Scidata Consortium
An international open-access platform

SDC SciDataConsortium

Research Article

OPEN ACCESS

Service-Aware Continuous Prioritization for ICU Deterioration Under Queue and Attention Constraints

Michal Kaminski[1], Grzegorz Lewandowski[2], Rafal Kubiak[3]

[1] Siedlce University of Natural Sciences and Humanities, Institute of Computer Science, 3 Maja 54 Street, 08-110 Siedlce, Poland; [2] Medical University of Lublin, Department of Public Health and Epidemiology, Aleje Raclawickie 1 Street, 20-059 Lublin, Poland; [3] University of Bialystok, Faculty of Mathematics and Computer Science, Ciołkowskiego 1M Street, 15-245 Bialystok, Poland

Abstract

Anastomotic leak prediction sits at an unusual intersection of surgical risk modeling, postoperative surveillance, and institutional decision making. Background conditions in this domain make the fairness question especially difficult. Anastomotic leak is infrequent, variably defined, often detected late, and documented through workflows that differ across hospitals and patient populations. These features mean that a model can seem technically competent while still allocating reassurance, scrutiny, and clinical attention unevenly. In this setting, fairness cannot be reduced to whether a classifier uses protected variables or whether one summary metric differs across subgroups. It is shaped earlier, at the moment when the outcome is defined, the cohort is assembled, and the benchmark is declared adequate for translation. This paper develops an account of fairness in anastomotic leak prediction by treating it as an epistemic property of the full research pipeline. The central claim is that fairness in this area depends on how knowledge is produced as much as on how predictions are computed. Outcome construction, data provenance, missingness, surveillance intensity, external validation strategy, workflow design, alert burden, and override behavior all influence whether a model distributes benefit and error in a clinically acceptable way. The discussion therefore shifts attention away from narrow metric debates toward benchmark governance, institutional heterogeneity, implementation accountability, and reproducible audit design. Rather than assuming that fairness can be certified by a single statistical test, the paper argues that equitable leak prediction requires a layered evidentiary standard. Under that standard, a model is considered fair only when its target is clinically coherent, its labels are defensible, its subgroup behavior is visible, its deployment pathway is specified, and its claims remain stable when moved beyond the setting in which it was first developed.

1. Introduction

Anastomotic leak remains one of the most consequential complications in gastrointestinal surgery because it combines low incidence with high clinical cost and substantial downstream uncertainty [1]. The event is not merely a binary postoperative endpoint. It is an evolving clinical process shaped by tissue perfusion, operative technique, inflammatory response, rescue timing, diagnostic attention, and the institutional capacity to recognize deterioration before collapse. That complexity has made AL prediction a compelling target for machine learning and clinical risk modeling. If high-risk patients could be identified early enough, clinicians might intensify monitoring, revisit diversion strategy, adjust discharge thresholds, or allocate specialist review more effectively. Yet the promise of prediction is inseparable from the question of who benefits from that prediction and under what evidentiary conditions the model is permitted to influence care.

Fairness in AL prediction is often introduced as if it were a specialized extension of standard algorithmic ethics. That framing is too narrow. In many areas of prediction, fairness debates begin once a reasonably stable target variable already exists. In AL modeling, by contrast, the target itself is unstable. One dataset may define leak through diagnosis codes, another through reoperation, another through chart review, another through a mixture of drainage procedures, antibiotic exposure, imaging, and narrative adjudication. One hospital may detect subtle leaks through aggressive early imaging, whereas another

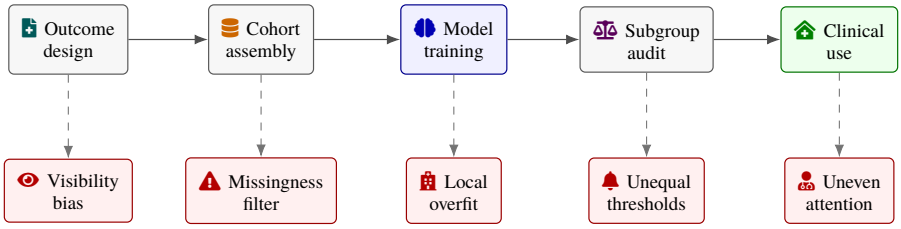

Figure 1: Epistemic fairness can be represented as a pipeline property rather than a late-stage model property. The figure separates the main translational path from five recurring failure points: outcome visibility bias, exclusion through missingness, institution-specific overfitting, threshold asymmetry in subgroup evaluation, and operationally uneven downstream attention.

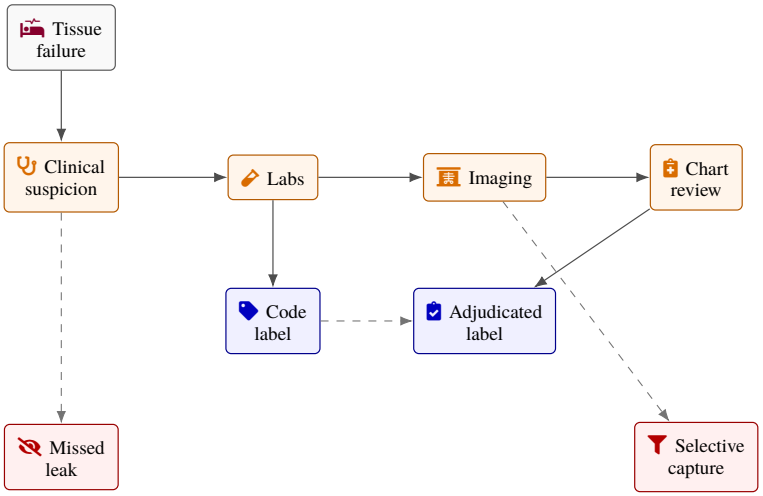

Figure 2: Outcome construction is shown as a progressive visibility chain rather than a direct readout of pathophysiology. The target becomes increasingly mediated by suspicion, tests, review, and coding, so label quality depends on how the clinical system recognizes and records events. Fairness problems emerge when recognition pathways differ systematically across patients, units, or institutions.

may recognize the same physiology only when sepsis or overt dehiscence forces intervention. A patient discharged rapidly after an apparently uncomplicated recovery may have a leak diagnosed elsewhere, while a patient kept in a high-acuity environment generates a denser trail of laboratory values, imaging, and clinical observations that increase the odds of recognition. The fairness question therefore does not begin with model training. It begins with the production of the label.

This immediately distinguishes AL prediction from many benchmark-driven machine learning tasks. The object being predicted is not observed uniformly, and the path to observation is itself conditioned by clinical workflow. A model can therefore learn at least three things at once: genuine physiologic vulnerability, institutional habits of surveillance, and the pathways through which some patients become more legible to the record than others. Those three forms of signal are not easily separable in retrospective structured data, and yet they have radically different fairness implications. A model that primarily learns physiologic vulnerability may support equitable intervention. A model that primarily learns which patients receive dense observation may reproduce inequity while appearing accurate [2]. A model that learns rescue pathways may perform well only inside the local clinical culture that generated its training data.

The difficulty is intensified by the practical uses to which AL models are proposed. Some models are imagined as preoperative stratifiers that inform counseling or optimization. Others are positioned intraoperatively to influence diversion or perfusion-related decisions. Many are effectively postoperative surveillance systems built from the first several hours or the first day of structured data. Each use case implies a different fairness problem. Preoperative prediction affects who receives preventive resources before the anastomosis exists. Postoperative prediction affects who receives intensified scrutiny after the surgical insult has already occurred. Intraoperative prediction affects action under time pressure, often in environments where the boundaries between human judg-

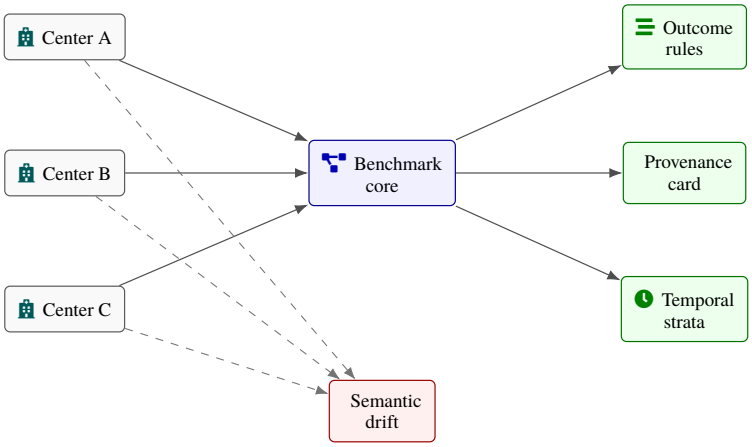

Figure 3: Benchmark governance for leak prediction should be anchored in multi-center aggregation but constrained by explicit metadata. The benchmark core is only trustworthy when outcome rules, data provenance, and temporal composition are documented alongside performance. The lower node highlights semantic drift, where variables retain the same name across sites but change meaning because workflow, surveillance, or documentation culture differs.

