likely bedside consequences.

5. Results

Table 1. Cohort composition and leak incidence

Category	Procedures	Leaks	Incidence (%)
Total cohort	54,910	3,625	6.6
Right-sided colonic	15,904	541	3.4
Left-sided colonic	11,436	595	5.2
Low pelvic colorectal	9,128	1,059	11.6
Small bowel	9,502	447	4.7
Upper gastrointestinal	8,940	983	11.0

Table 2. Leak incidence across care settings

Care setting	Episodes	Leaks	Incidence (%)
Elective enhanced recovery	19,612	831	4.2
Elective conventional	13,784	845	6.1
Urgent inpatient	9,413	699	7.4
Emergency overnight	7,128	606	8.5
Transfer-in complex care	4,973	644	13.0

Table 3. Representative high- and low-risk domains

Domain	Procedures	Leaks	Incidence (%)
UGI transfer-in complex	–	–	16.8
Low pelvic emergency overnight	–	–	14.9
Low pelvic urgent inpatient	–	–	13.2
Right colon enhanced recovery	–	–	2.8
Elective colonic average	–	–	~4–6
Mixed pooled average	54,910	3,625	6.6

The 54,910 procedures in the analytic cohort yielded 3,625 clinically managed leaks within 30 days, for an overall incidence of 6.6%. The domain composition was markedly uneven. Right-sided colonic surgery accounted for 41.8% of the elective enhanced recovery setting but only 9.4% of transfer-in complex care. Upper gastrointestinal reconstruction accounted for

Table 4. Model performance under random-split validation

Model	AUROC	ECE	Brier
Elastic-net logistic	0.774	0.061	–
Gradient boosting	0.812	0.046	–
Domain-mixture	0.821	0.031	–
Domain-specific	–	–	–
Reference baseline	~0.75	~0.06	–

Table 5. Performance under leave-domain-out transport evaluation

Model	AUROC	PR-AUC	ECE
Elastic-net logistic	0.711	0.244	0.092
Gradient boosting	0.736	0.271	0.083
Domain-specific	0.702	–	–
Domain-mixture	0.781	0.302	0.034
Sensitivity analysis mean	↓0.02	↓	↑

34.7% of transfer-in complex care but only 8.1% of elective enhanced recovery. Low pelvic colorectal surgery was present in all settings, including 1,136 emergency overnight cases and 1,492 urgent inpatient cases [18]. These two low pelvic domains carried some of the highest observed leak burdens. Small-bowel procedures were concentrated in urgent inpatient and emergency overnight care and contributed much of the contamination burden in the cohort. This domain imbalance was the first indication that pooled performance metrics would be difficult to interpret without transport testing.

Feature distributions tracked clinical expectations while highlighting nontrivial overlap. Preoperative albumin was lowest in upper gastrointestinal transfer-in complex care, where the median was 3.1 g/dL, and highest in right-sided colonic enhanced recovery, where the median was 3.8 g/dL. Operative duration was shortest in right-sided colonic enhanced recovery and longest in upper gastrointestinal transfer-in complex care, with low pelvic colorectal elective procedures occupying an intermediate but wide range. Postoperative lactate displayed both procedure and setting effects. Emergency overnight domains had the highest median lactate overall, but upper gastrointestinal domains showed the widest right tail. Laboratory panel count, used as a proxy for monitoring intensity, was low in enhanced recovery domains and highest in transfer-in complex care and emergency overnight settings. ICU admission within the first 18 hours was routine in some upper gastrointestinal contexts and rare in enhanced recovery right-sided colonic care. These differences meant that the same measured variables were entering the model under different clinical conventions.

The domain transfer difficulty matrix reflected this heterogeneity. The smallest DTDI values occurred among elective colonic domains, especially transfer between right-sided and left-sided colonic surgery within elective settings. Moderate distances linked low pelvic colorectal elective domains to upper gastrointestinal conventional inpatient care, suggesting some shared complexity but divergent feature distributions. The largest distances consistently involved transfer from low-acuity elective colonic domains into upper gastrointestinal transfer-in complex care and low pelvic emergency overnight surgery. These target domains combined prevalence shift, covariate shift, and calibration deformation. Notably, some within-procedure transfers were harder than some cross-procedure transfers. For example, low pelvic enhanced recovery to low pelvic emergency overnight transport was more difficult than low pelvic en-

Table 6. Transport gaps in selected target domains

Target domain	GB AUROC	DM AUROC	Cal. slope (GB/DM)
Right colon elective	0.784	0.801	0.90 / 0.97
Left colon elective	0.776	0.793	0.88 / 0.96
Low pelvic emergency	0.694	0.755	0.66 / 0.95
UGI transfer-in	0.682	0.756	0.61 / 0.93
Small bowel emergency	0.721	0.768	0.72 / 0.94

Table 7. Contribution of domain shift components to transport loss

Component	Elastic-net	GB	Domain-mixture
Covariate shift	Moderate	High	Reduced
Prevalence shift	Moderate	Low–moderate	Reduced
Calibration deformation	Moderate	High	Low
Support size	Low	Moderate	Moderate
Residual effects	Present	Present	Lower

hanced recovery to upper gastrointestinal conventional elective transport, indicating that care setting altered predictor meaning substantially even when the operation family remained fixed [19].

