Fairness-Constrained Risk Modeling for Anastomotic Leak Prediction Under Distribution Shift, Label Noise, and Heterogeneous Clinical Workflows

Karim Haddad[1], **Rami Khoury**[2], **and Tarek Nassar**[3]

[1]Department of Computer and Communications Engineering, Lebanese German University, Rue Fouad Chehab, Sahel Alma, Jounieh, Mount Lebanon, Lebanon

[2]Faculty of Computer Science and Information Technology, Arts, Sciences and Technology University in Lebanon (AUL), Cola Intersection, Verdun Street, Beirut, Lebanon

[3]Department of Informatics Engineering, Islamic University of Lebanon, Khaldeh Main Road, Al Wardaniyeh District, Khaldeh, Mount Lebanon, Lebanon

ABSTRACT Anastomotic leak prediction has become an important target for clinical machine learning because the event is uncommon, severe, and strongly time-sensitive. The practical question is not only whether a model can separate patients who will later develop a leak from those who will not, but whether that separation is equitable across patient populations, hospitals, and care pathways. Fairness in this setting is difficult because outcome prevalence is low, labeling is imperfect, surveillance intensity varies, and post-operative management changes the very event that the model is asked to predict. These features make standard fairness language too narrow when used without attention to clinical utility and data generation. This paper develops a framework for fairness in anastomotic leak prediction that integrates statistical risk estimation, calibration, subgroup robustness, causal reasoning, and deployment governance. The discussion treats fairness as a property of the full decision pipeline rather than of a score alone. It examines how imbalance, missingness, treatment selection, procedure mix, and site-specific workflows can create systematic disparities even when sensitive attributes are excluded from the feature set. It then formulates fairness-aware optimization strategies that combine worst-group risk control, calibration preservation, and utility-sensitive threshold design. A complementary audit protocol is described for temporal validation, external transport, intersectional subgroup analysis, and prospective surveillance after deployment. The resulting perspective does not assume a single fairness criterion is sufficient. Instead, it argues that equitable anastomotic leak prediction requires a layered design in which model development, evaluation, and clinical integration are jointly constrained by the need to avoid uneven error burden and uneven access to beneficial intervention.

KEYWORDS

1. Introduction

Anastomotic leak, hereafter AL, occupies a difficult position in surgical prediction because it is clinically consequential, statistically sparse, and operationally entangled with decisions made before, during, and after the index procedure [1]. A risk model for AL is therefore not simply a classifier applied to a static dataset. It is a component of an adaptive care process

Article info:
Karim Haddad, Rami Khoury, and Tarek Nassar. *Fairness-Constrained Risk Modeling for Anastomotic Leak Prediction Under Distribution Shift, Label Noise, and Heterogeneous Clinical Workflows.* Grovesocieties .

in which surveillance intensity, rescue intervention, discharge timing, reoperation thresholds, and documentation practices all shape the target variable. The same prediction system may perform acceptably at the level of aggregate discrimination and yet behave inequitably across sex, age, race, insurance category, operative urgency, anatomical site, or hospital context. In low-incidence surgical events, such inequity is easy to miss because pooled performance metrics are numerically stable long before subgroup performance estimates become reliable. Fairness in AL prediction is thus not a rhetorical extension of generic machine learning ethics. It is a technical requirement for any decision tool that may differentially alter access to attention, imaging, monitoring, or rescue care.

A further complication is that AL prediction sits at the boundary between prognostic modeling and intervention policy. A model may be used preoperatively to identify candidates for intensified optimization, intraoperatively to support protective diversion decisions, or postoperatively to modulate monitoring density, imaging thresholds, and escalation pathways. Each use case changes the relevant notion of fairness. A preoperative model touches allocation of preventive resources and informed consent. An intraoperative model influences surgical strategy in real time, where false positive and false negative errors carry asymmetric procedural costs. A postoperative surveillance model shapes who receives additional testing and when clinicians respond. These are not interchangeable settings. A fairness definition that is coherent for one role may be incoherent for another because the action set, benefit profile, and background prevalence are different.

The difficulty is magnified by the structure of surgical data. Electronic health records are neither passive measurements nor simple reflections of physiology. They are traces of interactions between patients, clinicians, workflows, institutional conventions, and billing systems. Laboratory density may vary by unit acuity [2]. Imaging reports may depend on whether a patient is deemed concerning enough to scan. Administrative labels may encode treated complications more readily than occult events. Protective measures such as diversion, antibiotics, or intensive monitoring can reduce observed leak rates in groups flagged as high risk, thereby hiding latent vulnerability while simultaneously making a model appear less fair or more fair depending on the metric chosen. A score that equalizes one error rate after such intervention may still yield unequal clinical benefit.

The emerging AL prediction literature has largely focused on discrimination, calibration, and feasibility, with fairness receiving little direct treatment. An open EHR-based proof-of-concept study underscored the issue by showing that AL models can achieve meaningful predictive performance while subgroup bias analysis remains unperformed and equitable validation remains a necessary next step, which places fairness at the center of future methodological work rather than at its margin [3]. That observation is notable because it reframes the problem. The challenge is no longer whether structured data can support AL risk estimation at all. The challenge is whether such estimation can be made reliable and equitable across heterogeneous populations without obscuring clinically relevant signal or destabilizing care.

A useful fairness framework for AL prediction must begin by rejecting two unhelpful simplifications. The first is the assumption that fairness can be achieved by deleting sensitive variables from the design matrix. Excluding explicit group identifiers does not remove proxy structure. Procedure type, anatomy, insurance status, preoperative optimization opportunities, ICU admission, laboratory measurement frequency, and operative timing may encode social and institutional differences that continue to shape predictions. The second simplification is the belief that a single metric, typically demographic parity or equalized odds, can summarize fairness in a clinically acceptable way. Because AL prevalence differs across meaningful subpopulations and because intervention costs are asymmetric, forcing identical score distributions or identical threshold behavior can destroy calibration, reduce utility, or conceal disparities in downstream benefit. Fairness in this domain must therefore be multi-criteria and clinically conditioned.

The conceptual move adopted here is to treat fairness as an alignment problem among four objects: the latent clinical state that determines vulnerability to AL, the observational process that generates features and labels, the statistical model that maps observations to risk, and the decision policy that converts risk into action. If inequity can arise at each layer, then it must be evaluated at each layer. The latent clinical state may differ across groups because of anatomy, disease severity, comorbidity burden, or prior access to care [4]. The observational layer may differ because some patients are monitored more densely or coded more completely than others. The modeling layer may differ because standard empirical risk minimization favors majority environments and can underfit minority substructures. The policy layer may differ because a common threshold can translate into unequal clinical opportunity when the same score corresponds to different expected benefit across groups.

The remainder of this paper develops a technical account of fairness in AL prediction around that layered view. It first formalizes the prediction task, the role of sensitive and non-sensitive group structure, and the incompatibilities among standard fairness criteria in rare-event surgery. It then examines how inequity enters through selection, measurement, labeling, and treatment feedback. After that, it considers representation learning and causal structure under hospital and temporal shift, since fairness that does not transport is not fairness in practice. A constrained optimization framework is then described for balancing discrimination, calibration, worst-group robustness, and clinical utility. Finally, an evaluation and governance design is outlined for auditing performance before and after deployment. The goal is not to argue that fairness can be reduced to a single formula. The goal is to show that equitable AL prediction requires a mathematically explicit framework in which what is optimized, what is measured, and what is acted upon are all jointly specified.

2. Problem Formulation and Fairness Notions

Let each patient encounter be indexed by $i \in \{, \dots, n\}$. For each encounter, suppose there exists a feature vector $X_i \in \mathbb{R}^p$ comprising demographic variables, comorbidity indicators, lab-

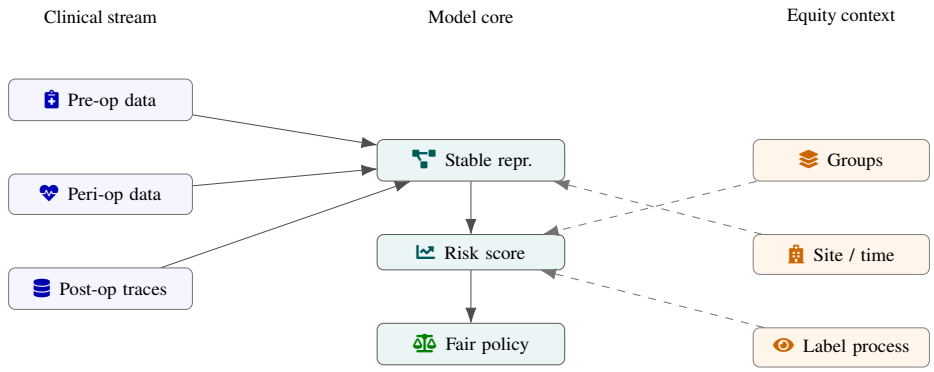

Figure 1 Layered fairness-aware AL prediction pipeline. Clinical observations are mapped into a stable representation and risk score, while group structure, hospital-time environment, and label generation remain explicit because equity depends on the full decision pathway rather than on a score alone.