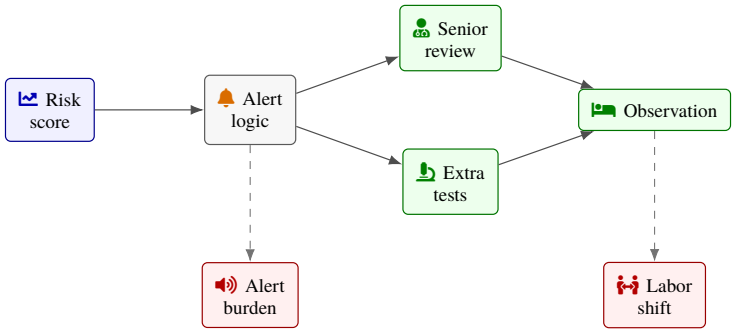

Figure 4: Operational fairness depends on how a score is converted into work. Alert logic can redistribute clinical attention by triggering review, testing, and prolonged observation, but that same pathway can also impose concentrated burden on specific teams or settings. A numerically acceptable score may still produce inequity if its workflow consequences are unevenly absorbable across units, staffing structures, or patient groups.

ment and algorithmic suggestion are most delicate. The same numerical score may therefore carry very different fairness implications depending on when it appears and what it is expected to trigger.

A further obstacle comes from the social structure of surgical data. Electronic health records are often discussed as if they were passive archives of patient state. In reality they are records of work. Their content depends on staffing, order sets, bed availability, surgeon preference, unit culture, documentation norms, and reimbursement logic. Missingness is rarely random in a clinically meaningful sense. A postoperative lactate may be absent because no one was worried enough to order it, or because the patient was discharged, or because the local pathway does not include it routinely, or because documentation was fragmented across systems. A ward patient and an ICU patient do not merely produce different values; they produce different possibilities for being known. In rare-event prediction that difference matters enormously. The denominator of patients with sufficiently rich data may already reflect a selective filtering process before any classifier is fit.

This paper approaches fairness in AL prediction by changing the level at which the problem is posed. Instead of asking which fairness metric should be optimized after a model is trained, it asks what kind of research pipeline is capable of generating fair knowledge in the first place [3]. That shift leads to a more demanding but also more clinically realistic perspective. Fairness becomes an epistemic property of the full modeling enterprise. It depends on whether the outcome definition corresponds to a coherent clinical phenomenon, whether subgroup variation is visible rather than averaged away, whether data provenance is transparent, whether external transport is taken se-

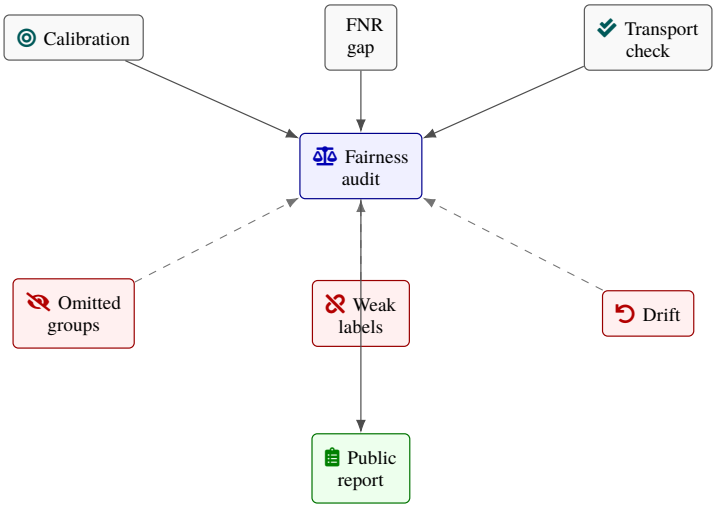

Figure 5: A reproducible audit for leak prediction should integrate classic model checks with benchmark-sensitive warnings. Calibration, false-negative asymmetry, and transport behavior form the positive evidentiary surface, while omitted groups, weak labels, and drift identify conditions under which fairness claims become fragile. Public reporting then becomes the mechanism that stabilizes these findings into accountable evidence rather than one-off internal evaluation.

Table 1: Key sources of fairness challenges in anastomotic leak prediction

Component	Issue	Fairness Implication
Outcome definition	Variable clinical criteria	Unequal event visibility
Data collection	Heterogeneous workflows	Biased representation
Missingness	Non-random absence	Skewed learning signal
Institutional practice	surveillance intensity	Differential detection rates

riously, whether implementation pathways are articulated, and whether post-deployment behavior is audited as part of the model rather than treated as external noise.

Such a perspective does not deny the value of conventional fairness metrics. Groupwise calibration, false negative asymmetry, thresholded recall, and related measures remain important. What it denies is their sufficiency. In AL prediction, a model can satisfy some familiar fairness criterion and still be unsuitable because the labels are unstable, the evaluation cohort excludes underdocumented patients, the benchmark privileges a single tertiary-care workflow, or the deployment plan channels additional attention toward populations already well served while leaving others with delayed recognition. The fairness of a leak predictor is therefore inseparable from the fairness of the benchmark and from the fairness of the workflow into which it will be inserted.

The sections that follow take up this broader view in a systematic way. The first body section argues that fairness in AL modeling is fundamentally an evidence problem, not only a classification problem. The next examines outcome definition and label construction as sites where inequity can be formalized and reproduced. The paper then turns to benchmark governance, institutional heterogeneity, and data provenance, emphasizing that a fairness claim tied to one center and one documentation culture remains weak. After that, the discussion shifts to clinical interface design, hidden redistribution of labor and alert burden, and the difference between a score that is numerically fair and a deployment that is operationally fair. The penultimate body sections develop a framework for reproducible auditing and for a forward research agenda organized around publicly defensible benchmarks, prospective implementation studies, and clinically accountable reporting. The conclusion returns to a simple but demanding proposition: equitable AL prediction is not achieved by tuning a model in isolation. It is achieved by constructing a research and deployment pathway in which what is counted, what is predicted, what is acted upon, and what is reviewed all remain open to technical scrutiny.

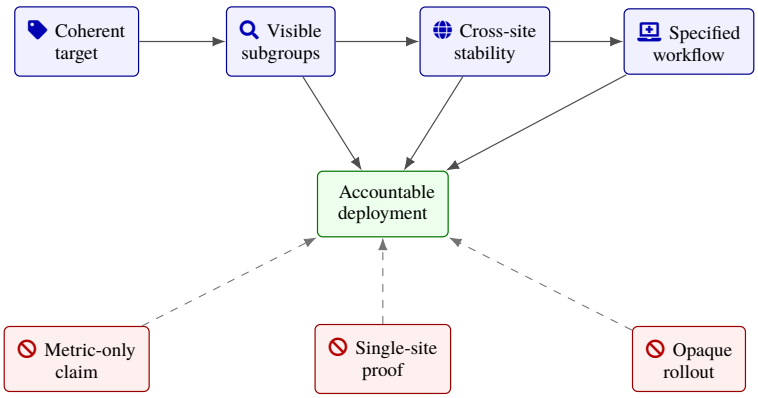

Figure 6: The layered evidentiary standard can be summarized as four positive conditions feeding an accountable deployment claim: the target must be clinically coherent, subgroup behavior must be inspectable, external stability must be demonstrated, and workflow integration must be declared in advance. Three common shortcuts are shown below as disqualifying counterexamples: relying only on post hoc metrics, treating one institution as sufficient proof, and deploying through an opaque operational pathway.

Table 2: Types of signals learned by predictive models

Signal Type	Description	Risk
Physiologic	True biological risk	Desired signal
Observational	Monitoring intensity	Encodes inequity
Workflow	Institutional processes	Poor generalization
Rescue pathways	Intervention patterns	Context dependence

2. Fairness as an Evidence Problem in Surgical Prediction

The usual grammar of fairness in machine learning assumes that the main epistemic work has already been done. A dataset exists, labels are treated as acceptable approximations of the underlying phenomenon, and the fairness task is to ensure that model behavior does not vary unjustifiably across groups. In AL prediction, that sequence is reversed. Before one can even ask whether a model behaves fairly, one must ask whether the evidence supporting the task is itself fair in how it makes the complication visible. This is why the fairness problem here is inseparable from study design [4]. The issue is not only whether a trained algorithm equalizes some error rate, but whether the research pipeline has decided, often implicitly, whose leaks count, whose records are legible enough to model, and which institutional realities are allowed to stand in for clinical truth.