Random-split validation overstated performance for all models. Under conventional pooled splits, elastic-net logistic regression achieved an area under the receiver operating characteristic curve of 0.774 and an expected calibration error of 0.061. Pooled gradient boosting achieved 0.812 and 0.046, respectively. The domain-mixture model achieved 0.821 and 0.031. Domain-specific models could not be meaningfully summarized under random splits because the design contradicted their local scope, but in large domains they often matched or slightly exceeded pooled models when evaluated within the same domain. These values might have been taken as evidence of near-deployment performance if transport had not been examined.

Under leave-domain-out testing, the picture changed substantially. Elastic-net logistic regression fell to a mean area under the receiver operating characteristic curve of 0.711, mean precision-recall area of 0.244, Brier score of 0.063, and expected calibration error of 0.092. Pooled gradient boosting retained more discrimination at 0.736 and a precision-recall area of 0.271, but its calibration deteriorated materially, with a Brier score of 0.060, expected calibration error of 0.083, and mean calibration slope of 0.74. The domain-specific models, trained only within source domains closest to the target when sufficient support existed, performed inconsistently. Their average target area under the receiver operating characteristic curve was 0.702, but the variance was large and performance in sparse or distant targets was poor. The primary domain-mixture model produced the strongest overall transport profile, with mean area under the receiver operating characteristic curve of 0.781, precision-recall area of 0.302, Brier score of 0.056, expected calibration error of 0.034, and mean calibration slope of 0.97. The improvement was not uniform across domains, but it was directionally consistent in nearly all of them.

Transport loss varied sharply by target domain. For pooled gradient boosting, the smallest transport gap occurred in right-sided colonic conventional elective care, where the area under the receiver operating characteristic curve declined by only 0.028 from random-split expectation. Similarly small losses were observed in left-sided colonic elective conventional

Table 8. Threshold performance variability across domains

Model/strategy	PPV range	Sensitivity range	Alert burden
GB global threshold	0.14–0.46	0.28–0.71	Highly variable
DM cluster-calibrated	0.26–0.36	0.44–0.61	More uniform
Elastic-net global	0.12–0.40	0.25–0.65	Variable
Domain-specific	Inconsistent	Inconsistent	Unstable
Reference midpoint	∼0.30	∼0.50	Balanced

Table 9. Key predictors in invariant vs contextual subspaces

Feature	Invariant importance	Contextual importance	Role
Albumin	High	Low	Baseline physiology
Lactate	High	Moderate	Perfusion marker
WBC count	High	Low	Inflammation
Lab panel count	Low	High	Monitoring intensity
ICU admission	Low	High	Care setting proxy
Operative duration	High	Moderate	Technical burden

care and right-sided colonic urgent inpatient care. The largest transport gaps occurred in upper gastrointestinal transfer-in complex care, where the pooled gradient boosting model fell to 0.682 with a calibration slope of 0.61, and in low pelvic colorectal emergency overnight care, where it fell to 0.694 with a calibration slope of 0.66. In these targets the model systematically underpredicted high-risk patients and compressed probability estimates toward the middle [20]. The domain-mixture model improved both discrimination and scale. In upper gastrointestinal transfer-in complex care it achieved an area under the receiver operating characteristic curve of 0.756 and calibration slope of 0.93. In low pelvic colorectal emergency overnight care it achieved 0.755 and 0.95, respectively. Those gains are clinically relevant because these are precisely the domains where false reassurance would be hardest to justify.

The observed errors were not reducible to baseline prevalence differences alone. A simple intercept recalibration applied to the pooled gradient boosting model improved expected calibration error modestly, from 0.083 to 0.065 on average, but produced little change in discrimination and left the hardest targets poorly aligned. Upper gastrointestinal transfer-in complex care, for instance, remained underdispersed after intercept correction, with many truly high-risk patients assigned probabilities only slightly above moderate-risk elective upper gastrointestinal cases. Similarly, low pelvic colorectal emergency overnight care retained both rank errors and scale errors after baseline adjustment. This indicates that label shift contributed to transport loss but did not dominate it.

The transport regression clarified this point. For elastic-net logistic regression, prevalence divergence measured by Jensen-Shannon distance explained a meaningful portion of the transport gap, but covariate geometry and conditional-shift proxies were also important. For pooled gradient boosting, the strongest predictors of transport loss were Wasserstein distance and the calibration-slope disparity term, while pure prevalence divergence had a smaller coefficient. In the final specification, approximately 47% of the explained transport gap was attributable to feature-space distance, 21% to prevalence divergence, 18% to calibration deformation, and the remainder to support size and residual interaction. For the domain-mixture model, the same explanatory variables had attenuated coefficients, especially the calibration component. That pattern is consistent with a model that does not eliminate domain shift

but is less vulnerable to it.