Table 1 Summary of Fairness Notions in AL Prediction

Concept	Definition	Clinical Relevance	Limitation
Calibration	$\Pr(Y = \mid R = r, A = g) = r$	Reliable risk interpretation	Ignores error disparity
Equalized Odds	Equal TPR/FPR across groups	Balanced detection burden	Conflicts with calibration
Utility Parity	$U_g(\tau)$ comparable	Aligns benefit across groups	Requires utility specification
Ranking Fairness	Within-group ordering accuracy	Triage prioritization	Threshold not addressed

oratory values, operative descriptors, postoperative physiologic measurements, and workflow-derived indicators. Let A_i denote a vector of sensitive or equity-relevant attributes, not restricted to legally protected categories alone but including variables such as sex, race, age band, insurance class, hospital, care setting, and intersectional group membership when appropriate. Let $Y_i \in \{\varkappa, \}$ represent the latent binary outcome of clinically meaningful AL within a defined horizon, and let $\tilde{Y}_i$ denote the observed label available for model fitting. The distinction between Y_i and $\tilde{Y}_i$ is essential, because observed labels in surgical datasets frequently mix diagnosis codes, reintervention patterns, antibiotic exposure, imaging findings, and chart abstractions, none of which perfectly recover the underlying event.

A prediction model outputs a score $R_i = f_\theta(X_i)$, with $R_i \in [\varkappa,]$ interpreted as estimated risk under a specified target estimand. Even this interpretation requires care. One possible estimand is the probability of observed AL under prevailing care, another is the probability of latent AL under no additional intervention beyond standard practice, and another is the probability of clinically consequential leak requiring rescue treatment. Fairness analysis depends on which estimand is chosen. A model trained on administrative labels may equalize errors with respect to coding practice rather than with respect to pathophysiology. A model trained on rescue-treated events may learn clinician behavior as much as patient risk [5]. Therefore fairness must be tied to the clinical meaning of Y_i rather than to the convenience of available labels.

A threshold policy maps risk to action. Let $D_i(\tau) = \mathbb{I}\{R_i \geq \tau\}$ denote an intervention trigger such as enhanced monitoring, early imaging, senior review, intensive care retention, or protective surgical strategy. In many discussions fairness is assessed only at the score level, but in AL prediction the decision layer matters at least as much as the score layer because the harm of an error is mediated by the action it induces. False positives consume scarce resources, expose patients to unnecessary testing, and can alter surgical management in ways that carry morbidity. False negatives delay recognition of a complication whose prognosis worsens with time. A fairness criterion that ignores these downstream asymmetries is incomplete.

A utility-based expression makes this explicit. Let B_i denote the clinical benefit of correct intervention for patient i, let C_i denote the burden of unnecessary intervention, and let M_i denote the missed-opportunity cost of failing to intervene before deterioration. Then the expected policy utility under threshold

Table 2 Sources of Inequity in Clinical Data

Source	Mechanism	Example	Impact
Selection Bias	Cohort construction differences	Tertiary vs community centers	Skewed prevalence
Measurement Bias	Unequal feature observation	ICU vs ward labs	Feature imbalance
Label Noise	Imperfect outcome capture	Coding variability	Misleading fairness
Surveillance Bias	Detection depends on monitoring	Imaging frequency	Inflated outcomes
Treatment Effects	Intervention alters outcome	Early antibiotics	Distorted labels

Table 3 Fairness-Aware Optimization Components

Component	Objective Term	Purpose	Trade-off
Predictive Loss	L_{pred}	Overall accuracy	Ignores minorities
Worst-Group Loss	L_{wg}	Protect subgroups	May reduce average
Calibration Penalty	L_{cal}	Reliable probabilities	Hard in rare events
Disparity Penalty	L_{disp}	Reduce inequity	Metric-dependent

τ may be written as

$$U(\tau) = \mathbb{E}\Big[B_i D_i(\tau) Y_i - C_i D_i(\tau)(-Y_i) - M_i(-D_i(\tau)) Y_i\Big]. \quad (1)$$

This expression is deliberately generic. It accommodates settings in which intervention is mild and monitoring-oriented, making C_i relatively small, and settings in which intervention is invasive or resource intensive, making C_i larger. Fairness then concerns not only whether the score predicts equally well across groups, but whether the induced utility is distributed in a clinically acceptable way across those groups.

Standard statistical fairness notions can still be useful, but only when interpreted carefully. Calibration within groups requires that for every group g and score value r, the conditional event frequency equals the predicted probability:

$$\Pr(Y = \mid R = r, A = g) = r. \quad (2)$$

If this condition holds, a risk estimate of $\varkappa.\mathbb{C}$ has the same frequency interpretation within each audited group. This property matters in surgery because clinicians often consume model output as a risk estimate rather than as an ordinal rank. Yet calibration alone does not constrain error-rate disparities. Two groups can both be calibrated while receiving different false negative or false positive burdens if base rates differ.

Equalized odds attempts to control such burdens by requiring equal true positive and false positive rates across groups at a chosen threshold [6]. For groups g and g', define

$$\Delta_{\mathrm{TPR}}(g, g') = \Pr(D = \mid Y = , A = g) - \Pr(D = \mid Y = , A = g'), \quad (3)$$

$$\Delta_{\mathrm{FPR}}(g, g') = \Pr(D = \mid Y = \varkappa, A = g) - \Pr(D = \mid Y = \varkappa, A = g'). \quad (4)$$

The closer these gaps are to zero, the closer the policy is to equal opportunity and equalized burden at the thresholded decision level. However, when prevalence differs across groups, simultaneous calibration and equalized odds are generally incompatible unless prediction is perfect. In AL prediction, prevalence can differ for clinically meaningful reasons related to anatomy, disease extent, urgency, or perioperative instability. If such differences are genuine, forcing equalized odds may require miscalibration or active distortion of risk. If such differences are artifacts of surveillance or treatment, preserving raw calibration may entrench inequity. This tension is one reason fairness in AL prediction must be contextualized rather than metric-maximalist.

A more clinically coherent alternative is utility parity or net-benefit parity across groups. Let $U_g(\tau)$ denote expected

Table 4 Evaluation Metrics by Group

Metric	Definition	Use Case	Caveat
AUC_g	Ranking accuracy	Discrimination	Ignores calibration
ICE_g	Calibration error	Risk reliability	Sensitive to bins
TPR/FPR	Threshold metrics	Decision fairness	Threshold dependent
Net Benefit	Utility-based score	Clinical value	Needs cost ratio

Table 5 Representation Learning Objectives

Term	Description	Goal	Risk
Invariant Loss	Reduce env. dependence	Transportability	Over-smoothing
Fairness Regularizer	Penalize disparities	Equity	Signal removal
Clinical Signal Z^{clin}	Stable features	True risk capture	Hard to isolate
Nuisance Z^{nuis}	Workflow artifacts	Remove bias	Leakage possible

utility restricted to group g. A fairness audit can then examine

$$\Delta_U(g, g') = U_g(\tau) - U_{g'}(\tau), \tag{5}$$

possibly after standardizing for clinically relevant case mix. This criterion does not replace calibration or error-rate analysis. It complements them by asking whether the decision policy produces systematically lower expected benefit for one group even when aggregate performance appears acceptable. In surgical monitoring, such a disparity may be more consequential than a modest difference in raw false positive rate.

Another useful distinction is between fairness in ranking and fairness in thresholding. Many AL models are first used to rank patients for review rather than to trigger fully automated action. In that case a threshold-independent measure such as within-group concordance or pairwise ranking loss may be relevant. Define the groupwise ranking risk

$$L_g^{\text{rank}}(\theta) = \Pr\big(R_i \le R_j \mid Y_i = , \\ Y_j = \varkappa, A_i = A_j = g\big). \tag{6}$$

A model can achieve similar global area under the curve while showing poor ranking behavior in minority groups due to small sample size, noisier measurements, or feature miscalibration [7]. Worst-group ranking performance therefore matters even when the deployment policy is not purely threshold-based.

Intersectionality introduces an additional layer. AL risk and its observation are shaped not by a single axis but by combinations such as age with frailty, sex with pelvic anatomy, insurance with hospital transfer patterns, or race with resource constraints. Let $\mathcal{G}$ denote a collection of intersectional groups. Fairness is then not adequately described by marginal parity across one attribute at a time. One must consider the worst-case subgroup risk

$$R_{\max}(\theta) = \max_{g \in \mathcal{G}} R_g(\theta), \tag{7}$$

where $R_g(\theta)$ may refer to cross-entropy, Brier risk, missed-event cost, or another clinically chosen loss. The challenge is statistical as much as ethical, because fine-grained subgroups rapidly become small. This means fairness analysis in AL prediction is inseparable from hierarchical estimation and uncertainty quantification.

A final issue is that the sensitive attribute set A should not be treated as static and purely social. Some variables, such as hospital identity or postoperative care unit, are not protected attributes but are essential equity-relevant environments because they mediate how information is measured and how action is delivered. A model that is fair across demographic groups but unstable across hospitals with different resource levels may still generate inequitable care if deployment is concentrated in under-resourced settings. Thus the fairness problem should be framed over a broader partition of the data-generating process, one that includes patient groups, procedural strata, and institutional environments.