This matters because leakage between knowledge production and fairness assessment is pervasive in surgical modeling. Retrospective datasets frequently arrive already filtered by operational constraints. A study cohort may exclude cases without complete laboratory windows, omit transfers lacking clean follow-up, or rely on coding schemes that perform differently in elective and emergent pathways. Each of those decisions shapes the population about whom the model can speak. When such exclusions are described merely as preprocessing, the fairness implications are hidden under technical routine. Yet preprocessing choices in AL studies are often tantamount to eligibility judgments about whose postoperative course is coherent enough to enter the benchmark. A benchmark that systematically favors richly documented tertiary-care patients while underrepresenting fragmented or lower-resource care trajectories can still produce excellent internal performance and still fail the populations that most need equitable surveillance.

The central technical point is that fairness cannot be appended to a benchmark whose representational structure is already distorted. If the data-generating process privileges certain hospitals, certain care units, certain procedure types, or certain documentation habits, then group-wise performance estimates computed downstream will inherit that distortion. Even sophisticated audit dashboards may simply quantify disparities within a population that has already been selectively rendered visible. In this sense, fairness analysis that begins only after cohort assembly is a late-stage intervention. It may diagnose some problems, but it cannot recover the fairness properties of a study design that made unequal visibility foundational.

Table 3: Outcome construction strategies and limitations

Method	Strength	Limitation
Diagnosis codes	Easy extraction	Documentation bias
Reoperation-based	High specificity	Misses mild cases
Chart review	Clinical richness	Resource intensive
Composite labels	Higher capture	Hides heterogeneity

Table 4: Benchmark design considerations

Aspect	Requirement	Risk if ignored
Data provenance	Transparent sourcing	Misinterpretation
Institution diversity	Multi-center inclusion	Poor transportability
Temporal coverage	Period stratification	Hidden drift
Cohort definition	Explicit exclusions	Selection bias

The evidence problem becomes sharper when one considers the distinction between event occurrence and event recognition. AL is not measured the same way that body temperature or serum sodium is measured. It is inferred. Recognition may depend on imaging, operative exploration, drainage, inflammatory progression, or clinician suspicion. These pathways are themselves influenced by local practice and by the social distribution of clinical attention. A model trained to predict recognized leak is therefore trained on an endpoint that partially reflects how the institution watches. If an institution watches some patients more intensely than others, the observed event landscape is already patterned by differential scrutiny. The fairness question then becomes double [5]. One must ask whether the model reproduces unfairness in the observed labels and whether the labels themselves are unfair approximations of the clinical phenomenon of interest.

This is where AL prediction departs from a simplistic meritocratic story of data abundance. More data do not automatically produce fairer models. If data accumulation intensifies around patients who are already monitored aggressively, the model may become more certain precisely where the institution is already attentive and less informative where surveillance is sparse. Greater data volume under unequal observation can sharpen rather than solve inequity. A fairness-aware evidence framework must therefore be sensitive not only to sample size and class balance but also to observational density, to the conditions under which postoperative information becomes available, and to the gap between latent complication and recorded endpoint.

An instructive implication follows. In AL research, the fairness status of a model cannot be read directly from whether sensitive variables are included. Excluding race, sex, age, insurance category, or hospital identity does not neutralize inequity when the outcome label, missingness pattern, and feature space continue to encode differential observation. Procedure type, postoperative bed location, measurement frequency, timing of laboratory collection, and reintervention history can all function as carriers of structured disparity. The evidentiary pipeline therefore needs a richer taxonomy than sensitive versus non-sensitive variables. It needs to distinguish variables that summarize biologically relevant surgical state, variables that mainly describe institutional workflow, and variables whose meaning shifts depending on the context in which they are measured. Without that taxonomy, fairness debates become trapped in the superficial question of whether demographic attributes appear in the design matrix.

An evidence-centered perspective also changes how one thinks about generalizability. External validation is often described as a robustness requirement. In the fairness context it is more than that. It is a test of whether the model's knowledge claims survive movement across institutions where the pattern of legibility changes. A score that performs similarly in multiple centers with different staffing structures, different coding conventions, different patient populations, and different postoperative pathways provides stronger evidence that it is learning something closer to clinically stable signal. A score that collapses outside its home institution may not simply be overfit in an abstract statistical sense. It may be overfit to a local regime of observation whose fairness properties cannot be presumed transferable.

The same logic applies to subgroup analysis [6]. There is a tendency in clinical machine learning to regard subgroup reporting as an optional refinement once an overall signal has been established. In AL prediction this attitude is particularly risky because aggregate perfor-

Table 5: Subgroup evaluation dimensions

Dimension	Example	Purpose
Demographic	Age, sex	Equity check
Clinical	Procedure type	Context sensitivity
Institutional	Hospital type	Transport analysis
Pathway	ICU vs ward	Workflow effects

Table 6: Deployment-related fairness factors

Factor	Description	Impact
Alert burden	False positives workload	Resource strain
Interface design	Score presentation	Interpretation bias
Override behavior	Clinician response	Uneven adoption
Capacity limits	Resource availability	Inequitable action

mance can conceal systematic underperformance in precisely the populations whose event recognition is already less consistent. A recent proof-of-concept paper made this absence visible by noting that existing AL models have not been examined for subgroup bias and by arguing that explicit bias analysis, chart-validated outcomes, and diverse multi-center validation are necessary elements of the next stage of research, an observation that places fairness at the level of evidentiary architecture rather than as a cosmetic add-on to model reporting Sadanandan (2024) [7].

Once fairness is understood as an evidence problem, several consequences follow. First, internal validation is ethically thin even when technically polished, because it cannot tell whether a model's fairness profile depends on local habits of surveillance and documentation. Second, outcome definition becomes a fairness decision rather than a neutral preprocessing step, because it governs whose complications become visible to the model. Third, missingness is not merely a statistical inconvenience; it is part of the institutional structure of knowability. Fourth, deployment cannot be separated from validation, because a model introduced into workflow changes how future evidence will be produced. Finally, fairness is no longer something a model possesses independently. It becomes a property of the relationship among data collection, label construction, model design, and clinical action.

This reframing is technically useful because it clarifies why conventional fairness debates often seem unsatisfying in the surgical setting. They are addressing a later slice of the problem than the one that determines whether the benchmark is informative. It is entirely possible to have a carefully calibrated, threshold-optimized classifier on top of an endpoint that poorly represents latent leak across settings and groups. In that case the model may be fair relative to the benchmark and unfair relative to the clinic. The task for AL research, then, is not simply to design a fair learner. It is to build a benchmark ecology in which fairness claims are worth making at all.

3. Outcome Definition, Label Construction, and the Politics of Clinical Visibility

Few elements of an AL study are more technically consequential and less conceptually neutral than the outcome definition. Researchers often speak of leak as if it were a naturally given endpoint, but in practice it is assembled. Coding rules, procedure logs, radiology reports, narrative adjudication, antibiotic exposure, drainage patterns, and reoperation decisions may each enter the label. The resulting endpoint is not merely a reflection of tissue failure. It is a negotiated artifact that compresses pathophysiology, institutional practice, and clinical interpretation into a binary indicator [8]. Because fairness depends on what the model is asked to predict, outcome construction is one of the primary sites at which inequity can be introduced, stabilized, or obscured.

The first difficulty is heterogeneity of clinical meaning. Some leaks are radiographically subtle, some produce fulminant sepsis, some are managed conservatively, and some are recognized only when a patient deteriorates dramatically. A label derived from reoperation privileges severe or intervention-requiring cases and may miss smaller events treated non-operatively. A label derived from diagnosis codes may capture institutional billing and documentation habits more than bedside phenomenology. A chart-reviewed endpoint may be clinically richer but still dependent on the adjudicator's interpretation of incomplete records. When studies compare model performance across such differently assembled labels, they often act

Table 7: Validation components for fair assessment

Component	Method	Importance
Calibration	Groupwise curves	Reliable risk estimates
Discrimination	AUC, sensitivity	Ranking ability
External validation	Multi-site testing	Generalizability
Threshold analysis	Workload estimation	Practical fairness

Table 8: Label design levels

Level	Focus	Limitation
Operational	Extractability	Low clinical meaning
Clinical	Phenotype accuracy	Ignores inequity
Equity-aware	Cross-group validity	Higher complexity

as though the underlying target were identical. It is not. The fairness implications of each label can differ sharply because the criteria for recognition are differentially sensitive to surveillance, procedure type, and group-specific clinical pathways.

