External Validation, Domain Shift, and Institutional Non-Exchangeability as the Primary Translational Constraint in Clinical Predictive Modeling

Nazri Hakim[1] AND Farid Azman[2]

[1] Department of Computer Science, Universiti Sultan Zainal Abidin, Jalan Gong Badak, 21300 Kuala Nerus, Terengganu, Malaysia
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, Bukit Ilmu, 18500 Machang, Kelantan, Malaysia

ABSTRACT

Clinical prediction research has entered a period in which the technical capacity to generate risk models is expanding faster than the capacity to establish whether those models will remain useful outside their development setting. Hospitals now produce extensive streams of perioperative, physiological, administrative, imaging, and laboratory data, and these data support increasingly sophisticated forms of statistical learning. Yet the environments in which such models are expected to operate are not interchangeable. Institutions differ in referral pathways, measurement intensity, coding practices, escalation thresholds, postoperative surveillance, and treatment timing. As a result, a model that appears accurate in retrospective development may fail to preserve calibration, ranking stability, or decision usefulness when transported to another clinical context. This paper develops the claim that generalizability and external validation should be treated as the principal translational problem rather than as a secondary reporting requirement. The analysis frames deployment as an exercise in estimation under institutional non-exchangeability, where the target quantity is not merely source-domain prediction loss but expected performance under heterogeneous, evolving, and intervention-sensitive environments. A formal account of transport failure is presented using multi-domain risk functionals, spectral representations of institutional variation, and decision-theoretic threshold analysis. The paper further examines how selection mechanisms, missingness structure, measurement heterogeneity, temporal drift, and intervention feedback destabilize apparently strong predictors. It then proposes a transport-first validation architecture centered on multi-site estimation, domain-wise calibration analysis, uncertainty quantification, and prospective policy testing. The broader conclusion is that translational success depends less on extracting maximal local fit than on demonstrating that model behavior remains coherent under the forms of variation that characterize real clinical systems.

I. Introduction

The contemporary literature on clinical prediction contains a revealing asymmetry [1] [2]. Considerable methodological effort is devoted to discovering predictive signal, refining model architectures, tuning hyperparameters, and increasing discrimination in retrospective cohorts, while substantially less effort is devoted to the question that determines whether any of those gains are clinically consequential: will the model remain reliable when it enters an environment that did not generate the data on which it was trained? The answer is often assumed rather than established [3]. A development cohort is treated as though it were an adequate surrogate for the broader clinical world, internal resampling is treated as though it were a test of deployment readiness [4], and an attractive retrospective performance profile is treated as though it were evidence of translational maturity. None of these inferences is logically secure. Clinical environments are not interchangeable containers of the same patient distribution. They are historically contingent systems whose workflows, technologies, thresholds for intervention, and outcome ascertainment practices actively shape the data from which models learn [5].

This problem becomes clearer when prediction is understood not merely as statistical approximation but as an attempt to estimate a quantity that will remain meaningful under deployment. A model trained on one institution's data estimates relations produced jointly by patient biology, measurement practice, treatment timing, and recordkeeping

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

conventions in that institution. Even when the clinical phenomenon is biologically stable, the recorded manifestation of that phenomenon may not be. A laboratory biomarker ordered liberally in one center and selectively in another carries different observational meaning. A postoperative complication documented through active chart review in one system and administrative coding in another is not the same target in any strict empirical sense. The central translational question is therefore not whether a model predicts well in the source environment, but whether the relation between recorded predictors and recorded outcomes survives the move into another environment with different observational and therapeutic structure.

The tendency to understate this difficulty is visible across many domains of clinical modeling. Predictive studies are often organized around the logic of discovery: identify a target event, assemble a retrospective cohort, fit candidate models, compare internally validated metrics, and present the highest-performing approach as a promising basis for implementation. Yet translation imposes a different logic. Once the model is expected to guide action in new settings, the key issues become calibration across domains, threshold stability, operational burden, and vulnerability to shifting case mix and treatment policy. A predictor that gains a modest increment in source-domain discrimination at the cost of greater dependence on local data idiosyncrasies may be less useful, not more useful, for real deployment. The dominant failure mode is frequently not insufficient complexity but insufficient transportability.

One surgical modeling review captures this broader pattern in a concrete way: despite a large number of proposed predictive studies, only a small minority underwent external validation, leaving reproducibility and deployment relevance underdeveloped even where apparent retrospective performance was acceptable Sadanandan (2024) [6]. The specific empirical context of that observation need not constrain its conceptual significance. It points to a more general condition of the field. The main obstacle to translation is not the absence of candidate algorithms, but the limited evidence that those algorithms estimate clinically stable relationships rather than local regularities tied to one setting's observational regime. In that sense, external validation is not a ceremonial supplement to model development. It is the principal empirical test of whether the model addresses a question that matters beyond the place in which it was trained.

This paper develops that claim in a different register from most discussions of predictive validity. Instead of treating generalizability as a property to be appended to a completed model, the analysis treats it as the organizing problem from the outset. The core idea is that clinical deployment should be framed as estimation under institutional non-exchangeability. Institutions differ not only in who is treated, but in how states are observed, how interventions respond to emerging risk, how labels are constructed, and how care processes evolve over time. These differences induce domain shift in ways that are not exhausted by familiar notions of prevalence change or covariate imbalance. They alter the meaning of both predictors and outcomes. Under such conditions, retrospective success in a source domain provides only partial evidence about future usefulness.