In summary, the formal prediction problem for AL involves a latent event, imperfect labels, heterogeneous environments, asymmetric action costs, and multiple group structures. Fairness

Table 6 Threshold Policy Outcomes

Outcome	Expression	Interpretation	Concern
Exposure	$\Pr(D = \mid A = g)$	Intervention rate	Resource burden
False Negative	$\Pr(D = \varkappa \mid Y =)$	Missed cases	Safety risk
False Positive	$\Pr(D = \mid Y = \varkappa)$	Over-treatment	Cost increase
Utility	$U_g(\tau)$	Net benefit	Requires modeling

Table 7 Distribution Shift Considerations

Factor	Description	Example	Effect
Temporal Drift	Practice evolution	ERAS adoption	Calibration shift
Site Variation	Hospital workflow	ICU usage	Feature meaning changes
Prevalence Shift	Different case mix	Urgent surgery rates	Metric instability
Measurement Shift	Data density changes	Lab frequency	Missingness bias

therefore cannot be reduced to removing sensitive features or matching one descriptive statistic. A coherent framework must track calibration, ranking quality, thresholded error burden, and policy utility, all conditional on groups and environments that matter clinically. The next step is to examine how inequity enters the pipeline before any optimization is attempted.

3. Sources of Inequity in AL Prediction

The primary sources of inequity in AL prediction are not confined to the model architecture. They arise upstream through who enters the dataset, what gets measured, how the outcome is encoded, and which interventions are applied before the label is finalized [8]. Because AL is infrequent, these distortions can be consequential even when they affect only a modest fraction of encounters. Small systematic differences in surveillance or coding can produce large proportional differences in observed event rates for some groups, thereby altering both the target and the apparent fairness profile.

Selection bias begins with the surgical cohort itself. Patients who undergo anastomosis are already filtered by access to care, preoperative workup, referral patterns, disease stage, hospital capability, and clinician judgment. A model trained in a tertiary center may be dominated by transferred, high-acuity, or oncologically complex cases. A model trained in a community setting may reflect a different operative mix and different rescue thresholds. If one group is overrepresented in centers with aggressive monitoring and another in centers with limited postoperative surveillance, the resulting dataset does not simply reflect different patients; it reflects different opportunities for leak detection and intervention. Fairness assessed only within a single dataset may therefore miss inequity induced by how that dataset was assembled.

Measurement bias is equally important. Postoperative laboratory values, vital signs, drain output documentation, radiology reports, and antibiotic exposure are not uniformly observed. More unstable patients are measured more often. Patients kept in the intensive care unit generate denser physiologic traces than those sent to the ward. Some hospitals obtain inflammatory markers routinely; others do not. Some surgeons use drains frequently; others do not, altering both early signals and the visibility of leaks. Missingness in such settings is not a nuisance variable but an informative process. A missing postoperative lactate is often evidence that a patient was not considered ill enough to warrant measurement. If measurement intensity differs by group because of acuity, resource allocation, or workflow conventions, then a model trained on observed covariates can inherit these inequities even when imputation is handled carefully.

The missingness mechanism can be expressed formally. Let $M_{ij} \in \{\varkappa, \}$ indicate whether feature j is observed for patient i. Then

$$\Pr(M_{ij} = \mid X_i^*, Y_i, A_i, H_i) \tag{8}$$

may depend on latent physiology X_i^*, the eventual outcome Y_i, group membership A_i, and hospital environment H_i. When this dependence includes A_i even after conditioning on clinically relevant severity, differential feature visibility arises [9]. Standard median imputation or complete-case analysis can then

Table 8 Governance and Deployment Elements

Element	Role	Example	Risk if Missing
Threshold Design	Action mapping	Imaging trigger	Unequal care
Override Logging	Track decisions	Alert dismissal	Hidden bias
Capacity Planning	Resource limits	ICU beds	Inequitable access
Monitoring	Drift detection	Sequential stats	Silent failure

Table 9 Intersectional Fairness Handling

Method	Description	Benefit	Limitation
Hierarchical Models	Partial pooling	Stable estimates	Assumption sensitive
Worst-Case Risk	$\max_g R_g$	Protect extremes	Conservative
Subgroup Audits	Fine-grained checks	Detect hidden bias	Small samples
Regularization	Shrinkage	Reduce variance	May mask issues

move the model toward majority measurement patterns. In AL prediction, this can manifest as better discrimination for patients whose laboratory trajectories are densely recorded and worse performance for those whose care leaves sparser traces.

Outcome misclassification poses an additional fairness hazard. AL is often not captured by a single indisputable marker. Datasets may define it through combinations of diagnostic codes, returns to the operating room, drainage procedures, antibiotic patterns, imaging evidence, or note review. Each component is subject to differential detection. A small leak treated conservatively may never receive a procedure code. A frail patient may be managed non-operatively and coded differently from a younger patient who undergoes reintervention. A hospital with lower imaging availability may underdetect radiographically subtle events. The observed label $\tilde{Y}$ is therefore a distorted proxy for the latent event Y.

A simple differential misclassification model makes the problem concrete:

$$\begin{aligned} \Pr(\tilde{Y} = \ | \ Y = , A = g, X = x) &= \alpha_g(x) \\ \Pr(\tilde{Y} = \ | \ Y = \varkappa, A = g, X = x) &= \beta_g(x). \end{aligned} \tag{9}$$

The function $\alpha_g(x)$ represents group- and covariate-dependent sensitivity of the observed labeling process, while $\beta_g(x)$ captures false positive labeling. If either varies by group, then fairness metrics computed on $\tilde{Y}$ may not reflect fairness with respect to the true event. For example, a model can appear to have a lower false negative rate in a group that actually has higher outcome ascertainment rather than better clinical performance.

Surveillance bias closely interacts with label noise. Patients judged high risk by clinicians may be imaged sooner, tested more frequently, or kept under closer observation. This increases the chance that a leak is diagnosed and coded. In contrast, patients perceived as low risk may be discharged earlier, scanned less often, or reviewed less intensively, which lowers the probability of detection unless the complication becomes dramatic. If perception of risk differs across groups because of access, communication, prior history, or implicit expectations, the observed outcome becomes partially endogenous to clinician behavior. A model trained on such outcomes learns not merely leak biology but the prior surveillance regime. This creates a feedback problem: using the model to guide future surveillance can amplify the patterns on which it was trained [10].

Treatment-induced label distortion is another source of inequity. Suppose a subgroup is more likely to receive protective diversion, extended antibiotics, or ICU observation because clinicians anticipate difficulty or because institutional policy is more conservative for certain patients. These interventions may reduce the incidence of observed leak or reduce the severity threshold at which it becomes coded. The model then encounters a label shaped by prior treatment. In causal terms, the treatment variable lies on a path from group membership and baseline covariates to outcome. Fitting a naïve predictor on post-treatment features can inadvertently penalize groups for receiving more or less intervention. The issue is not merely confounding but strategic feedback between predictions, actions, and observed endpoints.

Proxy discrimination also deserves attention. Excluding race or sex from the feature set does not prevent the model from inferring group membership through correlated variables such as insurance, zip code proxies, disease stage at presentation, nutritional status, timing of surgery, language-associated documentation patterns, or differential laboratory density. In AL

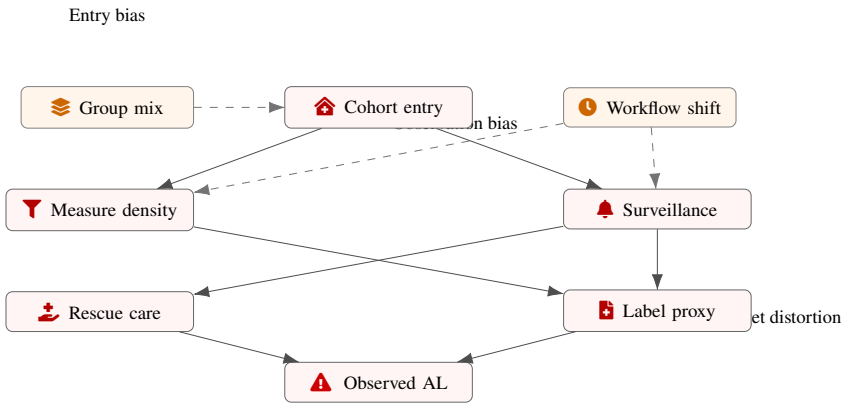

Figure 2 Upstream sources of inequity in AL datasets. Selection into the cohort, differential measurement density, uneven surveillance, rescue practices, and proxy-based outcome encoding jointly determine the observed target, so disparity can be created before model training begins.

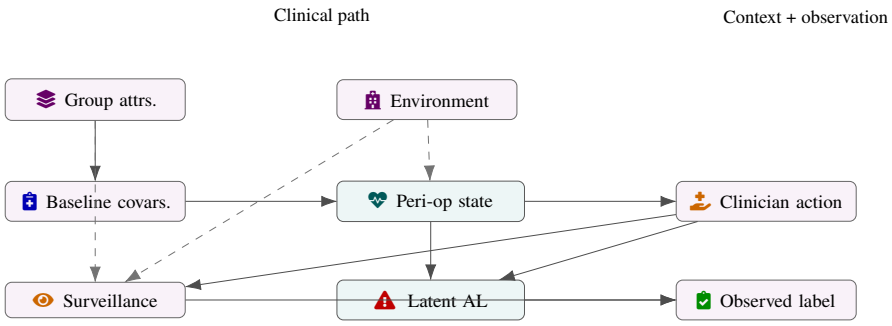

Figure 3 Compact causal sketch under distribution shift. Group attributes and environment influence baseline covariates, peri-operative state, clinician action, and surveillance intensity, while the observed label depends jointly on latent AL and the diagnostic process. Fairness, transport, and label validity are therefore linked.

prediction, anatomical and procedural variables complicate the picture because some group differences are clinically meaningful and should not be erased. The difficulty is to distinguish features that legitimately capture biological or procedural risk from features that encode inequitable access or measurement regimes. This distinction is rarely visible from correlation alone.