This instability matters because fairness is frequently discussed in relation to error rates. Yet the concept of error presupposes a defensible truth standard. If a leak label is more sensitive in one subgroup because that subgroup undergoes more imaging or stays longer in monitored settings, then a model judged against that label may appear more accurate for reasons unrelated to physiologic discrimination. A high-performing model in this context may simply be excellent at forecasting which patients are likely to pass through institutionally favored diagnostic pathways. That is not useless information, but it should not be conflated with equitable clinical prediction.

There is a deeper epistemic issue here. Labels do not merely summarize events; they discipline the research imagination by telling the field what counts as the complication worth predicting. If the dominant AL benchmark defines leak through hospital-recognized postoperative events within a constrained window, then late, transferred, or externally recognized complications are pushed outside the target. This move may simplify modeling, but it also narrows the ethical horizon of the work. The model becomes a predictor of leaks that fit institutional capture rather than a predictor of the broader clinical risk of anastomotic failure. The populations most vulnerable to incomplete follow-up, fragmented care, or delayed recognition are then disadvantaged twice: first in the clinic and again in the research evidence meant to improve the clinic.

A fairness-aware approach to label construction therefore needs to distinguish at least three levels of endpoint design [9]. The first is operational convenience, where labels are chosen because they are extractable from structured fields. The second is clinical coherence, where labels correspond to an interpretable event phenotype. The third is equity sensitivity, where label design is examined for how it performs across settings and patient groups with different patterns of observation. Conventional model development tends to stop at the second level when it reaches it at all. AL prediction requires the third. A label can be clinically coherent in aggregate and still inequitable in how it becomes observable across subpopulations.

Chart review is often proposed as the solution. It improves specificity of interpretation and can recover events that structured codes miss. Yet chart review is not an automatic fairness repair. Its effectiveness depends on whose records are complete enough to adjudicate, whether reviewers have access to outside-hospital follow-up, how equivocal cases are resolved, and whether adjudicators apply the same thresholds across procedure types and institutions. Chart review is best understood as a more explicit and inspectable form of label construction, not as a guarantee of fairness. Its contribution is that it exposes the inferential process rather than hiding it in administrative abstraction. That visibility is valuable because it allows disagreement, audit, and refinement.

Another important design choice concerns temporal framing. AL prediction studies often define the outcome within a fixed postoperative horizon, but the fairness implications of horizon choice are seldom explored. A short in-hospital window favors early-recognized complications and may undercount events emerging after discharge. A longer horizon improves capture but increases dependence on follow-up infrastructure and cross-system linkage. In hospitals or populations where post-discharge contact is fragmented, long-horizon outcome ascertainment may itself become uneven. Horizon choice is therefore not merely clinical. It governs which care pathways the label can see and which patients fall out of view.

Table 9: Research priorities for equitable prediction

Priority	Goal	Expected Benefit
Prospective benchmarks	Better labels	Reduced bias
Multi-center studies	Diverse data	Fair generalization
Workflow integration	Defined actions	Practical equity
Audit standards	Reproducibility	Trustworthy results

Table 10: Summary of fairness as an epistemic property :contentReference[oaicite:0]index=0

Layer	Question	Evaluation Need
Data	Who is represented?	Cohort audit
Label	What counts as leak?	Outcome validation
Model	How predictions behave?	Subgroup metrics
Deployment	What actions follow?	Workflow analysis

The problem becomes even more complex when prediction is intended for different moments in the clinical workflow [10]. A preoperative model that uses a postoperative-defined leak label may be judged against an event partly shaped by rescue actions that occur after surgery. A postoperative surveillance model using the first twenty-four hours of data may predict whether a leak will become visible under current monitoring intensity, not whether a latent dehiscence process is already underway. These distinctions affect fairness because groups may differ not only in intrinsic vulnerability but also in access to the forms of rescue and scrutiny that transform latent pathology into recorded outcome. If the research target blurs these layers, fairness analysis will likewise blur them.

A technical discipline of label documentation is therefore essential. Every AL benchmark should specify what clinical state the label is intended to represent, what sources contribute to its ascertainment, what time window is used, how equivocal cases are handled, which events outside the primary institution are captured, and where missing follow-up is likely to be concentrated. Without such documentation, fairness claims remain underdescribed because the relationship between group membership and label visibility cannot be evaluated. Transparent label documentation is not a matter of stylistic thoroughness. It is an analytic requirement for understanding whether model disparities arise from prediction, from endpoint construction, or from the interaction between the two.

One should also resist the temptation to resolve label heterogeneity by collapsing multiple definitions into a single composite endpoint without examining how each component behaves across groups. Composite labels can improve apparent event capture while hiding clinically relevant variation in recognition pathways. A patient counted through reoperation is not equivalent, from an equity perspective, to a patient counted through a late diagnosis code or a drainage procedure. These reflect different thresholds of severity, different levels of access to imaging or intervention, and different institutional scripts for what becomes documented. A fairness-oriented benchmark should preserve, not erase, as much of this pathway information as feasible. Multi-label or tiered outcome design may be more technically demanding, but it preserves distinctions that are essential for evaluating how the model interacts with the real care process.

This has consequences for how model performance should be interpreted. A predictor that excels at identifying severe intervention-requiring leaks may be clinically valuable even if it underperforms on subtler presentations, but its fairness profile cannot be inferred from overall discrimination alone. It may preferentially benefit groups whose complications are already likely to escalate visibly. Conversely, a model that captures early inflammatory signatures may improve surveillance equity if paired with appropriate action pathways, even if its overall precision is modest. The key is that fairness depends not only on whether the score matches the assembled label, but on whether the assembled label corresponds to a clinically and institutionally defensible understanding of whom the system is designed to protect.

The politics of clinical visibility are therefore embedded in the technical life of AL prediction [11]. Labels decide which events exist for the model, which absences are treated as non-events, and which forms of delayed recognition are excluded from the benchmarked world. A field that wishes to build fair leak prediction systems has to face this directly. The work begins not with a fairness penalty during training, but with a disciplined account of what the

outcome means, how it is known, and whose postoperative reality it excludes when it is rendered into a dataset.

4. Benchmark Governance, Data Provenance, and Institutional Heterogeneity

Once the outcome has been specified, the next fairness challenge concerns the benchmark itself. Benchmarks are often treated as neutral substrates on which models compete, but in clinical prediction they are governance instruments. They establish what kinds of evidence count as sufficient, what forms of generalization matter, and what kinds of failure remain invisible. In AL prediction, benchmark governance is especially important because the field has been built largely on single-center or otherwise narrow datasets, each carrying its own institutional logic of care, documentation, and postoperative escalation. A fairness claim issued inside one such benchmark is not meaningless, but it is fragile. It says at most that the model behaves in a certain way within one ecology of practice.

Data provenance is the first component of this governance problem. Provenance refers not only to where the data came from but to how they moved through extraction, curation, preprocessing, exclusion, harmonization, and temporal framing before they became a modeling cohort. Provenance determines whether a variable is interpretable, whether missingness is meaningful, and whether subgroup comparisons have common semantic ground. In many AL studies, provenance details are summarized only briefly, often because retrospective extraction from EHR systems is messy. Yet fairness depends on this mess. A postoperative white blood cell count drawn under one laboratory schedule does not mean exactly the same thing as a white blood cell count drawn under another schedule. ICU admission can indicate severity in one institution, pathway routine in another, bed scarcity in a third, and documentation artifact in a fourth. A variable lacking provenance can still be predictive, but it is difficult to know whether its predictive power is clinically stable or institutionally local.

This problem becomes acute when benchmark datasets are celebrated for scale without equal attention to their clinical geography. Large tertiary centers are attractive because they produce rich structured records and enough rare events for machine learning. But they are also distinctive environments. They may concentrate critically ill patients, support more intensive surveillance, and generate denser postoperative traces than community hospitals or safety-net settings [12]. A model trained in such data can inherit a highly specific view of what postoperative legibility looks like. When fairness is assessed only within that environment, the benchmark may mistake local representational abundance for universal clinical adequacy. This is why scale alone is not a fairness credential. Diversity of institutional environments matters at least as much as sample size.