The learned latent domain clusters were clinically interpretable. Four clusters emerged. The first contained low-acuity elective colonic domains with short operations, higher ward destination rates, and low monitoring density [21]. The second contained technically complex but controlled elective domains, including low pelvic enhanced recovery and upper gastrointestinal conventional elective care. The third contained urgent and emergency bowel surgery with high contamination and vasopressor exposure. The fourth contained transfer-in complex care and a subset of upper gastrointestinal and redo domains with low albumin, high monitoring density, and long operative times. Cluster-level calibration offsets reduced instability compared with exact-domain offsets. When exact-domain post-hoc scaling was attempted, sparse domains showed overcorrection and unstable probability tails. Cluster-based scaling preserved smoother calibration and proved more robust in nested validation.

Feature importance patterns differed between the pooled gradient boosting model and the domain-mixture architecture. In the pooled model, postoperative lactate, laboratory panel count, operative duration, albumin, ICU admission, diversion, vasopressor continuation, and contamination dominated the global ranking. In the domain-mixture model, the invariant subspace emphasized albumin, lactate, white blood cell count, operative duration, and hemoglobin drop, while the contextual subspace assigned greater weight to laboratory panel count, ICU destination, night-start status, diversion, and transfer-in admission. This partition aligns with clinical intuition. Some predictors carry cross-domain physiologic meaning, while others are partly markers of care-process context. The separation likely contributed to improved transport because it prevented the model from treating every high-intensity care-process variable as though its leak implication were uniform.

The effect of domain-aware learning was especially evident in probability calibration. At a nominal predicted risk of 0.20, the pooled gradient boosting model corresponded to observed event rates ranging from 0.11 in right-sided colonic enhanced recovery to 0.29 in upper gastrointestinal transfer-in complex care. The domain-mixture model narrowed this range to 0.17 through 0.23. This reduction matters even more than modest changes in discrimination once a probability is used to trigger imaging, prolonged observation, or senior review. Calibration curve inspection showed that the pooled model was too optimistic in high-risk emergency and transfer domains and too pessimistic in some low-risk elective domains. The transport-aware model pulled both extremes toward better numerical agreement without collapsing domain differences into a single average [22].

Threshold analysis provided the most direct operational picture. Using a single 8% alert threshold with pooled gradient boosting yielded positive predictive values ranging from 0.14 in right-sided colonic enhanced recovery to 0.46 in upper gastrointestinal transfer-in complex care. Sensitivity ranged from 0.28 to 0.71 across domains. In practice, this means the same threshold would create low-yield alert burden in some domains and miss a large share of events in others. Applying cluster-calibrated thresholds to the domain-mixture model reduced the positive predictive value range to 0.26 through 0.36 while keeping sensitivity between 0.44 and 0.61. Alert burden per 100 cases became substantially more uniform. This threshold harmonization did not equalize domains completely, nor would that be expected or necessarily desirable. It did, however, remove much of the arbitrary variation produced by transporting a pooled score without contextual calibration.

The performance of domain-specific local models highlights why transport-aware pooling was preferable to full localization. In high-support domains such as right-sided colonic enhanced recovery and left-sided colonic conventional elective care, domain-specific models sometimes

matched the domain-mixture model. Yet in low pelvic emergency overnight care and upper gastrointestinal transfer-in complex care, the local models were unstable, with calibration slopes varying widely across folds and average precision-recall area no better than the pooled gradient boosting model. These domains had sufficient events to matter clinically but not enough to sustain isolated model development without borrowing structure from related domains. The domain-mixture approach was specifically designed for this middle ground and appears to have used it effectively.

A secondary sensitivity analysis withheld, in addition to the target domain, the hospitals contributing the largest share of that domain. This created a more severe test by removing both domain examples and much of the hospital-specific context associated with them. Performance decreased for all models, but the relative ranking remained. Mean area under the receiver operating characteristic curve fell by 0.021 for elastic-net logistic regression, 0.027 for pooled gradient boosting, and 0.019 for the domain-mixture model [23]. Expected calibration error increased by 0.012, 0.015, and 0.009, respectively. The most affected targets were again low pelvic emergency overnight and upper gastrointestinal transfer-in complex care. The persistence of the model ranking under this stricter design suggests that the primary findings were not an artifact of residual hospital similarity.

Another sensitivity analysis removed laboratory panel count and ICU destination from the predictors, on the grounds that they might encode care-process intensity rather than patient state. This harmed performance in all models, but the harm was largest in the pooled gradient boosting model. Its mean transport area under the receiver operating characteristic curve fell from 0.736 to 0.719. The domain-mixture model declined from 0.781 to 0.768. Calibration also worsened. This pattern suggests that care-process features were not simply shortcuts causing overfitting. When modeled carefully, they contributed useful information about the environment in which physiology was being observed. Their inclusion became problematic mainly when the model lacked a mechanism to separate invariant from contextual meaning.