The argument proceeds in six stages. The next section characterizes institutional non-exchangeability and explains why a hospital or health system should be treated as a distinct data-generating mechanism rather than a nuisance grouping variable. The following section develops a statistical geometry of transport failure, emphasizing how source-specific representations can align with unstable institutional structure rather than clinically persistent signal. The paper then reframes external validation as an estimation problem under domain uncertainty and discusses what exactly should be estimated when multiple target settings are plausible. A later section examines temporal drift and intervention feedback, showing why deployment changes the very process the model aims to describe. The final analytic sections outline transport-first development principles and examine how failures of generalization propagate into operational burden, threshold misalignment, and decision regret [7]. The conclusion is intentionally moderate. It does not imply that all models are nontransportable or that local prediction is without value. It argues instead that the main translational barrier arises because evidentiary practices remain too local for the nonlocal claims models are asked to support.

II. Institutional Non-Exchangeability and the Clinical Data-Generating Process

The language of model development often encourages a misleading simplification: a dataset is treated as a sample from some population of patients, and differences between institutions are regarded as differences in who happened to be sampled. That view is inadequate for healthcare prediction because institutions do not merely sample patients from a common background distribution. They participate in generating the observations that enter the model. Referral structures determine which cases arrive. Staffing patterns determine how often patients are reassessed. laboratory ordering conventions determine what is measured and when. Information systems determine how events are encoded. Escalation pathways determine whether impending deterioration is recorded before or after intervention. A hospital is therefore not just a label attached to otherwise comparable cases. It is a mechanism that shapes both the observable covariates and the observed outcome process.

This point is important because exchangeability assumptions are often invoked implicitly in pooled multi-center studies. If observations were exchangeable across institutions after conditioning on a standard set of patient features, then a pooled model might estimate a broadly valid conditional risk surface. In practice, however, institutions differ in unrecorded ways that systematically affect both predictor semantics and

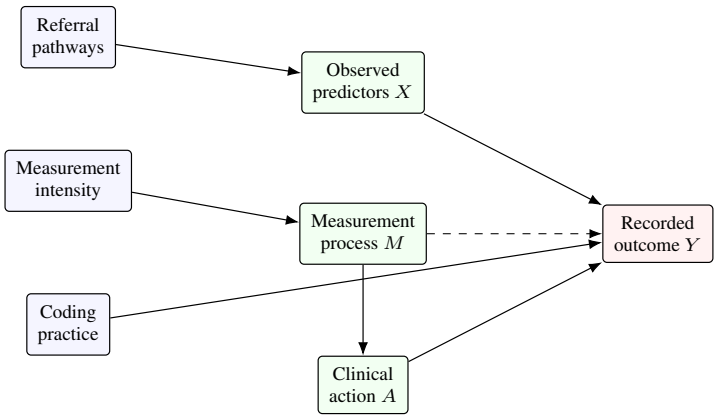

FIGURE 1: Institutional non-exchangeability viewed as a structured data-generating mechanism. The diagram emphasizes that hospitals do not merely sample patients; they shape what is measured, how risk is observed, when action occurs, and how the final label is recorded. Transport failure follows when a model trained on one factorization of predictors, measurement, action, and outcome is moved into another institutional process with different observational and therapeutic structure.

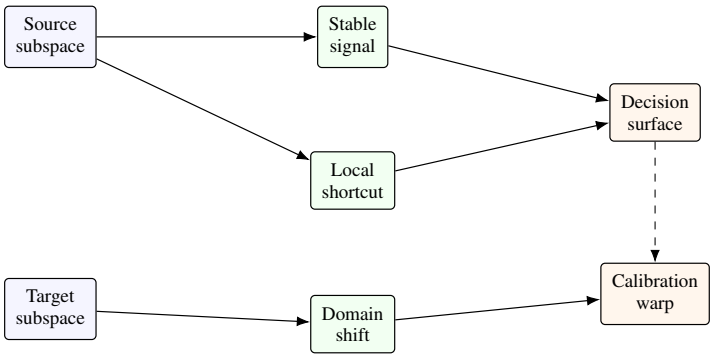

FIGURE 2: Geometric view of transport failure. During training, optimization can align the decision surface with a mixture of stable clinical structure and unstable local shortcuts. When the target domain rotates or distorts the predictor geometry, ranking may survive partially while calibration degrades, because identical score values no longer correspond to the same neighborhoods in the original feature space.

TABLE 1: Sources of Domain Shift in Clinical Prediction

Factor	Mechanism	Example	Impact
Referral patterns	Case selection bias	Tertiary centers	Prevalence shift
Measurement intensity	Observation frequency	ICU vs ward labs	Feature distortion
Coding practices	Label construction	Admin vs chart review	Outcome inconsistency
Treatment timing	Intervention policy	Early escalation	Altered risk relation

outcome realization. A postoperative fever recorded in a ward with aggressive early imaging may lead rapidly to diagnosis and intervention, while an apparently identical fever in another ward may be monitored longer before workup. The same observed pattern is therefore embedded in different treatment trajectories and different risks of recorded adverse outcome [8]. A model trained in the first setting learns from a relation partly stabilized by one rescue policy; when applied in the second, it is implicitly asked to operate under another.