Procedure mix heterogeneity adds a layer specific to surgery. Low pelvic anastomoses, urgent operations, inflammatory fields, malnutrition, and hemodynamic instability genuinely affect leak risk. If these factors are unevenly distributed across groups, then prevalence will differ for medically legitimate reasons. A fairness framework that interprets every prevalence difference as bias is therefore unsuitable. At the same time, if some groups systematically present later, sicker, or without optimization because of broader inequities, then treating those differences as purely biological can naturalize structural disadvantage. The statistical model sees only covariates [11]. The fairness analysis must decide when adjustment is clinically justified and when it masks inequity upstream of the operating room.

Temporal drift can make these problems worse. Surgical technique evolves, enhanced recovery pathways change discharge timing, diagnostic thresholds shift, and hospitals revise coding behavior. A model trained on one period may be less fair in a later period even if global performance remains acceptable. If adoption of fluorescence imaging, antibiotic protocols, or stoma practices occurs unevenly across hospitals or patient groups, then the relationship between features and outcome changes in group-specific ways. Static fairness audits therefore understate the challenge. Equitable performance must be monitored over time.

There is also a less discussed but important issue of denominator inequity. Many AL datasets are generated only for patients with adequate documentation, complete laboratory panels, or confirmed procedure coding. Excluding incomplete cases may disproportionately remove patients from particular groups, not because they are statistically inconvenient but because their care was documented under different constraints. In rare-event prediction this can radically alter subgroup composition. A model may appear fair on the retained sample while being effectively undefined for patients most likely to experience fragmented documentation.

Taken together, these mechanisms imply that unfairness in AL prediction can originate long before parameter estimation. Selection filters the population, measurement policies shape the covariates, surveillance practices shape the labels, treatment changes the target, and temporal drift alters all of the above. The standard machine learning abstraction of i.i.d. samples with fixed labels is therefore a poor approximation. Fairness-aware AL modeling must instead treat the dataset as the product of a clinical process in which observation itself is structured by

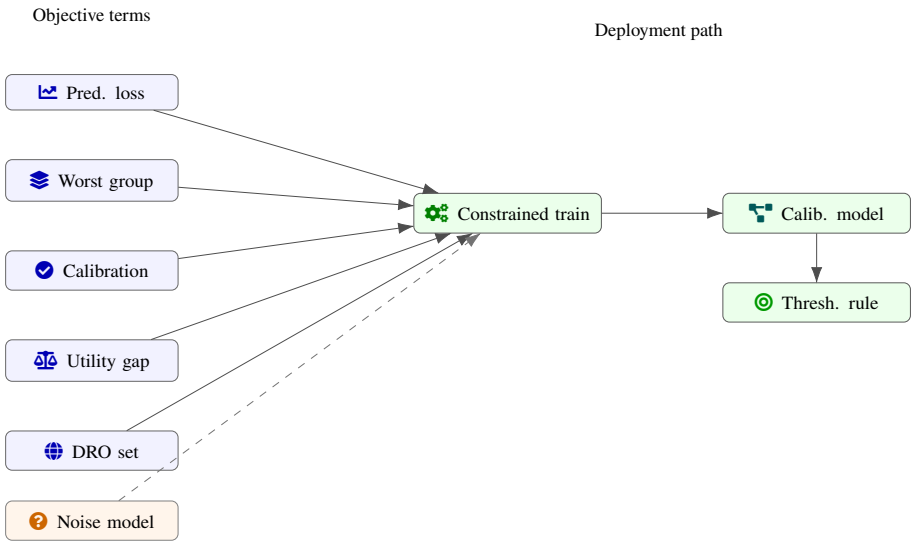

Figure 4 Fairness-constrained estimation template. Average predictive loss is combined with worst-group control, calibration preservation, utility-sensitive disparity penalties, distributional robustness, and a label-noise component before threshold design is chosen for clinical use.

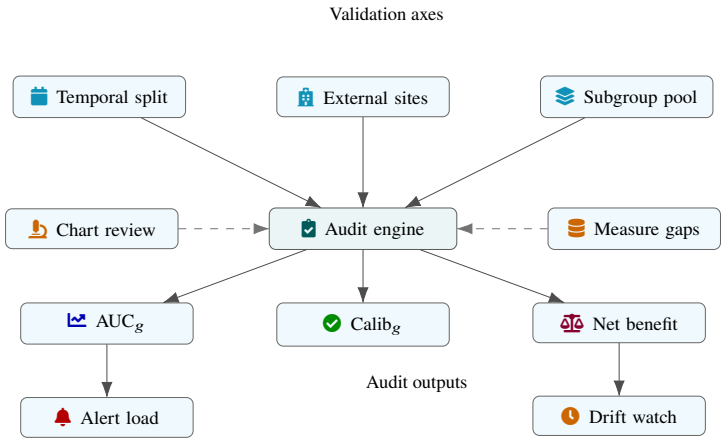

Figure 5 Audit architecture for rare-event fairness evaluation. Temporal validation, external transport, subgroup pooling, chart review, and measurement-gap checks feed a multi-metric audit spanning discrimination, calibration, net benefit, alert burden, and post-deployment drift.

care. This is why representation learning and causal structure matter: they offer ways to distinguish stable clinical signal from environment-specific artifact.

4. Representation Learning, Causal Structure, and Distribution Shift

A useful causal sketch for AL prediction includes latent patient physiology U, baseline observed covariates X^{pre}, intraoperative and immediate postoperative variables X^{peri}, group attributes A, environment variables E such as hospital and time period, clinician actions T, surveillance intensity S, latent AL outcome Y, and observed label $\tilde{Y}$. The arrows are dense rather than sparse. Group attributes and environment influence who reaches surgery, what preoperative optimization occurs, how operation type is selected, how often measurements are ordered, and what rescue measures are deployed [12]. Clinician actions affect both the outcome and the probability that the outcome is documented. The observed data are thus generated by a structured clinical graph, not by a single direct map from features to outcome.

This matters because fairness under distribution shift depends on which parts of the graph remain invariant across environments. A model can appear fair in one hospital because it has learned a stable mapping from postoperative physiology to leak risk, or because it has learned a brittle shortcut such as the meaning of ICU admission within that hospital's workflow. Only the former is likely to transport. The latter may fail when another center uses ICU beds differently or when postoperative observation protocols change. A fairness claim tied to non-transportable shortcuts is weak. If minority groups are concentrated in the environments where such shortcuts break, the model becomes unfair precisely where reliability is most needed.

Representation learning offers one possible response. Let $\phi_\omega(X)$ denote a learned representation and let h_θ map this representation to risk. The objective is to extract information predictive of Y while suppressing nuisance variation associated with environment-specific measurement regimes or spurious

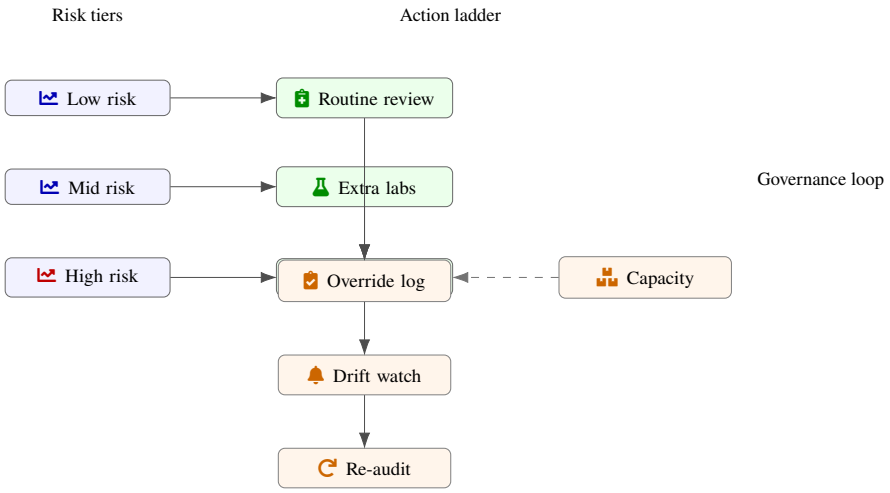

Figure 6 Clinical integration and governance loop. Risk tiers map to an action ladder rather than a single binary alert, capacity constraints remain visible, overrides are logged, and sequential monitoring triggers re-audit when subgroup burden or calibration begins to drift.

group proxies. A general formulation is

$$\begin{aligned}\min_{\omega,\theta} \quad & \sum_{e\in\mathcal{E}} R_e(h_\theta \circ \phi_\omega) \\ & + \lambda\, \Omega_{\text{inv}}(\phi_\omega; E) \\ & + \gamma\, \Omega_{\text{fair}}(h_\theta \circ \phi_\omega; A),\end{aligned} \tag{10}$$

where R_e is predictive risk within environment e, Ω_{inv} penalizes environment dependence in the representation, and Ω_{fair} penalizes group-specific inequity in the resulting score or policy. This formulation is intentionally broad. The invariant term might encourage matched feature moments across hospitals, similar gradients across time periods, or stability of conditional residual structure. The fairness term might penalize worst-group calibration error, thresholded disparity, or utility difference.