A final set of analyses examined model behavior within the highest-risk procedure families. Restricting to low pelvic colorectal surgery alone, pooled gradient boosting performed adequately within elective domains but degraded sharply in urgent and emergency settings, especially at the high-probability tail. The domain-mixture model retained better tail calibration and showed a 0.049 mean gain in precision-recall area across the five low pelvic domains. Restricting to upper gastrointestinal reconstruction, the pooled model underpredicted transfer-in complex care and overpredicted elective conventional care. The domain-mixture model reduced this asymmetry and achieved a 0.058 mean gain in area under the receiver operating characteristic curve across upper gastrointestinal domains. These family-restricted findings reinforce the broader result that transport is not merely a hospital issue. Within a single surgery family, setting still changes how well a model travels [24].

Taken together, the results support three main empirical statements. First, random internal validation materially exaggerated the apparent readiness of all models for deployment across heterogeneous postoperative gastrointestinal care. Second, transport loss was concentrated in specific surgery-setting domains rather than evenly distributed, with the largest failures arising where technical complexity and organizational complexity overlapped. Third, a model designed to share invariant signal while representing contextual drift explicitly preserved more discrimination and much better calibration than either naive pooling or full localization. The improvement was not a universal solution. It did not eliminate heterogeneity, and it did not make all domains equally easy. It did, however, convert some of the hardest domains from poorly calibrated targets into reasonably interpretable ones.

6. Discussion

The central result of this study is that heterogeneity of anastomotic leak risk across surgery types and care settings is also heterogeneity of model transport. That is a more specific claim than saying that some procedures and settings are riskier than others. It means that the statistical relationship between early postoperative features and leak is itself domain-sensitive, and that a model validated on pooled samples cannot be assumed to preserve meaning when moved into a new surgery-setting context. The empirical evidence for this claim appears in two linked observations. First, conventional random-split validation overstated performance for every model. Second, the discrepancy between pooled validation and target-domain performance was concentrated in clinically recognizable domains rather than scattered at random. Those domains were not obscure edge cases. They included low pelvic colorectal emergency surgery and upper gastrointestinal transfer-in complex care, both settings in which risk communication and surveillance are especially consequential.

One implication is methodological. In postoperative leak prediction, internal validity and transport validity are not near substitutes. A model can be internally strong yet operationally unreliable because the average validation sample contains the same mixture of procedure families and care settings as the training sample [25]. Such evaluation is not wrong, but it answers a different question. It estimates performance for a future sample drawn from the same domain mixture. Hospitals and service lines rarely deploy models into such abstract mixtures. They deploy them into specific pathways, often precisely the pathways whose prevalence and feature distributions differ from the training average. For this reason, leave-domain-out evaluation is not an exotic robustness exercise here. It is closer to the actual deployment problem than the more familiar random split.

A second implication concerns the nature of domain shift. The results show that prevalence shift mattered, but it was not the dominant source of transport loss. If prevalence shift had been primary, intercept recalibration would have solved most of the problem. It did not. Covariate geometry and calibration deformation carried larger explanatory weight, especially for pooled gradient boosting. In clinical terms, this means that a higher-risk target domain was not hard merely because more leaks occurred there. It was hard because the feature landscape and the meaning of the features changed. ICU admission, laboratory density, diversion, and even lactate functioned differently across domains. Some of these variables are obvious candidates for contextual reinterpretation. Others, such as albumin or duration, might appear more stable but still interacted with setting in practice. This is the statistical footprint of a clinically familiar reality: the same number can mean something different in a different service environment.

The transport-aware model improved performance by separating invariant physiology from contextual process [26]. That distinction is not perfect, and no representation can identify it completely from observational data alone. Even so, the empirical behavior of the model suggests the separation was useful. Albumin, white blood cell count, hemoglobin drop, and lactate remained important in the invariant subspace, while laboratory density, ICU destination, and night-start status were treated more contextually. A purely pooled model does not know that some variables should be interpreted partly as signs of observation intensity rather than as direct physiological surrogates. By assigning contextual weights through gating and calibration clusters, the domain-mixture model reduced the risk that high-monitoring environments would be misread as uniformly high-risk environments, or that low-observation environments would be read as reassuring simply because fewer abnormal values were measured.

The threshold analysis may be the most practically important part of the study. Prediction models are often reported as though the probability scale were self-explanatory. In clinical workflows, it is not. A risk estimate becomes meaningful only when linked to a decision threshold, and threshold behavior can vary far more across domains than summary discrimination suggests. Here a single global threshold created large, clinically awkward differences in positive predictive value and sensitivity. In one domain it produced many false alarms. In another it missed a substantial share of true leaks. This is not merely a calibration concern in the narrow technical sense. It translates directly into who undergoes additional imaging, who remains under higher observation, and how many low-yield alerts a team must tolerate before trust in the model decays. Domain-cluster calibration did not make the score universally uniform, but it substantially reduced arbitrary threshold distortion. That is a strong argument for viewing calibration as part of implementation design rather than as a post hoc reporting statistic.