Benchmark governance should therefore require explicit characterization of the environments represented in the training and evaluation data. This includes hospital type, surgical volume, care-unit structure, baseline surveillance intensity, discharge pathways, coding conventions, and patient population differences that are clinically relevant for AL risk. Such information is rarely integrated into modeling tables with the seriousness it deserves. Yet without it, external validation has a tendency to become a ritual rather than a test. A model moved from one tertiary academic center to another may technically cross institutional boundaries while still remaining inside a narrow band of documentation cultures. Fairness under deployment demands a more ambitious notion of heterogeneity, one that includes variation in resource constraints and in the social organization of postoperative care.

The benchmark question also includes what kinds of patients are missing altogether. Community-hospital cases, lower-acuity pathways, fragmented follow-up, and care episodes with sparse structured documentation are often underrepresented or excluded. The resulting benchmark can then privilege a style of patient who is institutionally legible, continuously monitored, and densely coded. When subgroup fairness is later reported, it is reported within that already filtered universe. The field may then congratulate itself on equity among the patients most visible to the data infrastructure while remaining silent about those rendered peripheral by the benchmark design. A rigorous fairness framework has to treat benchmark omission itself as a form of bias, not as a background condition beyond critique.

One practical consequence is that AL research needs a stronger culture of benchmark documentation and stewardship. Public or semi-public benchmarks should not merely provide train-test splits. They should include extensive metadata describing the origin, exclusions, outcome rules, variable provenance, and known representational limitations of the dataset. They should state which patient pathways are likely overrepresented, which follow-up patterns are incomplete, and which subgroup analyses are statistically weak because the benchmark remains narrow. Without such stewardship, the benchmark becomes a silent authority [13]. Researchers optimize against it,

compare against it, and claim fairness within it, while the benchmark's own structure escapes serious technical scrutiny.

Institutional heterogeneity also raises the issue of semantic instability across features. Variables that seem standard are often context-dependent in practice. Albumin may reflect chronic nutritional status in one setting and acute resuscitation context in another. ICU admission may encode surgeon preference or postoperative bed availability as much as clinical concern. Drain placement may be rare in one hospital and routine in another, changing the meaning of drain-related measurements. These variations matter for fairness because models trained to rely on locally stable proxies can behave unpredictably when transferred. A benchmark governance framework attentive to fairness should therefore include feature cards or variable-level provenance statements that describe not only data type and missingness but also plausible sources of institutional semantic drift.

Another neglected issue is temporal provenance. Benchmarks often pool multiple years of data to gain enough events, especially for rare complications. But surgical pathways, perioperative optimization practices, documentation systems, and coding habits change over time. A pooled benchmark can thus compress different eras of clinical work into a single learning problem. If those changes disproportionately affect certain patient groups or procedure categories, temporal pooling can hide fairness-relevant drift. Benchmark governance should include temporal stratification and explicit disclosure of period effects, not merely because drift harms prediction, but because drift can reconfigure which patients are diagnosed, how quickly they are recognized, and which covariates remain informative.

A benchmark-centered fairness agenda also requires rethinking what constitutes a meaningful baseline. Comparing models solely by discrimination metrics encourages the idea that the benchmark's job is to rank patients. In AL prediction, the benchmark should also test whether a model's output is clinically interpretable, subgroup- transparent, and operationally mappable to action. Benchmarks that reward only AUC or similar aggregate measures invite overinvestment in ranking performance and underinvestment in the harder work of fair cohort construction and deployment specification. A stronger benchmark would require subgroup calibration displays, uncertainty intervals by clinically meaningful strata, documentation of missingness handling, and explicit articulation of intended use. Such requirements would not make the benchmark less technical. They would make it a more faithful instrument for translational evaluation [14].

This leads to an institutional point. Fairness in AL prediction cannot be fully delegated to individual model developers. It is partly a property of the shared standards the field sets for what counts as adequate evidence. Journals, benchmark maintainers, surgical collaboratives, and data custodians all participate in this governance whether they describe it in those terms or not. If a field routinely publishes models without subgroup reporting, without outcome provenance, without multi-center transport analysis, and without implementation pathways, then unfairness is not simply the failure of isolated researchers. It is an artifact of collective evidentiary norms. Conversely, if benchmark governance begins to require these elements, individual modeling work will shift accordingly.

A mature fairness framework for AL prediction must therefore operate at the level of benchmark design and stewardship. It must insist that data provenance be treated as first-order technical content, that institutional diversity be regarded as a fairness requirement rather than a luxury, and that missing pathways and omitted populations be recognized as substantive failures of representational adequacy. Only within such a governed benchmark ecology do subgroup metrics and model audits become persuasive. Otherwise fairness remains bound to the local assumptions of the dataset that happened to be available.

5. Clinical Interface, Alert Burden, and the Hidden Redistribution of Work

The fairness profile of an AL model does not end with the score. Once the model is inserted into a clinical pathway, it begins to redistribute attention, labor, and uncertainty. Those redistributions are rarely captured in traditional validation reports, yet they are among the main mechanisms through which a seemingly acceptable model can produce inequitable care. A risk score that enters a dashboard, triggers a page, prompts imaging, delays discharge, or demands senior review is not merely an informational object. It is a device that alters workflow. The fairness question therefore extends beyond statistical behavior and into the domain of operational allocation.

This matters because clinical systems do not respond uniformly to alerts. Different care teams have different thresholds for concern, different access to imaging, different ability to prolong observation, and different levels of trust in automated recommendations. A risk estimate that is operationally light in one setting may be burdensome in

another [15]. For some patients, an alert may yield earlier consultant attention and rapid diagnostics. For others, the same alert may yield only additional documentation, repeated but non-actionable blood tests, or prolonged anxiety without meaningful escalation. Fairness in deployment is therefore not just about who is flagged. It is about what the flag can do in the setting where it appears.

Alert burden is especially salient in rare-event prediction. Because AL is uncommon, even reasonably discriminative models may produce many false positives at clinically useful sensitivity levels. This creates a workload problem. Extra monitoring, extra imaging discussions, additional rounds of chart review, and delayed discharge decisions consume finite clinical energy. If those burdens are distributed unevenly across patient groups or hospital contexts, the model may create a new layer of inequity. One subgroup may bear disproportionate surveillance friction without commensurate benefit, while another may receive most of the actionable attention because the local system is better able to respond to alerts in that group. A model evaluated only through sensitivity and specificity provides no visibility into this hidden redistribution.

The interface through which the score is delivered matters as well. A risk estimate presented as a single opaque number may be interpreted differently from one accompanied by clinically legible supporting signals. If clinicians cannot understand why a patient has been flagged, they may rely on preexisting assumptions to decide whether to act. Those assumptions can themselves be patterned by procedure type, age, sex, comorbidity, or perceived reliability of the patient's postoperative narrative. In this way, a numerically uniform score can meet a socially uneven interpretive environment. Fairness then depends partly on explanation design. An interface that surfaces the clinical basis for concern may reduce arbitrary override. An interface that obscures the model's logic may invite selective trust and selective skepticism.

There is also a temporal dimension. AL surveillance is not a single decision but a sequence of decisions [16]. A patient may be low risk immediately after surgery, intermediate risk on the ward, and high risk at discharge planning. Another may receive a high early score because of transient physiologic instability that resolves. The workflow burden of monitoring those trajectories depends on how frequently predictions update and how strongly the system encourages action at each stage. Some groups may generate more uncertain trajectories because their measurements are sparser or because their postoperative course is less continuously visible. If the system treats uncertainty as low priority rather than as a call for more careful review, fairness can fail by indifference rather than by explicit disparity.

Human override makes this problem even more complex. Clinicians often and appropriately resist algorithmic automation in surgery, especially when recommendations conflict with bedside judgment. But override behavior is not neutral. Alerts may be taken more seriously in some service lines than others, by some clinicians more than others, and for some patient populations more than others. If an AL model is introduced without monitoring who acts on its outputs and under what conditions, fairness analysis remains incomplete. A model can be fair in the narrow sense of score behavior and unfair in the broader sense that its recommendations are systematically discounted for certain groups or in certain institutional niches. This is one reason implementation science cannot be separated from fairness science in this domain.

Resource limits intensify these concerns. Imaging, ICU beds, advanced nursing observation, and specialist review are not infinitely available. Under resource pressure, alert systems operate less like uniform aids and more like triage mechanisms. The risk score becomes a way of deciding who gets scarce attention. In such settings, ranking performance is not enough. One must examine whether access to scarce intervention is being allocated in a way that is clinically defensible and not simply self-reinforcing of existing institutional priorities. A model trained in a high-capacity setting may generate recommendations that cannot be actioned equitably in a lower-capacity one. Without this consideration, deployment can widen rather than reduce the distance between hospitals that already differ in complication rescue capability.