However, representation invariance can be overused. If the model is forced to ignore all group-associated information, it may discard real clinical predictors. Pelvic anatomy, procedure type, inflammatory burden, or malnutrition may legitimately correlate with group structure while also carrying genuine leak risk [13]. The correct goal is therefore not unconditional invariance but selective invariance: remove group dependence that operates through inequitable measurement or access pathways while retaining dependence that reflects clinically relevant state. This is fundamentally a causal identification problem. Without assumptions about which paths are admissible, fairness regularization can become a form of indiscriminate information deletion.

One way to make the distinction explicit is through an orthogonal decomposition of representation into stable clinical components and environment- or group-specific nuisance components. Let

$$\begin{aligned}\phi_\omega(X) &= Z^{\text{clin}} + Z^{\text{nuis}}, \\ \langle Z^{\text{clin}}, Z^{\text{nuis}} \rangle &= \varkappa.\end{aligned} \tag{11}$$

The aim is to estimate Z^{clin} so that it predicts Y across environments, while Z^{nuis} absorbs measurement density, coding artifacts, hospital-specific workflow signatures, and other factors that help in-sample but undermine transport. In practice, exact orthogonality is unattainable and the decomposition is only approximate. Still, the concept is valuable because it shifts the fairness question from "Should the model use group-correlated information?" to "Which group-correlated information remains stable and clinically justified under deployment?"

Hospital and temporal shift further complicate fairness because subgroup prevalence and measurement processes can move in different directions over time. Enhanced recovery pathways may reduce length of stay, causing postoperative measurements to truncate earlier for some patients. New diagnostic modalities may increase ascertainment of subtle leaks in one period but not another. Changes in operative technique may lower base rates for some procedure classes but not others. If a model is calibrated on pooled historical data, these shifts can create subgroup-specific calibration drift even when ranking remains stable.

An environment-aware calibration function can partially address this. Let R be the raw score and $C_{g,e}$ a group- and environment-specific recalibration map. Then

$$\begin{aligned}R^* &= C_{g,e}(R) \\ &= \sigma\big(a_{g,e} + b_{g,e}\,\text{logit}(R)\big),\end{aligned} \tag{12}$$

where σ denotes the logistic function. This preserves ranking within a group-environment cell while adjusting score scale to recover local calibration. Yet recalibration alone cannot fix omitted-variable inequity or label distortion. It is a necessary but limited tool.

A more structural approach is to model the observation process jointly with the outcome. Suppose S denotes surveillance intensity and $\tilde{Y}$ the observed label. Then one can write

$$\begin{aligned}\Pr(\tilde{Y} = \,|\, X, A, E) = \sum_{y\in\{\varkappa,\}} & \Pr(\tilde{Y} = \,|\, Y = y, S, A, E) \\ & \times \Pr(Y = y \,|\, X, A, E, S).\end{aligned} \tag{13}$$

This decomposition highlights that observed positivity mixes true risk with ascertainment [14]. If surveillance S is partially observed through imaging orders, laboratory density, or unit transfers, joint modeling can reduce bias by separating physiological risk from diagnostic opportunity. The fairness benefit is substantial: group differences in measured performance can then be interpreted against differential surveillance rather than taken at face value.

Distribution shift also raises the question of what fairness should transport. It is unrealistic to expect identical subgroup metrics across every hospital if case mix genuinely changes. A more credible target is bounded degradation. Let $R_g^{(e)}$ denote a chosen loss for group g in environment e. Then a transport-fair model might be required to satisfy

$$R_g^{(e)} \le R_g^{\mathrm{ref}} + \delta_g \tag{14}$$

for each deployment environment relative to a reference. Here δ_g is a clinically negotiated tolerance rather than a purely statistical threshold. This formulation treats fairness as stability under deployment rather than parity at a single historical snapshot.

The representation problem in AL prediction is therefore inseparable from causal structure and environment heterogeneity. A fair model must preserve clinically meaningful signals, suppress spurious workflow shortcuts, recalibrate under local drift, and remain interpretable enough that clinicians can recognize when the data-generating process has changed. The next question is how to encode these aims directly in estimation and optimization rather than leaving them as post hoc aspirations.

5. Fairness-Constrained Estimation and Optimization

Classical empirical risk minimization minimizes average loss across the pooled training sample. In the AL setting this is often inadequate because the majority group, the majority environment, and the most richly measured workflow dominate the gradient. The model may therefore achieve strong average performance by allocating representational capacity away from smaller groups or noisier contexts. A fairness-aware objective must explicitly resist this tendency.

Let $\ell(y, r)$ denote a proper scoring loss, such as binary cross-entropy, and let G index audited groups. A basic worst-group objective is

$$\min_{\theta} \quad \max_{g \in G} \left\{ \frac{}{n_g} \sum_{i:A_i=g} w_i \, \ell(\overline{Y}_i, f_\theta(X_i)) \right\}, \tag{15}$$

where w_i may include class-imbalance weights, inverse-propensity weights for selection correction, or costs reflecting clinical action asymmetry [15]. This minimax formulation discourages the optimizer from sacrificing the hardest or smallest group to improve the pooled average. In rare-event surgery, worst-group control is often more relevant than mean performance because clinicians care about whether the model fails systematically in a particular population, not whether the average score improves by a few points.

Yet worst-group loss alone is not enough. A model can equalize cross-entropy while remaining poorly calibrated or clinically misaligned. Thus one may consider a composite objective

$$\begin{aligned} \min_{\theta} \quad & L_{\mathrm{pred}}(\theta) \\ & + \lambda L_{\mathrm{wg}}(\theta) \\ & + \lambda L_{\mathrm{cal}}(\theta) \\ & + \lambda_{\mathbb{A}} L_{\mathrm{disp}}(\theta), \end{aligned} \tag{16}$$

where L_{pred} is pooled predictive loss, L_{wg} a worst-group or high-quantile group loss, L_{cal} a calibration penalty, and L_{disp} a disparity penalty on thresholded or utility-based outcomes. Each term corresponds to a different dimension of fairness. The calibration term preserves interpretability of risk, the worst-group term protects minority performance, and the disparity term constrains unequal error burden or unequal intervention value.

Because AL is rare, the choice of weighting is nontrivial. If positive examples receive heavy class weights, the optimizer emphasizes sensitivity, which can be beneficial for groups with historically underdetected leaks. But if label noise is differential, upweighting positives may also amplify mislabeled cases in certain groups. A robust alternative is to use group-conditional focal weighting or smoothed inverse prevalence weights. Let π_g denote the observed positive rate in group g. Then a class-group weighting scheme may take the form

$$w_i = \begin{cases} (\pi_{A_i} + \epsilon)^{-/}, & \overline{Y}_i = , \\ (- \pi_{A_i} + \epsilon)^{-/}, & \overline{Y}_i = \varkappa, \end{cases} \tag{17}$$

which avoids the instability of full inverse-frequency weighting while still moderating the dominance of common negatives in high-prevalence or low-prevalence groups.

Fairness constraints may also be written explicitly rather than penalized softly. Suppose the deployment goal is groupwise calibration within tolerance and bounded recall disparity at a candidate threshold τ. One can pose

$$\begin{aligned} \min_{\theta, \tau} \quad & L_{\mathrm{pred}}(\theta) \\ \text{subject to} \quad & |\widehat{\mathrm{ECE}}_g(\theta)| \le \varepsilon_g \quad \hbar g \in G \\ & |\widehat{\mathrm{TPR}}_g(\theta, \tau) - \widehat{\mathrm{TPR}}_{g'}(\theta, \tau)| \le \delta_{gg'} \quad \hbar g, g' \in G. \end{aligned} \tag{18}$$

Here $\widehat{\mathrm{ECE}}_g$ denotes a groupwise expected calibration error estimate. This formulation is appealing because it expresses fairness as an explicit requirement rather than a secondary preference [16]. Its drawback is feasibility: when prevalence differs, the constraint set may be empty or may admit only poorly calibrated, low-utility models. In practice one often solves the Lagrangian relaxation and inspects the trade-off curve rather than insisting on exact parity.

A clinically useful extension replaces raw recall disparity with disparity in net benefit. Let $\mathrm{NB}_g(\tau)$ denote a groupwise

net benefit function adapted from decision analysis:

$$\mathrm{NB}_g(\tau) = \frac{}{n_g} \sum_{i:A_i=g} \Big[D_i(\tau) Y_i - \kappa(\tau) D_i(\tau)(-Y_i) \Big], \tag{19}$$

where $\kappa(\tau)$ is the harm-to-benefit exchange rate implied by threshold τ. Parity in NB_g can be more meaningful than parity in false positive rate because it accounts for the clinical value of true positives. A surveillance-oriented AL model might tolerate more false positives in a subgroup if the induced early detections produce proportionately greater benefit. Conversely, equal false positive rates can still be inequitable if the same alert burden yields little benefit in one group and substantial benefit in another.