It is useful to formalize this intuition by treating each domain as a structured data-generating process. Let D denote institutional domain, X observed predictors, M the measurement process, A clinical action, and Y the outcome label. Then the joint law can be written as

$$P_d(X, M, A, Y) = P_d(Y \mid X, M, A)\, P_d(A \mid X, M) \times P_d(M \mid X)\, P_d(X). \quad (1)$$

This factorization is not intended as a complete causal model, but it clarifies that observed prediction targets are jointly shaped by baseline clinical state, measurement intensity, and intervention behavior. Differences between domains may occur in any factor. An institution with denser monitoring changes $P_d(M \mid X)$. An institution with lower thresholds

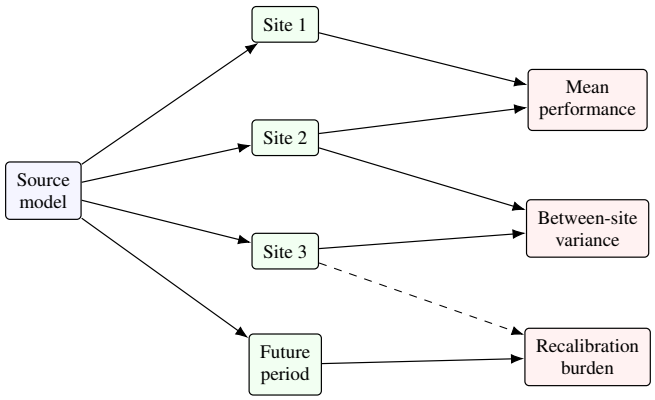

FIGURE 3: External validation framed as estimation under domain uncertainty. A single independent cohort estimates only one point on the deployment landscape, whereas multi-site and multi-period validation expose the quantities that matter translationally: average external performance, heterogeneity across plausible target domains, and the amount of recalibration required before the score becomes operationally interpretable.

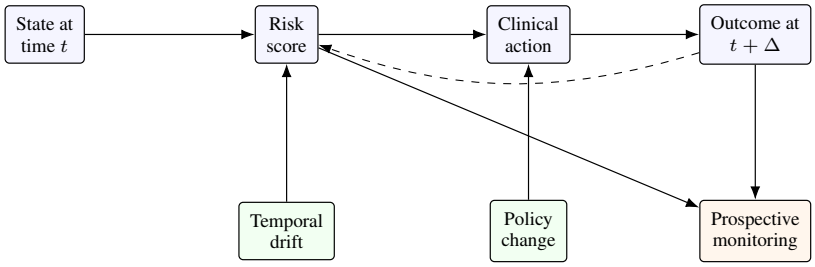

FIGURE 4: Temporal drift and intervention feedback. The model is inserted into a system that reacts to its own outputs, so predicted risk is not detached from treatment. Protocol updates, measurement drift, and policy response can all change the observed score–outcome relation over time, which is why live surveillance of calibration, workload, and action pathways is necessary after initial validation.

TABLE 2: Components of Institutional Data-Generating Process

Component	Symbol	Description	Variation Source
Predictors	X	Observed features	Measurement systems
Measurement	M	Observation process	Monitoring intensity
Action	A	Clinical intervention	Policy differences
Outcome	Y	Recorded event	Labeling practices

for escalation changes $P_d(A \mid X, M)$. An institution using a different adjudication protocol changes the effective construction of Y. A model developed under one such factorization cannot be presumed to estimate the same object in another domain merely because some variables share names and units.

Selection operates at the front end of this process. Many development cohorts emerge from tertiary referral settings, highly monitored units, or integrated health systems with unusually complete records. Such environments overrepresent certain disease severities, procedural complexities, or follow-up pathways. This creates a misleading impression of comprehensiveness. In truth, the model is learning from a carefully filtered region of the clinical state space, one in which the set of observable trajectories is itself conditioned by institutional capabilities and practice style. When the model is moved elsewhere, the problem is not simply that the new patients differ. It is also that the trajectories through which their risk becomes visible differ. A feature that carried predictive power in a dense-monitoring setting may lose that power when measurement intervals widen, not because the underlying physiology changed, but because the observational channel changed.

Missingness intensifies this problem because it is rarely random in operational healthcare data [9]. Whether a test is ordered, repeated, or omitted depends on local norms, clinician concern, patient access, and resource availability. The missingness pattern can therefore act as a compressed representation of institutional response. One site may omit a laboratory value mainly for low-risk patients, while another omits it because the test is unavailable overnight or restricted to certain service lines. In the first setting, absence of the

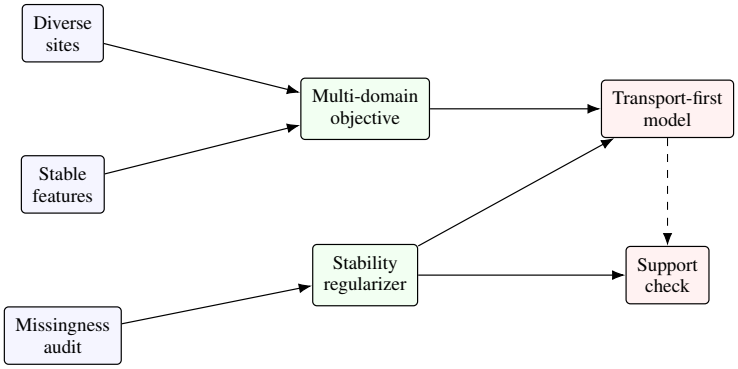

FIGURE 5: Transport-first development architecture. Instead of maximizing local fit in a single archive, the pipeline begins with heterogeneous sites, feature choices guided by semantic stability, and explicit auditing of observation processes such as missingness. The learning objective then combines predictive fit with stability constraints and support awareness so that the resulting predictor is optimized for plausible deployment variation rather than for one historical environment alone.