An often overlooked form of redistribution concerns clinician cognition [17]. High-risk surgical complications require interpretive effort. If a model increases the number of borderline cases requiring adjudication, that effort must come from somewhere. Teams with greater staffing flexibility may absorb it; overstretched teams may not. The same holds for patients and families. Additional testing, extended monitoring, and shifting discharge expectations may impose unequal emotional and logistical burdens, especially where language support, transport access, or social resources are limited. These are not peripheral humanistic concerns detached from technical fairness. They are part of the operational cost structure through which the model redistributes consequences.

Designing fair deployment therefore requires a more explicit mapping between score ranges and actions. What exactly happens when a patient is flagged? Does the alert trigger repeat labs, bedside review, imaging consideration,

delayed discharge, or surgical consultation? Who owns the response? How quickly must it occur? What happens when capacity is constrained? If these questions are not answered during development, then fairness cannot be evaluated meaningfully because the model's practical effect remains undefined. In AL prediction, undefined action pathways are particularly risky because the complication unfolds quickly enough that vague alerts can generate either alarm fatigue or missed opportunity.

A stronger design tradition would treat workflow mapping as part of model specification. It would examine how alert burden varies by subgroup, which populations experience the greatest ratio of surveillance burden to confirmed clinical benefit, and how override patterns modify those distributions. It would also recognize that some actions are lower burden than others. Additional bedside review is not equivalent to invasive testing or prolonged monitored admission. Fairness must therefore be action-tier specific rather than collapsed into a single alert concept. A group might reasonably receive more low-burden surveillance without inequity, while a pattern of disproportionate exposure to high-burden interventions would demand closer justification.

The hidden redistribution of work also creates feedback loops. Once clinicians know that a model tends to flag certain postoperative constellations, they may order more confirmatory tests in those situations. That changes future data and potentially the measured outcome distribution. Over time, the model may come to appear more or less accurate not because its underlying relation to physiology changed, but because the workflow adapted around it. If that adaptation differs across teams or patient groups, fairness can drift after deployment even when the code remains fixed [18]. Monitoring systems therefore need to track not only score performance but also the changing ecology of response.

In short, fairness in AL prediction cannot be assessed at the level of the score alone. It resides partly in the social and logistical pathway the score activates. A clinically acceptable model is one whose outputs can be translated into action without concentrating burden, skepticism, or opportunity in inequitable ways. That requires interface design, response ownership, capacity awareness, and override monitoring to be treated as core technical components of the modeling enterprise rather than as downstream operational details.

6. Reproducible Auditing, Reporting Discipline, and the Structure of Fair Validation

If fairness in AL prediction is to move beyond aspiration, the field needs an audit discipline that is both reproducible and clinically meaningful. Reproducibility here should not be understood narrowly as the ability to rerun a script. It also means that another team should be able to understand what outcome was constructed, how missingness was treated, which patients were excluded, what the intended use was, which subgroup definitions were prespecified, and how the model's outputs were expected to influence care. Without that broader reproducibility, fairness reporting remains fragile because the analytic object under audit is only partially disclosed.

The first requirement of such a discipline is that fairness reporting be integrated into the main validation narrative rather than relegated to supplemental tables. In rare-event surgical prediction, subgroup analysis is often treated as a statistical luxury because event counts become sparse once the population is partitioned. That concern is real, but it cannot justify silence. What it should justify is more careful uncertainty reporting, broader multicenter collaboration, and hierarchical or pooled estimation strategies that preserve visibility without pretending to a level of precision the data do not support. The response to subgroup sparsity cannot be to report only overall performance and infer equity by omission.

A reproducible fairness audit for AL prediction should therefore begin with a clear statement of the prediction moment, the action horizon, and the intended clinical role of the model. A preoperative counseling model, an intraoperative support model, and a postoperative surveillance model should not share the same validation template. Each has different fairness-relevant harms. The preoperative model may affect access to optimization or diversion. The postoperative model may affect access to monitoring and early rescue. The intraoperative model may affect irreversible procedural decisions. Reporting these distinctions is not a matter of prose polish [19]. It determines which subgroup performance characteristics matter most and how alert burden should be interpreted.

Next comes cohort accounting. Studies should specify, with numerical transparency, how many cases were excluded at each stage and which exclusion mechanisms may plausibly differ across patient groups or institutions. Exclusions for missing laboratory windows, incomplete coding, ambiguous operative identification, or absent follow-up are particularly important because they often remove

the very cases that challenge the benchmark's representational adequacy. Fair validation requires that readers can see whether the analyzed cohort is a filtered subset of clinically straightforward or richly documented trajectories. Without that view, good subgroup performance may simply be performance on a cleaned population.

Outcome auditing is equally essential. Researchers should report how the AL label was assembled, what sources were used, what adjudication standards applied, and where ascertainment is likely incomplete. When possible, studies should compare alternative label constructions rather than presenting one composite endpoint as definitive. If the model behaves very differently across these constructions, that instability is itself fairness-relevant because it suggests dependence on pathway-specific recognition rather than on robust clinical signal. In domains like AL, label pluralism is not merely methodological caution. It is a way to expose how much of the model's success relies on a particular way of making the complication visible.

Subgroup definition also deserves more discipline than it usually receives. Fairness reporting should include clinically meaningful strata such as procedure class, urgency, comorbidity burden, and care setting, alongside demographic and institutional categories. The reason is straightforward. In surgical prediction, equity concerns often arise at intersections between social position and procedural context. A model may behave acceptably across sex in aggregate while failing for patients undergoing specific low-volume or anatomically difficult operations. Another may perform evenly across hospitals overall while degrading sharply in the subgroup of transferred or underinsured patients. Prespecified subgroup plans help prevent opportunistic reporting and clarify whether the benchmark can support the claims being made.

Calibration and uncertainty reporting are particularly important. A model used for surveillance or risk communication should not be validated only through discrimination [20]. Groupwise calibration is needed because clinicians interpret risk values as meaningful estimates, not merely ranks. Yet calibration curves by subgroup are often omitted because they are visually messy or sample-limited. That omission is costly. In AL prediction, the upper risk tail is where action occurs, and miscalibration there can impose markedly different burdens across populations. Reproducible auditing should therefore include subgroup calibration displays, uncertainty intervals around them where possible, and explicit acknowledgment when data are too sparse for strong inference.

Threshold reporting should be equally disciplined. Too many studies select an operating point internally and present its sensitivity and specificity as if the choice were self-evident. In fair validation, threshold choice is a policy decision that should be justified in relation to clinical burden and actionability. What workload does the threshold generate? How many additional reviews or tests would it imply? Does that burden concentrate in particular groups? How robust is subgroup behavior when the threshold moves? These questions matter because fairness at one operating point does not guarantee fairness at another, and in low-prevalence complications thresholds often represent value judgments about missed events versus false alarms.

External validation should be treated not as a final ornamental step but as a central fairness audit. A model that performs well in one center and poorly in another is not simply unstable. It is providing unequal evidentiary support for the patients served by those institutions. Reproducible external validation should therefore document institutional context, feature semantics, outcome differences, and subgroup composition across sites rather than reducing transport to a single performance decrement. If subgroup disparities enlarge after transfer, the study should investigate whether the cause lies in label construction, variable meaning, workflow differences, or sample composition rather than merely noting performance loss.

A complete audit also needs post-deployment thinking even when prospective implementation has not yet occurred. Researchers should state how they expect the model to enter workflow, what kinds of actions it would trigger, and what new data-generating feedback loops it might create. This form of anticipatory reporting helps readers judge whether the fairness evidence presented actually bears on the proposed use. A model validated as a retrospective classifier may be inappropriate as a live alert if no account is given of alert burden, override, or capacity constraints. Fair validation is weakened when the model's imagined use drifts beyond the conditions under which it was evaluated.

Finally, reproducible fairness requires that methods and artifacts be inspectable. Code transparency matters, but so do variable definitions, extraction rules, and benchmark documentation [21]. In AL prediction, a great deal of apparent technical sophistication can hide quite simple but consequential study design choices. Releasing a model without clear extraction logic or subgroup definitions undermines fairness review because other teams cannot determine whether disparities reflect architecture, curation, or endpoint assembly. By contrast, even modest

models can contribute substantially to equitable progress if their entire evidentiary pipeline is inspectable and adaptable by others.

What emerges from this discussion is a validation philosophy in which fairness is not a separate appendix to an otherwise standard model report. It is an organizing principle for how the report is written, what is disclosed, and what kinds of uncertainty are permitted to remain visible. A fair AL model is not simply one that passes a metric threshold. It is one whose knowledge claims can be traced, challenged, reproduced, and situated within the clinical settings where it may eventually act.