The domains with the largest transport losses also tell a substantive clinical story. Low pelvic colorectal emergency surgery and upper gastrointestinal transfer-in complex care combined three features: high baseline event rates, wide physiologic dispersion, and care-process patterns that differed from the pooled average [27]. These are the settings where a simple pooled model is most likely to understate uncertainty. They are also the settings where clinical teams are least tolerant of false reassurance. In this sense, transport-aware modeling has ethical as well as statistical relevance. It does not eliminate risk or ambiguity, but it reduces the chance that the model's apparent confidence is borrowed from easier domains. That reduction matters even if the absolute gain in discrimination seems modest. In high-risk domains, calibration and uncertainty awareness often matter more than a few hundredths of area under the receiver operating characteristic curve.

The study also clarifies why full localization is not an adequate answer. Training separate models for each domain may seem aligned with the problem, but it is unstable in precisely the domains that need the most attention. Rare but important domains do not have enough support to sustain isolated modeling without substantial variance. Moreover, some signal genuinely is shared across procedures and settings. Physiologic derangement after an anastomosis is not wholly domain-specific. The challenge is not to abandon sharing but to share selectively. The domain-mixture model used common structure where it existed and contextual adaptation where the data suggested it was necessary. This seems better suited to the scale of the problem than either universal pooling or complete separation.

A related point concerns the definition of care setting. The study used episode-level settings such as enhanced recovery, urgent inpatient surgery, emergency overnight operation, and transfer-in complex care. These are not the only plausible settings, but they have two advantages. They are visible at the time prediction would be generated, and they map onto real workflow differences rather than purely administrative categories [28]. The choice matters because transport is ultimately about the circumstances of use. A model embedded in an overnight emergency pathway faces a different set of measurement and escalation dynamics than the same model embedded in a scheduled enhanced recovery pathway. Care setting, in other words, is not merely a confounder to be adjusted away. It is part of the object being predicted.

The results further suggest that cross-domain evaluation should become a routine reporting element for leak prediction studies. Aggregate performance can still be useful, especially for broad benchmarking, but it is insufficient for decisions about clinical use. A model might appear competitive overall and yet be poorly calibrated where its decisions would be most consequential. Reporting performance by procedure family alone would not solve the prob-

lem, because setting within family still mattered substantially. Reporting by hospital alone would not solve it either, because much of the shift was episode-level rather than institution-level. What seems necessary is a domain scheme that crosses clinically meaningful procedure categories with care contexts that plausibly alter observation, physiology, or intervention thresholds.

This paper does not imply that every hospital must build a sophisticated domain-mixture model before any form of leak surveillance can be attempted. A more modest implication is that deployment strategies should be matched to the degree of domain heterogeneity. In low-shift domains such as elective colonic surgery, a pooled model with careful local recalibration may be sufficient. In high-shift domains, especially those involving technically demanding procedures or operationally unusual care pathways, a pooled model should not be trusted without explicit transport testing and domain-aware calibration. That recommendation is less demanding than universal bespoke modeling and more realistic than universal score portability.

The study also contributes to a broader discussion in clinical machine learning. Many medical prediction tasks are now known to suffer from distribution shift across institutions. The present results show that meaningful shift can occur within the same technical network even when hospital systems share a common data architecture [29]. The clinically salient distribution shift is not always the hospital boundary. It may instead lie at the intersection of procedure, care pathway, urgency, and monitoring practice. Focusing only on institution-to-institution drift can therefore miss a major part of the portability problem. For postoperative anastomotic leak, the domain appears to be a more natural unit of transport than the hospital alone.

7. Limitations

Several limitations should guide interpretation. The first concerns observational design. Because the study used retrospective structured data, it cannot determine why a predictor changes meaning across domains. Some domain effects likely reflect biological differences in case mix and technical complexity. Others likely reflect care-process differences such as monitoring density, ICU triage rules, postoperative order sets, or thresholds for imaging and drainage. The model can adapt to these shifts empirically, but it cannot identify their causal proportions. The finding that transport-aware learning improves prediction does not establish which underlying mechanisms are responsible.

Second, the care-setting taxonomy, although clinically motivated, is still a simplification. Enhanced recovery, urgent inpatient care, emergency overnight surgery, and transfer-in complex care are broad categories. Important differences remain within each one. Two emergency overnight domains can differ in staffing, contamination source, or time to source control. Two transfer-in complex care episodes can differ in whether the transfer reflects specialized oncology care, redo surgery, or salvage after prior complications. The domains used here are therefore pragmatic rather than exhaustive. They were selected because they are observable, common enough for analysis, and relevant at the point of intended prediction [30]. Different taxonomies may reveal different structures of transport loss.

Third, the outcome definition emphasized clinically managed leak rather than every possible biological or radiographic failure. That choice aligns with the practical goal of postoperative escalation, but it may miss some lower-grade or conservatively managed leaks. It may also be affected by domain differences in diagnostic intensity and treatment thresholds. A domain with lower access to imaging or drainage could appear to have fewer clinically managed leaks

even if biological failure occurs at a similar rate. The study attempted to reduce this problem through adjudication and readmission linkage, yet residual differential misclassification remains possible.