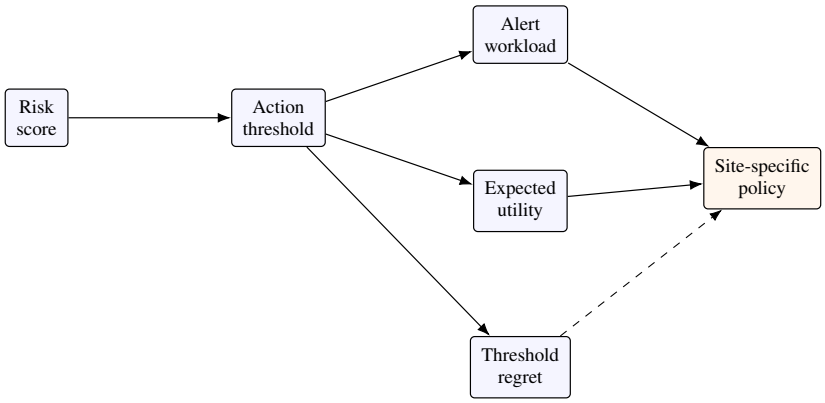

FIGURE 6: Decision-theoretic consequences of failed generalization. The same numerical threshold can induce very different action burdens and utility profiles across institutions because prevalence, calibration, and operational capacity shift together. A transportable model is therefore one whose score preserves threshold meaning sufficiently well that workload, actionability, and policy regret remain governable outside the development setting.

TABLE 3: Types of Transport Failure

Type	Cause	Symptom	Metric Affected
Covariate shift	$P(X)$ change	Different feature range	Calibration
Conditional shift	$P(Y\|X)$ change	Altered relationships	All metrics
Label shift	$P(Y)$ change	Prevalence variation	Thresholds
Representation shift	Feature semantics	Same name, diff meaning	Stability

value can itself predict low risk. In the second, the same absence carries a very different meaning. A model that treats missingness indicators as predictive features without asking whether their semantics are stable across domains may inadvertently encode local workflow. Internal performance can improve precisely because the model has captured a domain marker.

Measurement heterogeneity introduces a related but distinct instability. Clinical variables are not always the objective, context-free quantities that model inputs appear to assume. Even common measurements can differ by assay platform, timestamp granularity, normalization convention, or the event to which they are aligned. Free-text notes depend on template structure and documentation culture. Radiologic findings may reflect both acquisition protocol and interpretive threshold. Administrative codes are shaped by billing logic and institutional coding support. The same column name in two data warehouses may therefore correspond to different empirical objects. If a model exploits those differences unintentionally, its internal optimization objective encourages dependence on institution-specific representation rather than persistent clinical content.

The outcome side is no more stable. Many clinically interesting events are soft labels assembled from chart ab-

TABLE 4: Validation Levels Across Domains

Level	Data Source	Purpose	Limitation
Internal	Same dataset	Optimism control	No transport insight
Temporal	Same site later	Drift detection	Limited heterogeneity
External single	One new site	Basic transport test	No variance estimate
External multi-site	Multiple sites	Heterogeneity estimation	Higher complexity

TABLE 5: Calibration Assessment Across Domains

Measure	Symbol	Meaning	Interpretation
Intercept	α_t	Baseline shift	Prevalence change
Slope	β_t	Scale distortion	Over/underfitting
Curve	$m_t(s)$	Local calibration	Threshold reliability
Error	ECE	Avg deviation	Global miscalibration

TABLE 6: Adaptation Strategies in New Domains

Method	Change Type	Complexity	Use Case
None	Direct use	Low	Strong transport
Intercept update	Baseline adjust	Low	Prevalence shift
Slope recalibration	Scale adjust	Medium	Mild distortion
Refitting	Parameter update	High	Structural shift

TABLE 7: Sources of Temporal Drift

Source	Change Type	Example	Effect
Clinical protocols	Treatment	New pathways	Risk reduction
Measurement tools	Data capture	New assays	Feature shift
Documentation	Recording	EHR updates	Label change
Population	Case mix	Aging cohort	Prevalence drift

stractions, coding rules, thresholds for reoperation, escalation to intensive care, or patterns of postdischarge follow-up. Their prevalence across institutions may change because patients differ, because treatment succeeds differentially, because surveillance intensity differs, or because the event is defined differently in practice [10]. These mechanisms are not interchangeable. A rise in recorded complications may indicate higher true risk, more thorough detection, or more expansive adjudication. A model transported into such heterogeneity may seem poorly calibrated for reasons that reflect label non-equivalence rather than defective ranking. Yet from the standpoint of deployment, this distinction does not rescue the model. If the event label changes its meaning, bedside interpretation of the score becomes unstable.

A more compact representation of institutional non-exchangeability can be obtained by introducing a latent institutional coordinate Z_d that captures workflow, measurement, and treatment style. Then one may write

$$\begin{aligned} X &= X^{\mathrm{bio}} + L_d Z_d + \varepsilon_x, \\ Y &= g\big(X^{\mathrm{bio}}, A, Z_d\big) + \varepsilon_y, \end{aligned} \tag{2}$$

where X^{bio} denotes a latent biologically relevant state, L_d a loading matrix translating institutional structure into recorded covariates, and g an outcome surface that depends not only on latent biology but also on intervention and institutional context. The translational problem arises because empirical learning observes X, not X^{bio}, and can reduce source-domain loss by exploiting the $L_d Z_d$ component. When that component shifts outside the source domain, the learned relation is no longer aligned with a stable target. The failure is structural rather than incidental.

The consequence is that hospitals should not be treated as interchangeable sampling frames unless evidence supports that assumption. Institutional labels often mark differences in the observation process itself, and therefore in the statistical meaning of predictors. A model's apparent success can reflect its ability to exploit these local regularities. That is precisely why external validation is indispensable. It is the first empirical opportunity to test whether what the model learned is robust clinical structure or simply one institution's version of it.