7. A Research Agenda for Equitable AL Prediction Beyond Proof-of-Concept

The field of AL prediction has reached a point where additional proof-of-concept studies alone are unlikely to resolve its fairness problems. More retrospective classifiers trained on structurally similar datasets may refine performance estimates, but they will not by themselves answer whether the models distribute benefit equitably, whether the endpoints correspond to comparable clinical realities across institutions, or whether deployment can occur without hidden redistribution of burden. What is needed now is not simply model improvement but research architecture improvement.

The first priority is benchmark redesign around explicitly equitable evidentiary goals. Future benchmarks should be assembled prospectively whenever feasible, with outcome definitions that are preregistered, clinically interpretable, and subjected to adjudication procedures designed to reduce institution-specific ambiguity. Rather than treating chart review as an optional enhancement, studies should view adjudicated subsets as essential calibration instruments for understanding how much structured labels miss and for whom they miss it. Prospective benchmarking also permits explicit capture of follow-up outside the index institution, a major issue in postoperative complications that are often recognized after discharge or in transferred care settings.

The second priority is multi-center collaboration built around heterogeneity rather than around simple enlargement of sample size. Diverse participation should include hospitals with different resource profiles, documentation cultures, and patient populations, not only elite centers with compatible data infrastructures. The aim is not merely to stress-test generalization in an abstract sense. It is to ensure that the model is exposed during development to multiple regimes of clinical visibility. When the field builds only from data-rich environments, it risks producing models that are technically sophisticated but normatively provincial. Fairness is better served when the training and validation ecology reflects the plural reality of surgical care.

A third priority concerns variable strategy [22]. The field should move away from unexamined accumulation of any structured variable that improves retrospective discrimination and toward a more deliberate taxonomy of feature meaning. Variables should be categorized according to whether they primarily reflect baseline physiology, procedural exposure, institutional workflow, observation intensity, or rescue pathway. This taxonomy would not prohibit any category a priori. Instead, it would force researchers to justify how each type of variable is expected to behave under transfer and what fairness risks it carries. Features tightly bound to local workflow may still be useful, but their use should be transparent and accompanied by sensitivity analyses that reveal how much they drive subgroup and site-specific performance.

The fourth priority is prospective implementation research. Fairness in AL prediction cannot be established only through retrospective benchmarking because the model's use changes the very processes that generate future data. Studies therefore need implementation phases in which alert burden, clinician response, time-to-detection, and action uptake are measured directly. Randomized or quasi-experimental designs would be particularly valuable, not only to determine whether model-guided surveillance improves outcomes, but also to discover whether those improvements are distributed evenly. A model that reduces time-to-detection in well-resourced units while leaving other pathways unchanged may improve overall outcomes and still exacerbate inequity. Only prospective evaluation can expose that pattern convincingly.

Fifth, the field needs a stronger culture of negative and partial results. In rare-event clinical prediction there is a strong incentive to foreground the best aggregate performance and minimize the interpretive weight of subgroup instability, uncertain labels, or burdensome thresholds. That tendency is understandable but counterproductive. Fairness advances when studies openly report where the model fails, which subgroups remain uncertain, and which deployment scenarios appear operationally unrealistic. A mature literature should be able to say that a model is promising for one use, premature for another, and currently under-evidenced for a third. Such specificity is more useful than generalized optimism.

A sixth priority concerns institutional accountability. Hospitals and surgical programs adopting AL models should not rely exclusively on vendor or research-team assurances. They need local review structures capable of asking whether the benchmark resembles their own patient population, whether the proposed output fits existing workflow, and how subgroup performance will be monitored after deployment. This is especially important when models travel from academic development settings into hospitals with different staffing and resource constraints [23]. Equity requires not only better model development but also better institutional procurement and governance practices.

Seventh, educational culture within clinical AI needs adjustment. Fairness is still too often framed as either a narrow legal-compliance issue or as a late-stage ethical commentary. In AL prediction, it should be taught as part of core methodological competence. Surgeons, data scientists, quality leaders, and informaticians working in this area need a shared vocabulary for discussing outcome validity, cohort representativeness, benchmark limitation, and workflow burden. When fairness is integrated into ordinary technical conversation, it becomes less likely that teams will mistake high discrimination in a narrow benchmark for readiness for equitable clinical use.

An eighth priority is public methodological infrastructure. Shared reporting templates, variable provenance schemas, outcome documentation standards, and subgroup audit checklists would make it easier for researchers to produce comparable and inspectable evidence. These need not constrain creativity. On the contrary, they would allow innovation to be evaluated against a clearer background standard. In a field where heterogeneity of methods currently impedes cumulative learning, such infrastructure could help distinguish genuinely new advances from repeated rediscovery under slightly different retrospective conditions.

Ninth, the field should develop a more precise language for fairness trade-offs specific to surgery. Discussions often default to abstract debates over parity versus accuracy, but the concrete questions in AL prediction are more practical. Which errors delay rescue? Which errors consume scarce resources without benefit? Which variables encode legitimate operative risk and which encode unequal observation? Which deployment environments can act on the score and which cannot? A technical language tied to these realities would improve both model design and clinical interpretation. It would also reduce the temptation to import generic fairness slogans that fit poorly with the clinical stakes of postoperative complication management.

The final element of the research agenda is humility about what current models know. Proof-of-concept has value. Demonstrating that structured EHR data contain meaningful signal for AL prediction is an important step. But a field reaches maturity not when it declares the problem solved, but when it knows what remains unmeasured. In the present case, those unmeasured elements include the full range of outcome ascertainment pathways, the distribution of subgroup benefit under deployment, the behavior of models in resource-constrained environments, and the interaction between alerts and human judgment. A fair research agenda does not treat these as marginal complications to be cleaned up later [24]. It builds them into the central design of the next generation of studies.

What is at stake is not only the quality of individual models. It is the shape of evidence that will define acceptable surgical AI over the coming years. If AL prediction continues to advance mainly through isolated retrospective benchmarks, fairness will remain episodic and underpowered. If the field instead invests in governed benchmarks, adjudicated outcomes, transport-centered validation, and implementation-aware audits, it can begin to produce models whose claims are more than locally persuasive. That transition is not automatic, but it is technically feasible, and it is the level at which meaningful equity work in this area now needs to occur.

8. Conclusion

Fairness in AL prediction is often discussed as though it were a secondary constraint imposed on a basically settled modeling problem. The analysis here suggests the opposite. In this domain, fairness helps determine whether the modeling problem has been properly specified at all. A leak predictor is not merely a classifier trained on postoperative data. It is a claim about what counts as a leak, whose leaks are visible to the record, which clinical environments are represented in the benchmark, what kind of action the score is meant to support, and how that action will redistribute attention once the model enters care. If those layers remain underdescribed, then even sophisticated fairness metrics rest on weak foundations.

The central argument of this paper has been that equitable AL prediction must be understood as an epistemic achievement rather than as the output of a single algorithmic correction. Outcome construction, cohort assembly, benchmark stewardship, subgroup visibility, institutional heterogeneity, interface design, and implementation monitoring all shape whether a model behaves fairly in practice. This broader view does not replace statistical fair-

ness analysis. It gives that analysis a setting in which its conclusions can carry clinical meaning. Without such a setting, a model may appear fair relative to the benchmark while remaining unfair relative to the actual distribution of postoperative risk and recognition.

Several implications follow from this position. Outcome labels need to be treated as contested and constructed, not as self-evident truths. Benchmark design needs to foreground provenance, exclusions, and missing pathways rather than treating them as background technicalities. External validation needs to be understood as a fairness requirement because transport across heterogeneous institutions is one of the few ways to test whether the model is learning stable clinical signal rather than local habits of observation [25]. Workflow mapping and alert burden need to be evaluated because a score that is acceptable in isolation can become inequitable when translated into action under unequal capacity. Reproducible auditing needs to include subgroup uncertainty, threshold consequences, and deployment assumptions rather than limiting itself to pooled performance.

A further implication is methodological restraint. In a field shaped by rare events and imperfect labels, there is always pressure to let any available signal count as progress. That pressure is understandable, but it should not erase the distinction between proof-of-concept and clinically accountable evidence. Fairness requires the field to keep that distinction visible. A model can be technically interesting, statistically promising, and still underprepared for equitable deployment. Saying so is not a rejection of innovation. It is a way of preserving the link between predictive ambition and clinical responsibility.