Fourth, only variables available by 18 hours after surgery were included. This made the predictive task operationally coherent and kept the transport problem focused, but it excluded later postoperative physiology and narrative documentation that might narrow uncertainty in difficult domains. The choice should not be interpreted as evidence that richer data would be unnecessary. On the contrary, some of the remaining transport errors in upper gastrointestinal and low pelvic emergency domains may be reducible only through time-updated modeling or inclusion of unstructured text, imaging metadata, or operative narrative details. The present study examined how far one can go with a restrained early-postoperative structured feature set, not the maximal achievable performance.

Fifth, the domain-mixture model is more complex than the benchmark models. Although its components are interpretable at a high level, its full internal behavior is not as transparent as elastic-net logistic regression. Complexity alone is not a flaw, but it does create practical considerations for deployment, monitoring, and governance. A hospital may accept modestly worse transport performance in exchange for greater simplicity. The paper does not argue that the proposed model is the only acceptable form of domain-aware learning. It argues that some explicit transport strategy is necessary when heterogeneity is strong.

Sixth, the evaluation, while more rigorous than random splitting, still took place within a connected hospital network [31]. Data schemas, coding conventions, and some elements of practice were shared across sites. Transport into entirely separate health systems, especially those with different perioperative documentation or rescue infrastructure, may be more difficult. The observed gains from the domain-mixture model should therefore be interpreted as improvement under meaningful but not maximal shift. External validation in independent systems would remain necessary before broad adoption.

Finally, the study did not incorporate surgeon-level identifiers or fine-grained institutional pathway adherence. Those variables are often incomplete or inconsistently coded, but they may explain part of the domain-sensitive predictor meaning observed here. A prolonged operation in one surgeon's practice may not imply the same leak risk as in another's, and a transfer-in complex care episode in a center with rapid endoscopic rescue may differ from one without it. Some of this structure is indirectly absorbed by the domains and the observed features, yet more granular process data could change both the estimated distance between domains and the optimal transport strategy.

8. Conclusion

This study approached heterogeneity of anastomotic leak risk across surgery types and care settings from a different angle: not only as variation in event rates, but as variation in the validity of prediction models when they move across domains. In a multicenter cohort spanning twenty-five clinically defined surgery-setting domains, pooled models that appeared strong under conventional internal validation lost substantial accuracy and calibration when tested on held-out target domains. The losses were concentrated in technically and organizationally complex settings, especially low pelvic colorectal emergency surgery and upper gastrointestinal transfer-in complex care.

A transport-aware domain-mixture model preserved more discrimination and considerably better calibration than both naive pooling and full localization. The improvement arose less

from correcting prevalence alone than from acknowledging that some features carry shared physiological meaning while others are partly context-dependent. Threshold behavior also became more stable when calibration was performed at the level of latent domain clusters rather than with a single global score. A leak model should not be judged only by how well it fits a pooled postoperative cohort. It should also be judged by how far it can travel before its risk scale changes meaning. For this problem, transportability is not a peripheral robustness statistic. It is part of the clinical usefulness of the model itself [32].

References

[1] Hui Jiang Gao, Ju Wei Mu, Wei Min Pan, Malcolm V. Brock, Mao Long Wang, Bin Han, and Kai Ma. "Totally mechanical linear stapled anastomosis for minimally invasive Ivor Lewis esophagectomy: Operative technique and short-term outcomes." In: *Thoracic cancer* 11.3 (Feb. 3, 2020), pp. 769–776. DOI: 10.1111/1759-7714.13339.

[2] Alexandre Doussot et al. "Surgical Management and Outcomes of Rectal Cancer with Synchronous Prostate Cancer: A Multicenter Experience from the GRECCAR Group." In: *Annals of surgical oncology* 27.11 (June 4, 2020), pp. 4286–4293. DOI: 10.1245/s10434-020-08683-4.

[3] 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.

[4] Tina Bharani, Eric G Sheu, and Malcolm K Robinson. "Procedure Matters in Gender-Associated Outcomes following Metabolic-Bariatric Surgery: Five Year North American Matched Cohort Analysis." In: *Obesity surgery* 33.10 (July 15, 2023), pp. 3090–3096. DOI: 10.1007/s11695-023-06722-z.

[5] Koshi Kumagai et al. "A systematic review and meta-analysis comparing partial stomach partitioning gastrojejunostomy versus conventional gastrojejunostomy for malignant gastroduodenal obstruction". In: *Langenbeck's archives of surgery* 401.6 (June 23, 2016), pp. 777–785. DOI: 10.1007/s00423-016-1470-8.

[6] Binay Thakur, Mukti Devkota, Nikesh Bhandari, Shashank Shrestha, and Ashish Kharel. "Minimally invasive esophagectomy/ gastroesophagectomy for cancer - Long term results from a single institution". In: *Nepalese Journal of Cancer* 6.2 (Oct. 6, 2022), pp. 7–15. DOI: 10.3126/njc.v6i2.48754.