III. The Statistical Geometry of Transport Failure

To understand why source-domain success can coexist with poor transport, it is helpful to think geometrically. Clinical prediction maps a high-dimensional representation of

TABLE 8: Decision-Theoretic Quantities

Quantity	Symbol	Role	Sensitivity
Threshold	c	Action cutoff	Calibration
Workload	$W_d(c)$	Alert rate	Score distribution
Utility	$U_d(c)$	Net benefit	Costs/benefits
Regret	$\mathcal{L}_{thr}$	Decision error	Local miscalibration

TABLE 9: Transport-Oriented Development Principles

Principle	Goal	Method	Tradeoff
Multi-domain data	Coverage	Diverse sites	Coordination cost
Feature stability	Robustness	Audit variables	Lower local fit
Regularization	Consistency	Penalize variance	Bias increase
Support awareness	Reliability	Abstention rules	Reduced coverage

patients into a scalar or vector of estimated risks. During training, the model learns directions in feature space that reduce empirical loss under the source distribution. Nothing in that objective guarantees that those directions coincide with clinically stable axes of variation [11]. If the source environment contains institutional artifacts that correlate strongly with the outcome, then optimization will align itself with those artifacts because they are predictive where the model is being judged. Transport failure occurs when these learned directions are not preserved in target domains.

Let $X_d \in \mathbb{R}^{n_d \times p}$ denote the predictor matrix for domain d. A singular value decomposition

$$X_d = U_d \Sigma_d V_d^\top \tag{3}$$

separates source-domain variation into orthogonal directions carried by the columns of V_d. In a translationally idealized setting, the dominant predictive directions would correspond to stable clinical mechanisms, and therefore the subspaces associated with large predictive gradients would align across domains. In practice, the leading directions of predictive usefulness may be mixtures of biological signal and institution-specific representation. One domain may emphasize trajectories generated by dense surveillance, another may emphasize variables stabilized by standard order sets, and a third may encode severity through transfer patterns not present elsewhere. If a model loads strongly on such directions, the geometry of its decision surface becomes domain-dependent.

This can be quantified by comparing predictor subspaces across domains. If $V_d^{(k)}$ and $V_t^{(k)}$ are the first k right singular vectors in domains d and t, one may define a subspace discrepancy

$$\mathcal{G}_{d,t}^{(k)} = \left\| V_d^{(k)} V_d^{(k)\top} - V_t^{(k)} V_t^{(k)\top} \right\|_F . \tag{4}$$

A large value indicates that the principal axes of observed variation differ between the two environments. This alone does not imply transport failure, because not all dominant variance is predictive. The more relevant quantity concerns the alignment of predictive gradients with those subspaces. Let a differentiable score function be $f_d(x)$. Then the expected gradient covariance

$$H_d = \mathbb{E}_{P_d}\left[\nabla f_d(X)\, \nabla f_d(X)^\top\right] \tag{5}$$

describes directions in feature space to which the score is most sensitive. If the leading eigenspaces of H_d align with unstable institutional components of X_d, the model's prediction geometry is unlikely to transport even when the raw feature distributions appear broadly similar.

Calibration failure can also be understood geometrically. A scalar risk score compresses a high-dimensional observation into one dimension [12]. In the source domain, this compression induces a calibration surface that maps score values to empirical event frequency. If target-domain observations project differently onto the same score axis, then points with identical scores may correspond to different neighborhoods in the original feature space, and therefore to different risks. The score remains numerically familiar while its semantic neighborhood has changed. This is one reason a model can preserve some ranking ability and yet lose calibration badly. The score axis exists in both domains, but the local geometry around that axis differs.

One may formalize this using conditional score densities. Let $S = f(X)$. For event class $y \in \{0, 1\}$, the source and target domains induce densities $p_d(s \mid y)$ and $p_t(s \mid y)$. Discrimination depends on separation between these class-conditional score densities, whereas calibration depends on the posterior

$$\Pr(Y = 1 \mid S = s, D = t) = \frac{\pi_t\, p_t(s \mid 1)}{\pi_t\, p_t(s \mid 1) + (1 - \pi_t)\, p_t(s \mid 0)}. \tag{6}$$

Even if the ordering induced by S changes little, modest distortions in $p_t(s \mid y)$ relative to $p_d(s \mid y)$ can produce substantial posterior differences, especially near clinically relevant thresholds. The model's geometry has not collapsed in a dramatic global sense; it has warped enough to alter the clinical meaning of a given score.

Finite-sample training makes this worse because empirical risk minimization is opportunistic. Suppose the true target

is a stable signal component X^{stb}, but the source domain also contains a local proxy X^{loc} highly correlated with the outcome through one institution's workflow. With limited sample size or flexible function classes, the model can reduce source loss by overloading on X^{loc}. One may write an estimator in a linear form

$$\hat{f}_d(x) = x^\top \hat{\beta} = {x^{\mathrm{stb}}}^\top \hat{\beta}_{\mathrm{stb}} + {x^{\mathrm{loc}}}^\top \hat{\beta}_{\mathrm{loc}}. \tag{7}$$

Even if the stable component is present, transport depends critically on the magnitude of $\hat{\beta}_{\mathrm{loc}}$. Internal resampling will not reliably expose the problem because the local proxy is reproducible within the source distribution. The model can appear well behaved under cross-validation and yet degrade when X^{loc} shifts or disappears in the target. This is not overfitting in the narrow sense of fitting random noise. It is overcommitment to unstable structure.

A useful decomposition of target-domain risk makes this point explicit. Let $\mathcal{R}_d(f)$ and $\mathcal{R}_t(f)$ denote source and target risk under a loss ℓ. Then one may write

$$\begin{aligned}\mathcal{R}_t(f) &= \mathcal{R}_d(f) + \Big(\mathcal{R}_t(f) - \mathcal{R}_d(f)\Big) \\ &= \mathcal{R}_d(f) + \Delta_{\mathrm{marg}}(f) + \Delta_{\mathrm{cond}}(f),\end{aligned} \tag{8}$$

where $\Delta_{\mathrm{marg}}(f)$ summarizes shift induced by changes in the predictor distribution and $\Delta_{\mathrm{cond}}(f)$ summarizes shift induced by changes in the conditional relation between predictors and outcome. Many validation discussions focus only on the first term and describe the problem as one of case mix. But in clinical transport, the second term is often central because interventions, measurement conventions, and label construction change the conditional structure itself [13]. Once $\Delta_{\mathrm{cond}}(f)$ becomes material, source-domain risk provides a weak guide to target performance.