There is also the question of group-specific versus group-agnostic thresholds. A single threshold is often preferred for procedural simplicity and perceptions of formal equality. However, if calibration differs slightly by group or if the clinical utility of action varies because of anatomy, procedure class, or follow-up capacity, a common threshold can lead to unequal effective treatment standards. Group-specific thresholds can improve utility parity and, under some circumstances, error-rate parity as well. But they are ethically and operationally contentious, especially when the grouping variable is protected. In AL prediction a cautious compromise is to use clinically justified stratified thresholds for procedure or care pathway while maintaining groupwise calibration audits across demographic attributes. This does not solve every fairness issue, but it reduces the risk of hidden inequity from a superficially uniform decision rule.

Distributionally robust optimization is particularly attractive in this domain because hospitals and subgroups can be seen as environments over which uncertainty persists. Consider the ambiguity set

$$Q = \Big\{ q \in \Delta^{|G|} : D_\phi(q, u) \le \rho \Big\}, \tag{20}$$

where u is the empirical group distribution and D_ϕ a divergence measure. A distributionally robust objective seeks

$$\min_\theta \max_{q \in Q} \sum_{g \in G} q_g R_g(\theta). \tag{21}$$

This formulation protects against reweightings toward underrepresented or high-risk groups within a radius ρ. The fairness value is that the model is not allowed to rely excessively on the observed training mix. If deployment sees a different subgroup composition, performance degrades less sharply [17].

Because observed labels may be noisy, fairness-aware optimization should ideally include noise correction. Suppose groupwise label sensitivity and specificity estimates, α_g and β_g, are available from chart review or adjudicated subsamples. Then corrected risk contributions may be formed by replacing the raw loss with an unbiased or approximately unbiased surrogate under differential label noise. While exact correction can be unstable, even partial adjustment helps prevent the optimizer from treating ascertainment patterns as truth. In AL modeling this is especially relevant when one group undergoes more imaging or more aggressive reintervention, thereby inflating observed positivity relative to latent event occurrence.

Another practical design is multi-task learning. Instead of predicting AL alone, the model can jointly predict related postoperative states such as sepsis, reintervention, prolonged ileus, unexpected ICU escalation, or imaging-confirmed intra-abdominal complication. Shared structure across these tasks can stabilize representation for small groups and reduce overfitting to idiosyncratic AL coding. A fairness-aware multi-task loss may be written as

$$\min_\theta \sum_{k=}^{K} \alpha_k L^{(k)}(\theta) + \lambda L_{\mathrm{fair}}(\theta), \tag{22}$$

where one task is AL and others capture clinically adjacent states. This approach is not a guarantee of fairness, but it can make subgroup estimation more data-efficient by borrowing signal from correlated outcomes.

The final optimization issue concerns interpretability. In high-stakes surgery, fairness is not only about metric compliance but also about whether clinicians can understand why a patient was flagged and whether an observed disparity is clinically sensible or suspicious. Sparse additive models, monotone gradient boosting, generalized additive models with pairwise interactions, or hierarchical logistic models may therefore be preferable to opaque architectures if they allow transparent examination of groupwise behavior. A modest sacrifice in pooled discrimination may be acceptable if it yields stronger calibration, clearer error analysis, and easier detection of nontransportable shortcuts. Fairness and interpretability are not identical, but in AL prediction they are often aligned because hidden shortcut learning is a common path to both unfairness and deployment failure.

6. Evaluation, Calibration, and Audit Design

A fairness claim for an AL prediction model is only as credible as the audit supporting it. Because the event is rare and the data-generating process is heterogeneous, conventional single-number reporting is inadequate. Evaluation must be layered, uncertainty-aware, and prospective enough to anticipate deployment conditions rather than merely summarize retrospective fit.

The first requirement is disaggregated discrimination [18]. Overall area under the receiver operating characteristic curve is too blunt. One needs groupwise discrimination estimates, ideally for both ranking and thresholded operation. For each group g, define

$$\mathrm{AUC}_g = \Pr(R_i > R_j \mid Y_i = , Y_j = \varkappa, A_i = A_j = g). \tag{23}$$

This quantity should be reported with uncertainty intervals, not point estimates alone, because small subgroup counts make

nominal differences unstable. Where sample size is limited, partial pooling through hierarchical models is preferable to unregularized subgroup estimates, since the goal is not to maximize apparent disparity but to estimate it responsibly.

Calibration requires equal attention. Groupwise calibration error can be assessed through binning, smoothing, or parametric recalibration, but a fairness audit should look beyond a single expected calibration error statistic. In rare-event settings, coarse bins may conceal high-risk tail miscalibration, which is exactly where clinical intervention occurs. A groupwise integrated calibration error can be written as

$$\mathrm{ICE}_g = \int_{\varkappa} \left| \Pr(Y = \ \mid R = r, A = g) - r \right| dF_g(r), \tag{24}$$

where F_g is the group-specific score distribution. Tail-focused variants that upweight the upper quantiles of risk may be more relevant for postoperative surveillance. Calibration plots should be stratified by group, environment, and time period whenever feasible.

Thresholded performance must be expressed in clinically interpretable terms. Recall, positive predictive value, false positive rate, and false negative rate should all be reported by group at candidate intervention thresholds. However, because thresholds are policy choices, audits should not freeze fairness analysis at a single operating point chosen on pooled data. Instead, the threshold-performance curve for each group should be examined, together with groupwise net benefit and alert volume. This allows stakeholders to see whether a common threshold imposes disproportionate burden on a subgroup or whether threshold tuning produces reasonable trade-offs.

Decision-focused fairness can be summarized by groupwise expected utility and intervention exposure [19]. Let

$$\begin{aligned} \mathrm{Exp}_g(\tau) &= \Pr(D = \ \mid A = g), \\ \mathrm{FN}_g(\tau) &= \Pr(D = \varkappa \mid Y = , A = g). \end{aligned} \tag{25}$$

A fairness audit should inspect whether groups with similar latent need receive similar intervention exposure and whether groups with greater vulnerability bear disproportionate missed-event burden. Such analysis is necessarily value-laden because one must define what counts as need and what action burden is tolerable. Still, making these choices explicit is better than hiding them behind aggregate discrimination.

Intersectional analysis is indispensable. Reporting fairness across sex alone or race alone can mask failures in small but clinically important subgroups such as older frail women after urgent left-sided resection or patients with limited preoperative optimization in resource-constrained settings. Yet exhaustive intersectional auditing is statistically unstable if handled naïvely. A practical approach is hierarchical shrinkage. Let η_g denote a subgroup effect for calibration slope, log loss, or thresholded recall. Then

$$\eta_g \sim \mathcal{N}(\mu, \tau) \tag{26}$$

allows partial pooling across related subgroups, stabilizing estimates while still permitting detection of true outliers. The result is not a replacement for raw counts but a principled way to avoid overreacting to sparse noise or underreacting to systematic deviation.

Temporal validation is another non-negotiable component. Fairness that exists only in a random split of historical data can vanish when practice patterns change. At minimum, one should train on earlier periods and evaluate on later periods, tracking groupwise calibration drift and thresholded disparities. If the model will be deployed across centers, leave-one-site-out or external-center validation is needed. The fairness question under transport is not simply whether aggregate AUC falls, but whether the fall is concentrated in specific populations or workflows. A model that remains globally adequate while becoming poorly calibrated for one hospital or one demographic group is not deployment-ready.

The audit must also address uncertainty from label noise. If chart review is available for a stratified subset, subgroup fairness estimates should be recalculated on adjudicated labels and compared with the main analysis. When chart review is unavailable, sensitivity analysis is still possible [20]. Suppose group-specific label sensitivity lies within $[\underline{\alpha}_g, \overline{\alpha}_g]$ and false positive rate within $[\underline{\beta}_g, \overline{\beta}_g]$. Then fairness metrics can be bounded under those ranges rather than asserted with unwarranted precision. In AL prediction, where observed outcomes may depend on reintervention thresholds and coding conventions, such sensitivity analysis is more honest than reporting subgroup disparities as if the label were exact.

Auditing should also include measurement fairness. For each key variable, one can estimate groupwise missingness, timing, and measurement density:

$$\begin{aligned} m_{jg} &= \Pr(M_j = \varkappa \mid A = g), \\ d_{jg} &= \mathbb{E}[N_j \mid A = g], \end{aligned} \tag{27}$$

where N_j is the number of times variable j was measured in the relevant window. Large disparities may indicate that model fairness is being judged on unequal information quality. In such cases a model can be "fair" only for patients who generate sufficient data, which is a restricted and clinically unsatisfactory sense of fairness.

Calibration after deployment deserves separate treatment because closed-loop effects can change the label process. Once clinicians begin acting on the model, true event rates among flagged patients may fall due to successful rescue, while diagnosis rates may rise due to increased surveillance. This can make the model look miscalibrated even when it is clinically helpful, or conversely, can hide deterioration if actions differ by group. Therefore prospective monitoring should track both outcome incidence and action pathways. A dynamic calibration monitor may update groupwise residual summaries over time:

$$\begin{aligned} G_{t,g} = \rho G_{t-,g} \\ + (- \rho)\Big(Y_{t,g} - R_{t,g}\Big), \end{aligned} \tag{28}$$

where ρ is a smoothing factor and the notation suppresses aggregation within the incoming batch. Persistent drift in $G_{t,g}$ indicates calibration change that may be group specific.