[7] Kostas Toutouzas, Eleftheria S Kleidi, Panagiotis G. Drimousis, Margarita Balla, Metaxia N Papanikolaou, Andreas Larentzakis, Dimitrios Theodorou, and Stylianos Katsaragakis. "Anastomotic leak management after a low anterior resection leading to recurrent abdominal compartment syndrome: a case report and review of the literature." In: *Journal of medical case reports* 3.1 (Nov. 14, 2009), pp. 125–125. DOI: 10.1186/1752-1947-3-125.

[8] Frances J. Puleo, Nitin Mishra, and Jason F. Hall. "Use of Intra-Abdominal Drains". In: *Clinics in colon and rectal surgery* 26.3 (Aug. 19, 2013), pp. 174–177. DOI: 10.1055/s-0033-1351134.

[9] Hirotoshi Kikuchi, Hideki Endo, Hiroyuki Yamamoto, Soji Ozawa, Hiroaki Miyata, Yoshihiro Kakeji, Hisahiro Matsubara, Yuichiro Doki, Yuko Kitagawa, and Hiroya Takeuchi. "Impact of Reconstruction Route on Postoperative Morbidity After Esophagectomy: Analysis of Esophagectomies in the Japanese National Clinical Database." In: *Annals of gastroenterological surgery* 6.1 (Sept. 6, 2021), pp. 46–53. DOI: 10.1002/ags3.12501.

[10] Amir Aryaie, Jordan L. Singer, Mojtaba Fayezizadeh, Jon Lash, and Jeffrey M. Marks. "Efficacy of endoscopic management of leak after foregut surgery with endoscopic cov-

ered self-expanding metal stents (SEMS)". In: *Surgical endoscopy* 31.2 (June 17, 2016), pp. 612–617. DOI: 10.1007/s00464-016-5005-8.

[11] Martin Pölcher, Meike Eckhardt, Christoph Coch, Matthias Wolfgarten, Kirsten Kübler, Gunther Hartmann, Walther Kuhn, and Christian Rudlowski. "Sorafenib in combination with carboplatin and paclitaxel as neoadjuvant chemotherapy in patients with advanced ovarian cancer." In: *Cancer chemotherapy and pharmacology* 66.1 (Mar. 5, 2010), pp. 203–207. DOI: 10.1007/s00280-010-1276-2.

[12] L Lobbes, S Berns, K Beyer, and B Weixler. "Software-based perfusion assessment of the ileal J-pouch – a standardized and objective approach of interpreting intraoperative indocyanine green near-infrared fluorescence". In: *British Journal of Surgery* 109.Supplement$_3$ (May 31, 2022). DOI: 10.1093/bjs/znac181.011.

[13] Raquel Ortigão, Brigitte Pereira, Rui Silva, Pedro Pimentel-Nunes, Pedro Bastos, Joaquim Abreu de Sousa, Filomena Faria, Mário Dinis-Ribeiro, and Diogo Libânio. "Anastomotic Leaks following Esophagectomy for Esophageal and Gastroesophageal Junction Cancer: The Key Is the Multidisciplinary Management." In: *GE Portuguese journal of gastroenterology* 30.1 (Dec. 14, 2021), pp. 38–48. DOI: 10.1159/000520562.

[14] Subhashini Ayloo, Young Hoon Roh, and Nabajit Choudhury. "Laparoscopic, hybrid, and totally robotic Roux-en-Y gastric bypass". In: *Journal of robotic surgery* 10.1 (Jan. 25, 2016), pp. 41–47. DOI: 10.1007/s11701-016-0559-y.

[15] Waresijiang Yibulayin, Sikandaer Abulizi, Hongbo Lv, and Wei Sun. "Minimally invasive oesophagectomy versus open esophagectomy for resectable esophageal cancer: a meta-analysis". In: *World journal of surgical oncology* 14.1 (Dec. 8, 2016), pp. 304–304. DOI: 10.1186/s12957-016-1062-7.

[16] Naseem Akhtar. "IDDF2018-ABS-0242 Sleeve omentopexy over pancreatico jejunostomy – a new technique". In: *Clinical Gastroenterology* 67 (June 8, 2018), A78.2–A78. DOI: 10.1136/gutjnl-2018-iddfabstracts.170.

[17] Rachel B. Scott, Lane A. Ritter, Amber L. Shada, Sanford H. Feldman, and Daniel E. Kleiner. "Endoluminal Vacuum Therapy for Ivor Lewis Anastomotic Leaks: A Pilot Study in a Swine Model". In: *Clinical and translational science* 10.1 (Nov. 11, 2016), pp. 35–41. DOI: 10.1111/cts.12427.

[18] Yukinori Kurokawa, Hitoshi Katai, Haruhiko Fukuda, and Mitsuru Sasako. "Phase II Study of Laparoscopy-assisted Distal Gastrectomy with Nodal Dissection for Clinical Stage I Gastric Cancer: Japan Clinical Oncology Group Study JCOG0703". In: *Japanese journal of clinical oncology* 38.7 (June 26, 2008), pp. 501–503. DOI: 10.1093/jjco/hyn055.