These geometric considerations imply a different interpretation of complexity. A more flexible model class is not inherently advantageous for translation. Flexibility increases the capacity to adapt to stable nonlinear clinical signal, but it also increases the capacity to capture institution-specific representational quirks. The relevant question is whether the function class is constrained by a stability principle. If not, source-domain optimization alone does not distinguish a clinically meaningful manifold from a local shortcut manifold. The model may discover a sharp decision boundary because the source domain makes one available, not because that boundary corresponds to something durable.

This perspective also helps explain why simple models sometimes transport as well as or better than more complex ones. A parsimonious model built on relatively stable features may ignore certain local shortcuts and therefore incur slightly higher source loss but lower cross-domain variance. Conversely, a richly parameterized model may dominate internally by exploiting many weak source-specific regularities that are invisible to global summary metrics. The comparison should therefore not be framed as complexity against simplicity in the abstract. It should be framed as local fit against stability of the learned geometry under domain perturbation.

What external validation reveals, in geometric terms, is whether the model's decision surface has been anchored to source-specific coordinates. A transported model that preserves ranking, calibration, and threshold behavior across institutions suggests that its dominant predictive directions align with something more stable than local workflow. A model that fails externally indicates that its source-domain geometry was too entangled with one institution's observational or therapeutic structure. This is why external validation is not simply a more stringent test. It is a different kind of test, one that interrogates the coordinate system in which the model's apparent success was originally defined.

IV. External Validation as Estimation Under Domain Uncertainty

External validation is often described procedurally, as though one simply evaluates the trained model on an independent dataset and reports the resulting metrics. That description misses the inferential problem [14]. The relevant deployment environment is rarely a single fixed target. A model may be intended for use across multiple hospitals, across future years within the same hospital, or across service lines that differ in measurement density and intervention patterns. The practical question is therefore not merely whether the model performs in one external cohort, but what should be estimated about its behavior under a family of plausible target domains. External validation is best viewed as estimation under domain uncertainty.

Let $\mathcal{T}$ denote a collection of target domains that the model might encounter. For each domain $t \in \mathcal{T}$, define a performance functional $G_t(f)$, which may represent discrimination, calibration loss, decision utility, or a vector of such quantities. The translational object of interest is not a single number but a distribution over $G_t(f)$ induced by uncertainty about which domain will be encountered and how that domain differs from the source. A basic summary is the expected target performance

$$\bar{G}(f) = \int_{\mathcal{T}} G_t(f)\, d\Pi(t), \tag{9}$$

where Π is a weighting distribution over anticipated deployment settings. Yet the mean is not sufficient. Two models with the same $\bar{G}(f)$ may differ profoundly in stability, and the one with lower between-domain dispersion may be operationally preferable. Accordingly, one also needs

$$\mathrm{Var}_\Pi\big(G_t(f)\big) = \int_{\mathcal{T}} \big(G_t(f) - \bar{G}(f)\big)^2\, d\Pi(t). \tag{10}$$

The translationally interesting model is often not the one with the best average performance in the literature, but the one with acceptable mean behavior and lower variance across plausible deployment environments.

A single external dataset can estimate at most one point of this landscape. It may be useful, but it cannot by itself characterize performance heterogeneity. This is why multi-site

or multi-period validation is fundamentally different from one-off external testing. With multiple target domains, the model's performance profile becomes an estimable random-effects structure rather than a solitary statistic. If g_j is the observed performance in external domain j, one may use a hierarchical model

$$\begin{aligned} g_j &= \mu_g + b_j + \varepsilon_j, \\ b_j &\sim \mathcal{N}(0, \tau_g^2), \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma_j^2), \end{aligned} \tag{11}$$

where μ_g is mean external performance, τ_g^2 is between-domain heterogeneity, and σ_j^2 is within-domain estimation variance. This decomposition matters because a model with favorable μ_g but large τ_g^2 is less predictable in deployment than a slightly weaker but more stable alternative. In many clinical settings, uncertainty about where performance will land is itself a major source of implementation risk [15].

Calibration deserves special treatment in this framework because it mediates how numerical scores are interpreted across domains. Suppose the source model outputs a probability-like score $s = f(X)$. In target domain t, one can characterize calibration by

$$\Pr(Y = 1 \mid S = s, D = t) = m_t(s), \tag{12}$$

with m_t estimated nonparametrically or via a logistic recalibration model

$$\operatorname{logit} m_t(s) = \alpha_t + \beta_t \operatorname{logit}(s). \tag{13}$$

Here α_t captures baseline risk displacement and β_t captures scale distortion. External validation that reports only a target-domain c-statistic leaves the most operationally relevant part of transport unmeasured. A score can maintain ordering while requiring substantial recalibration before it becomes interpretable for threshold-based action. The size and variability of (α_t, β_t) across target settings are therefore key estimands, not supplementary diagnostics.

There is a second distinction that external validation should make explicit: pure transport versus transport after local adaptation. A model evaluated without modification in a new domain tests whether its learned relation is directly portable. A model that undergoes intercept adjustment, slope recalibration, or partial refitting tests a different proposition, namely whether the original model can serve as a stable scaffold for local updating. Both questions matter, but they should not be conflated. If a model requires substantial local correction everywhere it goes, it may still be useful, but its translational burden is greater than that of a model requiring little adaptation. The amount of adjustment needed is itself evidence about the model's stability.