This perspective also reshapes how progress should be measured. The next important advances in AL prediction may not come from a dramatic jump in aggregate discrimination. They may come from better outcome adjudication, stronger multi-center benchmarks, clearer variable provenance, more transparent subgroup auditing, more realistic implementation studies, and better governance of how scores enter workflow. Those developments are less marketable than a headline metric, but they are more likely to produce systems that deserve trust across diverse patient populations and institutional contexts.

In practical terms, the future of equitable AL prediction depends on whether the field is willing to make its evidentiary scaffolding explicit. Models should be built in a way that allows others to inspect what the label means, who the cohort excludes, how subgroup uncertainty behaves, what the score is supposed to trigger, and how the institution will know if the system begins to drift after deployment. A prediction tool that cannot answer those questions may still be computationally impressive, but it is not yet a fair clinical instrument.

For that reason, fairness in AL prediction should not be imagined as an optional ethical refinement added once predictive power is adequate. It is part of the discipline required to produce knowledge that can move responsibly from retrospective signal extraction into real surgical care. When fairness is treated at that level, the research task becomes more demanding, but also more honest. The result is not a simpler model-development story. It is a better one [26].

References

[1] L. J. Harris, B. Phillips, P. J. Maxwell, G. A. Isenberg, and S. D. Goldstein, "Outcomes of low anterior resection anastomotic leak after preoperative chemoradiation therapy for rectal cancer.," *The American surgeon*, vol. 76, pp. 747–751, 7 2010.

[2] G. Mazzarella, E. M. Muttillo, B. Picardi, S. Rossi, and I. A. Muttillo, "Complete mesocolic excision and d3 lymphadenectomy with central vascular ligation in right-sided colon cancer: a systematic review of postoperative outcomes, tumor recurrence and overall survival," *Surgical endoscopy*, vol. 35, pp. 4945–4955, 5 2021.

[3] Y. Mazni, A. Syafiuddin, and A. S. Putranto, "Intraoperative pancreatic assessment in pancreaticoduodenectomy the correlation with pancreatic fistula formation," *The New Ropanasuri : Journal of Surgery*, vol. 5, pp. 12–15, 6 2020.

[4] K. Kumagai, I. Rouvelas, A. Ernberg, S. Persson, A. Analatos, D. Mariosa, M. Lindblad, M. Nilsson, W. Ye, L. Lundell, and J. A. Tsai, "A systematic review and meta-analysis comparing partial stomach partitioning gastrojejunostomy versus conventional gastrojejunostomy for malignant gastroduodenal obstruction," *Langenbeck's archives of surgery*, vol. 401, pp. 777–785, 6 2016.

[5] C. A. D. Pasqual, V. Mengardo, F. Tomba, A. Veltri, M. Sacco, S. Giacopuzzi, J. Weindelmayer, and G. de Manzoni, "Effectiveness of endoscopic vacuum therapy as rescue treatment in refractory leaks after gastro-esophageal surgery," *Updates in surgery*, vol. 73, pp. 607–614, 11 2020.

[6] M. Fornasiero, G. Geropoulos, K. S. Kechagias, K. Psarras, K. K. Triantafyllidis, P. Giannos,

G. Koimtzis, N. A. Petrou, J. Lucocq, C. Kontovounisios, and D. Giannis, "Anastomotic leak in ovarian cancer cytoreduction surgery: A systematic review and meta-analysis.," *Cancers*, vol. 14, pp. 5464–5464, 11 2022.

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

[8] M. A. ÇAPARLAR, Şeref DOKCU, and S. DEMİRCİ, "Clinicopathological characteristics and management of patients with early readmission to our surgical oncology clinic," *The European Research Journal*, vol. 8, pp. 710–715, 9 2022.

[9] K. Knight, P. G. Horgan, D. C. McMillan, C. S. Roxburgh, and J. H. Park, "The relationship between aortic calcification and anastomotic leak following gastrointestinal resection: a systematic review," *International journal of surgery (London, England)*, vol. 73, pp. 42–49, 11 2019.

[10] A. S. Soares, N. T. Clancy, S. Bano, I. Raza, M. Diana, L. B. Lovat, D. Stoyanov, and M. Chand, "Interobserver variability in the assessment of fluorescence angiography in the colon.," *Surgical innovation*, vol. 30, pp. 45–49, 11 2022.

[11] M. C. Hernandez, C. Chen, C. S. Carlin, N. S. Seth, B. Yuh, Z. Eftekhari, A. Nguyen, and L. L. Lai, "Machine learning and postoperative complications: Determining risk before surgery in patients with cancer.," *Journal of Clinical Oncology*, vol. 41, pp. 6575–6575, 6 2023.

[12] X. Bonet, M. C. Moschovas, F. F. Onol, K. R. S. Bhat, T. Rogers, G. Ogaya-Pinies, B. Rocco, M. C. Sighinolfi, T. L. Woodlief, F. Vigués, and V. R. Patel, "The surgical learning curve for salvage robot-assisted radical prostatectomy: a prospective single-surgeon study.," *Minerva urology and nephrology*, vol. 73, pp. 600–609, 12 2020.

[13] F. Klevebro, M. Konradsson, S. Han, J. Luttikhold, M. Nilsson, M. Lindblad, M. Andersson, and D. E. Low, "Eras guidelines-driven upper gastrointestinal contrast study after esophagectomy can detect delayed gastric conduit emptying and improve outcomes.," *Surgical endoscopy*, vol. 37, pp. 1838–1845, 10 2022.

[14] S. H. Emile, H. Gilshtein, and S. D. Wexner, "Outcome of ileal pouch-anal anastomosis in patients with indeterminate colitis: A systematic review and meta-analysis.," *Journal of Crohn's & colitis*, vol. 14, pp. 1010–1020, 1 2020.

[15] A. L. Shada, L. H. Rosenberger, M. J. Mentrikoski, M. A. Silva, S. H. Feldman, and D. E. Kleiner, "Endoluminal negative-pressure therapy for preventing rectal anastomotic leaks: A pilot study in a pig model," *Surgical infections*, vol. 15, pp. 123–130, 1 2014.

[16] I. Kamaleddine, M. Popova, A. Alwali, and C. Schafmayer, "Endoscopic vacuum therapy for treating an esophago-pulmonary fistula after esophagectomy: A case report and review of the literature.," 2 2023.

[17] L.-H. Wang, F. Fang, C.-M. Lu, D. Wang, P. Li, and P. Fu, "Safety of fast-track rehabilitation after gastrointestinal surgery: systematic review and meta-analysis.," *World journal of gastroenterology*, vol. 20, pp. 15423–15439, 11 2014.

[18] H. Majhi, "Hypoalbuminemia is an important risk factor for surgical wound healing," *Journal of Medical Science And clinical Research*, vol. 7, 11 2019.

[19] L. Huang, Y. Yin, Y. Liao, J. Liu, K. Zhu, X. Yuan, L. Xue, and H. Pan, "Risk factors for postoperative urinary retention in patients undergoing colorectal surgery: a systematic review and meta-analysis.," *International journal of colorectal disease*, vol. 37, pp. 2409–2420, 11 2022.

[20] 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.

[21] A. Wolthuis, F. Penninckx, S. Fieuws, and A. D'Hoore, "Outcomes for case-matched single-port colectomy are comparable with conventional laparoscopic colectomy.," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 14, pp. 634–641, 3 2012.

[22] S. DHARMAYAN, "Impact of major urological complications on graft outcomes after renal transplantation," 2 2017.

[23] M. Pölcher, M. Eckhardt, C. Coch, M. Wolfgarten, K. Kübler, G. Hartmann, W. Kuhn, and C. Rudlowski, "Sorafenib in combination with carboplatin and paclitaxel as neoadjuvant chemotherapy in patients with advanced ovarian cancer.," *Cancer chemotherapy and pharmacology*, vol. 66, pp. 203–207, 3 2010.

[24] A. Elnahas, D. R. Urbach, G. Lebovic, M. Mamdani, A. Okrainec, F. A. Quereshy, and T. Jackson, "The effect of mechanical bowel preparation on anastomotic leaks in elective left-sided colorectal resections.," *American journal of surgery*, vol. 210, pp. 793–798, 6 2015.

[25] A. Mathew, G. Kaushal, D. Ramachandra, N. R. Rakesh, and P. Dhar, "Staple line dehiscence in gastric conduit following esophagectomy: A complication conspicuously missing a mention it deserves.," 1 2022.

[26] S. Kusamura, "Peer review report for: Hyperthermic intraperitoneal chemoperfusion with high dose oxaliplatin: Influence of perfusion temperature on postoperative outcome and survival [version 2; peer review: 1 approved, 2 approved with reservations]," 11 2015.