Conformal uncertainty sets or prediction intervals can also support fairness, though not by themselves. If the model outputs a risk score together with a calibrated uncertainty band, clinicians can modulate reliance when subgroup data are sparse or environment shift is suspected. Group-conditional conformal calibration may be useful when sample size permits, since it avoids pooling away subgroup uncertainty. In rare-event surgery this is attractive because many fairness failures are failures of overconfidence.

Finally, evaluation must remain connected to downstream action [21]. A model can satisfy many fairness metrics and still worsen care if it redistributes attention away from patients who are clinically difficult to evaluate. Therefore simulation of workflow impact is important. One should estimate how many additional scans, ICU days, senior reviews, or prolonged observations are triggered in each group under candidate policies, and compare those exposures to expected benefits. Fairness is undermined when one population bears the burden of increased monitoring without commensurate gain, even if statistical parity looks acceptable.

An AL fairness audit is thus necessarily multidimensional. It must report groupwise discrimination, calibration, thresholded burden, utility, uncertainty, measurement quality, transport stability, and post-deployment drift. Anything simpler is likely to miss the mechanisms through which inequity actually enters surgical decision support.

7. Clinical Integration, Governance, and Human Factors

Even a technically careful fairness-aware model can fail in practice if integration is poorly designed. In AL prediction the output rarely acts alone. Surgeons, intensivists, ward teams, radiologists, and advanced practice clinicians interpret risk in the context of an evolving postoperative course. The fairness of the system therefore depends on how the score is displayed, who receives it, when it appears, what actions it recommends, and how overrides are handled.

Timing is the first governance decision. A preoperative fairness problem concerns who is flagged for nutritional optimization, counseling, diversion consideration, or referral to higher-level centers. A postoperative fairness problem concerns who receives intensified surveillance or earlier imaging. The same patient may warrant different fairness considerations at these stages because the relevant counterfactual action differs. If a model is updated dynamically over time, the system should preserve continuity of interpretation so that a rising score reflects changed state rather than arbitrary rescaling. Otherwise clinicians may learn to distrust subgroup differences because the meaning of the score drifts across phases.

Dynamic AL surveillance can be expressed as a state-space update. Let X_{it} denote time-indexed postoperative features and let S_{it} be a latent severity state. Then

$$\begin{aligned} S_{it} &= \alpha S_{i,t-} + \beta^\top X_{it} + \xi_{it} \\ R_{it} &= \sigma\big(\gamma_\varkappa + \gamma S_{it}\big). \end{aligned} \tag{29}$$

A fairness-aware deployment should monitor whether the state update behaves comparably across groups when data arrive at different frequencies [22]. If one group generates fewer measurements, the latent state may update more slowly, which can create delayed alerts despite comparable underlying deterioration. This is an integration issue as much as a modeling issue, because the institution can choose to standardize measurement intervals for patients within certain risk bands.

The user interface should present both risk and rationale. In surgery, black-box outputs often fail not because clinicians demand full mechanistic truth, but because they need to know whether the score is responding to plausible signals or to workflow artifacts. A fairer interface is one that helps detect unfairness. If the model highlights elevated risk because of persistent leukocytosis, rising lactate, and delayed bowel recovery, the clinician can judge whether the recommendation is clinically coherent. If it highlights a cluster of administrative or context-dependent variables, concern about shortcut learning is reasonable. Local explanations are imperfect, but they can support accountability and help surface group-specific anomalies.

Override analysis is particularly important. Clinicians will and should ignore model recommendations at times. However, override behavior itself can become inequitable. If alerts for certain groups are more likely to be dismissed, or if false alarm tolerance differs by unit or surgeon, then the downstream policy may become unfair even if the model is fair. Governance should therefore record alerts, responses, actions taken, and documented reasons for override. Equity review should examine whether similar risk patterns lead to similar responses across groups and environments.

Resource constraints cannot be ignored. Fairness in AL prediction is often discussed as though every flagged patient could receive the same enhanced pathway. In reality, access to CT scanning, ICU beds, overnight senior review, laboratory turnaround, and interventional radiology varies. A fair model deployed into unequal capacity can still yield unequal benefit. This implies that fairness evaluation must include capacity-aware simulation [23]. If only a fixed number of enhanced surveillance slots are available, then ranking fairness becomes crucial because lower-ranked patients effectively receive no intervention. Conversely, if capacity varies by hospital, a shared threshold may create very different action rates across sites.

An operationally meaningful governance model should therefore define an intervention cascade rather than a binary alert. For example, moderate risk may trigger repeat examination and laboratory review, higher risk may trigger imaging consideration, and very high risk may prompt senior consultation or prolonged monitored stay. Fairness can then be assessed for each action tier. This is preferable to treating every alert as identical, because the burden and benefit of each action differ. A subgroup may reasonably receive more low-burden monitoring without implying unfair over-intervention, whereas disproportionate exposure to invasive follow-up would raise stronger concern.

Prospective fairness surveillance after deployment should be institutionalized rather than ad hoc. A monitoring statistic can be maintained for groupwise alert burden, confirmed leak capture, and calibration drift. One possible sequential detector

is

$$H_{t,g} = \max\left\{\varkappa,\ H_{t-,g} + Z_{t,g} - k_g\right\}, \tag{30}$$

where $Z_{t,g}$ is a batch-level discrepancy statistic and k_g a tolerated drift level. When $H_{t,g}$ crosses a review threshold, the subgroup-environment cell enters audit. This kind of sequential monitoring is familiar in industrial quality control and is well suited to high-stakes clinical models because it detects deterioration without requiring complete model retraining.

Governance must also include representation from clinical and equity stakeholders. Decisions about audited groups, clinically acceptable disparity, threshold selection, and response pathways are not purely technical. They involve trade-offs among sensitivity, burden, resource use, and norms of equitable care. A fair AL system cannot be defined solely by developers because the meaning of false alarm burden, acceptable missed-event risk, and justified stratification depends on clinical values. At the same time, governance should avoid substituting vague ethical language for measurable commitments. Each fairness objective should be tied to an operational metric, a review cadence, and a response plan.

Data provenance and documentation are part of this governance layer [24]. For each model version, the institution should record the development population, label definition, missingness profile, audited groups, calibration status, external validation contexts, and known limitations. This is especially important for AL prediction because many features are tightly linked to local workflows. A model trained in one environment may rely on postoperative measurement patterns that do not exist elsewhere. Transparent documentation is not ancillary to fairness; it is one of the main ways clinicians can recognize when a fairness claim no longer applies.

Finally, human factors research should examine whether the system redistributes cognitive labor unevenly. If one subgroup produces more ambiguous alerts, clinicians may spend more time adjudicating those cases. If the interface is less interpretable for cases with sparse data, some patients may be effectively deprioritized despite formal inclusion in the model. Equitable deployment therefore requires not only mathematical fairness but also workflow designs that do not shift hidden burdens onto already disadvantaged patients or overextended clinicians.

A fair AL prediction system, then, is not just a constrained estimator. It is a sociotechnical arrangement in which data collection, score generation, explanation, escalation pathways, overrides, capacity limits, and post-deployment monitoring are jointly specified. Without that broader design, fairness metrics remain abstract and can even be misleading by suggesting equity where the actual intervention pathway remains unequal.

8. Conclusion

Fairness in AL prediction is a harder problem than fairness in many standard tabular prediction tasks because the outcome is rare, the labels are imperfect, the care pathway is adaptive, and the model's output can itself change future event ascertainment. These features make simplistic solutions unreliable. Removing sensitive variables does not remove inequity because proxies and environment-specific workflows remain. Optimizing a single fairness metric does not solve the problem because calibration, thresholded burden, utility, and transport stability cannot generally be equalized at once when prevalence and surveillance differ across groups. Treating fairness as a post hoc dashboard applied to a finished model is likewise insufficient, since many disparities arise from selection, measurement, and labeling before training even begins.

A more coherent approach views fairness as a layered property of the entire AL decision pipeline. At the data layer, one must confront informative missingness, differential surveillance, label uncertainty, and site-specific measurement conventions. At the modeling layer, one must constrain empirical risk minimization so that minority groups, rare environments, and high-uncertainty settings are not absorbed into average performance [25]. At the evaluation layer, one must report subgroup discrimination, calibration, thresholded burden, utility, uncertainty, and transport behavior rather than relying on pooled summary metrics. At the deployment layer, one must specify action tiers, capacity limits, override handling, prospective monitoring, and re-audit triggers. None of these layers can be neglected without weakening the fairness claim.

The mathematical perspective developed here suggests several practical principles. Worst-group and distributionally robust objectives are useful because pooled optimization is naturally inattentive to minority structures. Groupwise calibration remains indispensable because AL prediction is consumed as risk communication, not only as ranking. Utility-sensitive auditing is necessary because statistical parity can conceal unequal benefit or unequal burden. Representation learning should target selective invariance rather than indiscriminate group erasure, preserving clinically meaningful signal while discouraging workflow shortcuts that fail under transport. Dynamic monitoring after deployment is required because the model changes the care process and therefore changes the data-generating mechanism on which its fairness was initially judged.