This distinction can be represented through a nested sequence of estimators. Let $f^{(0)}$ be the original source-trained model, $f_t^{(1)}$ an intercept-corrected version, $f_t^{(2)}$ an intercept-and-slope recalibrated version, and $f_t^{(3)}$ a partially revised version with selected coefficients re-estimated. For a given target domain,

$$G_t\big(f^{(0)}\big), \quad G_t\big(f_t^{(1)}\big), \quad G_t\big(f_t^{(2)}\big), \quad G_t\big(f_t^{(3)}\big) \tag{14}$$

form an adaptation ladder. Comparing these values reveals whether poor raw transport is due mainly to prevalence shift, to broader calibration distortion, or to deeper coefficient instability [16]. Such information is far more actionable than a binary statement that the model did or did not validate externally.

Precision of estimation matters as much as point estimates. Many external validation studies are underpowered with respect to event counts, especially when the outcome is uncommon or subgroup analysis is attempted. In such settings, favorable results can be too imprecise to justify operational use, and unfavorable results can be ambiguous because estimation error overwhelms true signal. Since deployment decisions are threshold-sensitive, uncertainty intervals around calibration, sensitivity at operational thresholds, and workload fractions are essential. A model intended to trigger resource-intensive review should not be regarded as externally supported merely because a point estimate looks acceptable in a small sample.

External validation should also be aligned with the eventual action context. If the model is intended to inform bedside probability communication, calibration and interval estimates are central. If it is intended to prioritize chart review under limited staffing, rank stability and workload curves matter more. If it is intended to trigger prophylactic treatment, expected utility under plausible cost ratios and site-specific prevalence becomes essential. The same external cohort can support these analyses, but only if the validation plan is explicit about use. Otherwise, validation reports remain detached from the operational question they are supposed to settle.

Reframed in this way, external validation is not the last administrative step in model development. It is the empirical procedure that estimates whether a predictor trained under one domain family can support claims about another. It should therefore be designed with the same conceptual care as model fitting itself. Which domains matter, how they are weighted, what form of recalibration is allowed, what uncertainty is acceptable, and what actions depend on the output are all part of the estimand. Once this is recognized, the field's usual emphasis on internal optimization at the expense of domain-wise estimation looks increasingly misallocated. The central translational evidence is generated not when the model is first fit, but when it is first confronted with the heterogeneity of the world in which it is meant to be used.

V. Temporal Drift, Intervention Feedback, and the Endogeneity of Risk

Even if a model transported cleanly across institutions at one point in time, its translational problem would remain incomplete [17]. Clinical environments evolve. Protocols change, case mix shifts, measurement technologies are replaced, staffing patterns vary, and the introduction of the

model itself can modify behavior. The target of prediction is therefore not stationary. Risk in healthcare is endogenous to the system in which it is measured, because the same variables used for prediction often trigger interventions that alter the event process. This property distinguishes clinical deployment from many static prediction settings and makes temporal validation and prospective surveillance indispensable.

A useful starting point is to index time explicitly. Let X_t denote observed predictors at time t, A_t the action taken, and $Y_{t+\Delta}$ the future outcome. The predictive task is often represented as estimation of $\Pr(Y_{t+\Delta} = 1 \mid X_t)$. In reality, the outcome depends on a dynamic treatment trajectory that is partly induced by observed risk. A more honest formulation is

$$\Pr(Y_{t+\Delta} = 1 \mid X_t) = \sum_{a_t} \Pr(Y_{t+\Delta} = 1 \mid X_t, A_t = a_t) \Pr(A_t = a_t \mid X_t). \tag{15}$$

When a model is deployed, it can alter the second factor by changing clinician behavior. High-risk alerts may produce earlier imaging, prophylaxis, transfer, or consultation. If those actions reduce event incidence among high-scoring patients, the observed relation between X_t and $Y_{t+\Delta}$ will drift. A model can therefore make its own historical calibration obsolete. This is not a paradox. It is what one should expect when prediction is coupled to intervention.

Temporal drift also occurs in the absence of model-induced feedback. New documentation systems change the granularity of charted data. Assay revisions shift laboratory distributions. Enhanced recovery pathways alter postoperative trajectories. Coding updates change how complications are recorded [18]. Staffing shortages modify monitoring frequency. All of these phenomena affect either predictors, outcomes, or the link between them. A model developed on data from one era may appear internally robust because retrospective folds share the same operational context. Once applied years later, the same model encounters a new equilibrium between patient state, observation, and response. Temporal validation within the same institution is therefore not a weak substitute for geographic validation; it tests a distinct dimension of transport, namely whether the learned relation survives the institution's own evolution.

Dynamic instability can be represented with a state-space formulation. Let H_t denote the latent clinical state, evolving as

$$\begin{aligned} H_{t+1} &= F_t H_t + B_t A_t + \omega_t, \\ X_t &= C_t H_t + \nu_t, \end{aligned} \tag{16}$$

where F_t captures underlying disease and recovery dynamics, B_t maps intervention into state change, and C_t maps latent state into observations. If any of these operators vary over time, the predictive meaning of X_t shifts. The score produced by a model trained under one set of operators may then become misaligned with current risk. Particularly important is variation in B_t, because changes in treatment efficacy or response timing alter the extent to which a recorded high-risk pattern leads to observed adverse outcomes. A predictor that looked highly informative when rescue actions were slow may appear weaker after protocols improve, not because it was originally wrong, but because the system learned to intervene earlier.