[19] Malcolm Steel, Peter Danne, and Ian T. Jones. "Colon trauma: Royal Melbourne Hospital experience". In: *ANZ journal of surgery* 72.5 (May 27, 2002), pp. 357–359. DOI: 10.1046/j.1445-2197.2002.02408.x.

[20] Valerie P Bauer. "The Evidence against Prophylactic Nasogastric Intubation and Oral Restriction". In: *Clinics in colon and rectal surgery* 26.3 (Aug. 19, 2013), pp. 182–185. DOI: 10.1055/s-0033-1351136.

[21] Thomas A. Bowles and David A. K. Watters. "TIME TO CUSUM: SIMPLIFIED REPORTING OF OUTCOMES IN COLORECTAL SURGERY". In: *ANZ journal of surgery* 77.7 (June 28, 2007), pp. 587–591. DOI: 10.1111/j.1445-2197.2007.04156.x.

[22] J. Ragg, G. D. Guest, M. Thorne, David A. K. Watters, James Hurley, and S. Crowley. "RS06 INTRODUCTION OF LAPAROSCOPIC RESECTIONAL COLORECTAL SURGERY TO NAIVE HOSPITALS". In: *ANZ Journal of Surgery* 77.s1 (Apr. 24, 2007). DOI: 10.1111/j.1445-2197.2007.04128_6.x.

[23] K Arunsenthilnathan. *A Comparative study of Efficacy of Staplers Versus Hand Sewn Anastomosis in Bowel Surgeries*. May 1, 2019.

[24] Uberto Fumagalli, Alessandra Melis, Jana Balazova, Valeria Lascari, Emanuela Morenghi, and Riccardo Rosati. "Intra-operative hypotensive episodes may be associated with post-operative esophageal anastomotic leak". In: *Updates in surgery* 68.2 (May 5, 2016), pp. 185–190. DOI: 10.1007/s13304-016-0369-9.

[25] Nina Nederlof, Hugo W. Tilanus, Tahnee de Vringer, J. Jan B. van Lanschot, Sten P. Willemsen, Wim C. J. Hop, and Bas P. L. Wijnhoven. "A single blinded randomized controlled trial comparing semi-mechanical with hand-sewn cervical anastomosis after esophagectomy for cancer (SHARE-study)." In: *Journal of surgical oncology* 122.8 (Sept. 28, 2020), pp. 1616–1623. DOI: 10.1002/jso.26209.

[26] Gabriele Anania, R. J. Davies, Francesco Bagolini, N. Vettoretto, Justus J. Randolph, Roberto Cirocchi, and Annibale Donini. "Right hemicolectomy with complete mesocolic excision is safe, leads to an increased lymph node yield and to increased survival: results of a systematic review and meta-analysis". In: *Techniques in coloproctology* 25.10 (June 12, 2021), pp. 1099–1113. DOI: 10.1007/s10151-021-02471-2.

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

[28] Haytham Abudeeb, A.E. Hammad, Ajogwu Ugwu, Jamshid Darabnia, Lee Malcomson, Min Maung, Khurram Khan, Clare Mclaughlin, and Arijit Mukherjee. "Defunctioning stoma- a prognosticator for leaks in low rectal restorative cancer resection: A retrospective analysis of stoma database." In: *Annals of medicine and surgery (2012)* 21 (July 19, 2017), pp. 114–117. DOI: 10.1016/j.amsu.2017.07.044.

[29] Luis F. Tapias, Ashok Muniappan, Cameron D. Wright, Henning A. Gaissert, John C. Wain, Christopher R. Morse, Dean M. Donahue, Douglas J. Mathisen, and Michael Lanuti. "Short and Long-Term Outcomes After Esophagectomy for Cancer in Elderly Patients". In: *The Annals of thoracic surgery* 95.5 (Mar. 7, 2013), pp. 1741–1748. DOI: 10.1016/j.athoracsur.2013.01.084.

[30] Ying Li, Ji-Jun Deng, and Jun Jiang. "Relationship between body mass index and short-term postoperative prognosis in patients undergoing colorectal cancer surgery." In: *World journal of clinical cases* 11.12 (Apr. 26, 2023), pp. 2766–2779. DOI: 10.12998/wjcc.v11.i12.2766.

[31] Leonardo Solaini, Antonio Bocchino, Andrea Avanzolini, Domenico Annunziata, Davide Cavaliere, and Giorgio Ercolani. "Robotic versus laparoscopic left colectomy: a systematic review and meta-analysis." In: *International journal of colorectal disease* 37.7 (June 1, 2022), pp. 1497–1507. DOI: 10.1007/s00384-022-04194-8.

[32] Binay Thakur, Chun Shan Zhang, Yang Guo, and Robin Lama. "Comparison Between Concurrent Chemoradiation Followed by Surgery vs. Surgery for Locally Advanced Cancer of Esophagus." In: *The Indian journal of surgery* 73.2 (Nov. 19, 2010), pp. 111–115. DOI: 10.1007/s12262-010-0182-5.