Equitable AL prediction is therefore unlikely to emerge from a single architectural innovation. It will depend on careful outcome definition, better adjudicated datasets, multi-environment validation, hierarchical subgroup estimation, and governance that links fairness metrics to clinical action. The technical task is to build models that remain informative without becoming dependent on nontransportable artifacts. The clinical task is to ensure that those models alter monitoring and rescue pathways in ways that do not unevenly privilege some patients over others. The institutional task is to keep auditing after deployment rather than treating fairness as certified once and for all.

In this sense, fairness in AL prediction should be understood less as a constraint that narrows model development and more as a discipline that makes the modeling problem properly specified. A system that predicts leaks well only for richly observed patients, only in one hospital workflow, or only under one pattern of rescue intervention is not fully solving the prediction problem. It is solving a restricted and potentially inequitable version of it. A fairer framework widens the definition of success to include robustness across patient groups, transport across settings, interpretability of clinical rationale, and stability of

downstream benefit. That broader definition is more demanding, but it is also more aligned with what a surgical decision support tool must achieve if it is to be trusted and used responsibly [26].

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] B. C. Toh, J. Chong, B. P. Yeung, C. H. Lim, E. K. Lim, W. H. Chan, and J. T. Tan, "Endoscopic internal drainage with double pigtail stents for upper gastrointestinal anastomotic leaks: suitable for all cases?," *Clinical endoscopy*, vol. 55, pp. 401–407, 1 2022.

[2] H. M. Schardey, U. Wirth, T. Strauss, M. S. Kasparek, D. Schneider, and K. W. Jauch, "Prevention of anastomotic leak in rectal cancer surgery with local antibiotic decontamination: a prospective, randomized, double-blind, placebo-controlled single center trial," *International journal of colorectal disease*, vol. 35, pp. 847–857, 2 2020.

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

[4] P. K. Simoes, K. M. Woo, M. Shike, R. B. Mendelsohn, H. Gerdes, A. J. Markowitz, E. Ludwig, P. Shah, and M. A. Schattner, "Direct percutaneous endoscopic jejunostomy: Procedural and nutrition outcomes in a large patient cohort," *JPEN. Journal of parenteral and enteral nutrition*, vol. 42, pp. 898–906, 12 2017.

[5] B. K. Sah, M.-M. Chen, M. Yan, and Z. Zhu, "Reoperation for early postoperative complications after gastric cancer surgery in a chinese hospital," *World journal of gastroenterology*, vol. 16, pp. 98–103, 1 2010.

[6] A. D. Politano, T. Hranjec, L. H. Rosenberger, R. G. Sawyer, and C. A. T. Leon, "Differences in morbidity and mortality with percutaneous versus open surgical drainage of postoperative intra-abdominal infections: A review of 686 cases," *The American surgeon*, vol. 77, pp. 862–867, 7 2011.

[7] J. Halle-Smith, S. Kamarajah, R. Evans, and E. Griffiths, "Ogc o04 a preoperative scoring model to estimate the risk of conduit necrosis in oesophagectomy patients - results from the oesophago-gastric anastomosis audit," *British Journal of Surgery*, vol. 109, 12 2022.

[8] H. T. Bui and S. T. F. Chan, "Laparoscopic approach for esophagojejunal anastomotic leak in patients requiring relook intervention: How we do it," *Surgical Case Reports*, vol. 2021, pp. 1–3, 3 2021.

[9] R. Jaibaji, M. Ibrahim, A. Tsai, and C. Peters, "Ogc p67 extended use of feeding jejunostomy and reduction of weight loss in patients with total gastrectomy for malignancy," *British Journal of Surgery*, vol. 110, 11 2023.

[10] M. Fasting, D. Førland, C. Skagemo, and T. Mala, "181. safety and time usage of circular stapled anastomosis in the shift from mie to ramie," *Diseases of the Esophagus*, vol. 36, 8 2023.

[11] C. W. Hicks, R. A. Hodin, and L. Bordeianou, "Possible overuse of 3-stage procedures for active ulcerative colitis.," *JAMA surgery*, vol. 148, pp. 658–664, 7 2013.

[12] S. E. Hwang, M. J. Jung, B. H. Cho, and H. C. Yu, "Clinical feasibility and nutritional effects of early oral feeding after pancreaticoduodenectomy," *Korean journal of hepato-biliary-pancreatic surgery*, vol. 18, pp. 84–89, 8 2014.

[13] S. Sakuramoto, H. Katai, H. Katayama, S. Iwamoto, Y. Hasegawa, T. Nomura, S. Makino, M. Watanabe, T. Omori, T. Yoshikawa, T. Ojima, and M. Terashima, "Long-term outcomes after laparoscopy-assisted total or proximal gastrectomy with nodal dissection for clinical stage i gastric cancer: Japan clinical oncology group study (jcog1401).," *Journal of Clinical Oncology*, vol. 41, pp. 305–305, 2 2023.

[14] E. E. Hutchings, O. G. Townley, R. M. Lindley, and G. V. S. Murthi, "The role of stomas in the initial and long-term management of hirschsprung disease.," *Journal of pediatric surgery*, vol. 58, pp. 236–240, 10 2022.

[15] F. J. van der Sluis, A. M. Couwenberg, G. H. de Bock, M. P. W. Intven, O. Reerink, B. van Leeuwen, and H. L. van Westreenen, "Population-based study of morbidity risk associated with pathological complete response after chemoradiotherapy for rectal cancer," *The British journal of surgery*, vol. 107, pp. 131–139, 1 2020.

[16] Y. Liu, C. Gao, G. Wang, Y. C. Wang, X. Z. Lu, and G. Han, "The clinical values of neutrophil-to-lymphocyte ratio as an early predictor of anastomotic leak in postoperative rectal cancer patients," *Zhonghua zhong liu za zhi [Chinese journal of oncology]*, vol. 42, pp. 70–73, 1 2020.

[17] S. Y. Ong, Z. Z. X. Tan, N. Z. Teo, and J. C. Y. Ngu, "Surgical considerations for the "perfect" colorectal anastomosis.," *Journal of gastrointestinal oncology*, vol. 14, pp. 2243–2248, 10 2023.

[18] K. Tuncer, İsmail Sert, G. Kilinc, C. Tuğmen, and M. Emiroğlu, "The effect of preoperative nutritional support on postoperative morbidity and mortality in patients with gastric cancer: A single center retrospective study," *Medical Records*, vol. 3, pp. 214–219, 9 2021.

[19] Y. Rudnicki, I. White, V. Tiomkin, L. Lahav, B. Raguan, and S. Avital, "Intraoperative evaluation of colorectal anastomotic integrity: a comparison of air leak and dye

leak tests," *Techniques in coloproctology*, vol. 25, pp. 841–847, 4 2021.

[20] U. S. Sibia, S. B. Badve, A. C. Istl, J. R. Klune, and A. I. Riker, "Impact of frailty upon surgical decision-making for left-sided colon cancer.," *Ochsner journal*, vol. 23, pp. 120–128, 4 2023.

[21] Z. Adamová and R. Slováček, "Effects of preoperative carbohydrate drinks on postoperative outcome after colorectal surgery," *European Surgery*, vol. 49, pp. 180–186, 7 2017.

[22] S. Irtan, A. Maisin, V. Baudouin, Y. Nivoche, R. Azoulay, E. Jacqz-Aigrain, A. E. Ghoneimi, and Y. Aigrain, "Renal transplantation in children: critical analysis of age related surgical complications.," *Pediatric transplantation*, vol. 14, pp. 512–519, 1 2010.

[23] D. Opoku, A. Hart, D. T. Thompson, C. G. Tran, M. O. Suraju, J. Chang, S. Boatman, A. Troester, P. Goffredo, and I. Hassan, "Equivalency of short-term perioperative outcomes after open, laparoscopic, and robotic ileal pouch anal anastomosis. does procedure complexity override operative approach?," *Surgery open science*, vol. 9, pp. 86–90, 5 2022.

[24] A. K. Mathur, S. N. Nadig, S. Kingman, D. Lee, K. Kinkade, C. J. Sonnenday, and T. H. Welling, "Internal biliary stenting during orthotopic liver transplantation: anastomotic complications, post-transplant biliary interventions, and survival," *Clinical transplantation*, vol. 29, pp. 327–335, 2 2015.

[25] C. Baltin, F. Kron, A. Urbanski, T. Zander, A. Kron, F. Berlth, R. Kleinert, M. Hallek, A. H. Hoelscher, and S.-H. Chon, "The economic burden of endoscopic treatment for anastomotic leaks following oncological ivor lewis esophagectomy.," *PloS one*, vol. 14, pp. e0221406–, 8 2019.

[26] E. Paulus, C. Ripat, V. P. Koshenkov, A. T. Prescott, K. Sethi, H. Stuart, G. Tiesi, A. S. Livingstone, and D. Yakoub, "Esophagectomy for cancer in octogenarians: should we do it?," *Langenbeck's archives of surgery*, vol. 402, pp. 539–545, 3 2017.