Intervention feedback introduces a more subtle problem: selective label suppression. If clinicians act on high-risk cases effectively, the very patients the model flags may no longer experience the event at the historical rate. Evaluated naïvely, the model could appear overpredictive. Yet the score may still be identifying truly vulnerable patients, now rendered lower-risk by action. This complicates postdeployment validation. The observed outcome is no longer a passive truth about baseline risk [19]. It is a realized endpoint under a policy partly conditioned on the model. In such settings, retrospective notions of calibration need reinterpretation. One must ask whether the model estimates untreated or usual-care risk, and whether the postdeployment policy is expected to change that estimand.

This issue matters because many deployment failures are misdiagnosed as model decay when they are actually policy interactions. A hospital may report that predicted probabilities now exceed observed event rates after implementation. That observation could mean the model no longer transports, or it could mean the hospital is averting some events in alerted patients. Distinguishing these possibilities requires auxiliary analyses of action pathways, not merely updated discrimination metrics. Without that distinction, the institution may either abandon a useful model or retain a failing one under the mistaken belief that intervention explains the drift.

From a design perspective, the existence of drift and feedback implies that validation cannot be a one-time event. It must become a lifecycle process. Let g_t denote a rolling estimate of a performance functional such as calibration slope or workload fraction. A surveillance rule can be written as

$$Z_T = \sum_{t=1}^{T} w_t \left(g_t - g_0\right), \tag{17}$$

with alarm thresholds chosen to detect meaningful departure from baseline. Whether one uses cumulative sum methods, Bayesian updating, or sliding-window estimates, the goal is the same: monitor whether the model's operating behavior remains within acceptable bounds. Importantly, the monitored quantities should include not only discrimination but also calibration, subgroup performance, action rate, and missing-input frequency, because drift may first appear operationally rather than in a global ranking metric.

Temporal adaptation, however, is not free. Frequent updating can improve local fit while reducing comparability over time and increasing maintenance burden. It may also lock the model more tightly to one institution's current workflow, potentially reducing cross-site transport [20]. This

creates a familiar tension between local responsiveness and broader portability. A transport-first perspective suggests that updates should be documented as changes in the model's deployment scope. A heavily adapted local model may be quite useful, but it no longer supports broad claims about generalizability. Conversely, a globally stable model may require periodic recalibration to remain locally interpretable without abandoning its wider domain scaffold.

The larger point is that clinical risk prediction is embedded in adaptive systems. External validation remains central because it establishes whether the model survives movement across contexts, but temporal validation and postdeployment monitoring are equally necessary because the context itself does not remain fixed. Translation is not the transfer of a static object from one place to another. It is the attempt to maintain coherent predictive meaning while institutions, measurements, and responses evolve. Once prediction is understood in this dynamic way, the notion that a high-performing retrospective model is nearly ready for routine use becomes difficult to sustain.

VI. Transport-First Model Development and Robust Estimation Strategies

If generalizability is the main translational barrier, then development should be reorganized around that barrier rather than around local performance maximization. The prevailing pipeline often begins with a single rich dataset and asks which algorithm can extract the strongest internal signal. A transport-first pipeline begins by asking which sources of variation are likely to destabilize prediction across intended deployment settings, and then constructs the data, features, objectives, and reporting scheme to address those variations explicitly. The difference is not superficial. It changes what counts as a useful feature, a successful model, and a sufficient validation plan.

Data assembly is the first design choice that reflects this shift. More records from one institution are not always more valuable than fewer records from a strategically diverse collection of institutions. What matters for transport is not only sample size but coverage of institutional mechanisms. A development consortium spanning different care intensities, documentation systems, laboratory infrastructures, and patient pathways provides a richer basis for learning stable structure than a much larger monocentric archive. The aim is not simply to average over sites, but to expose the model to the forms of heterogeneity it will later encounter [21]. Without that exposure, even large-scale training can create an illusion of robustness while remaining narrow in domain terms.

Feature design should likewise be guided by stability considerations. Variables tightly coupled to irreversible or slowly varying physiology are often more promising than variables whose predictive value depends on local logistics. This does not mean excluding all workflow-sensitive features. Some may be operationally useful. It means distinguishing between features chosen because they summarize patient state and features chosen because they exploit site-specific artifacts. A modest reduction in source-domain fit can be acceptable if it yields a representation whose semantics are more likely to survive transport. In practical terms, this calls for stress-testing features across domains, auditing whether missingness patterns are institution markers, and examining whether measurement transformations behave similarly across sources.

A general multi-domain objective can be written as

$$\begin{aligned}\min_{f\in\mathcal{F}} \quad & \sum_{d=1}^{m} \pi_d \,\hat{\mathcal{R}}_d(f) \\ & + \lambda\,\Omega(f) \\ & + \gamma\,\mathcal{S}(f),\end{aligned} \tag{18}$$

where $\hat{\mathcal{R}}_d$ is empirical loss in domain d, $\Omega(f)$ a complexity penalty, and $\mathcal{S}(f)$ a stability regularizer. The stability term can take several forms: penalizing between-domain variance in calibration slope, encouraging invariance of representation moments, constraining coefficients to vary smoothly across sites, or discouraging excessive reliance on features whose predictive contribution is highly domain-specific. The essential point is that source fit is not the only objective. The model is explicitly rewarded for stable behavior across heterogeneous environments.

Hierarchical parameterizations are especially useful when domains differ but are not wholly unrelated. Suppose a generalized linear predictor with shared coefficients β and domain-specific deviations δ_d:

$$\begin{aligned}\operatorname{logit}\,\Pr(Y=1\mid X, D=d) &= \alpha+\alpha_d + X^\top\beta + X^\top\delta_d, \\ \delta_d &\sim \mathcal{N}(0,\Psi).\end{aligned} \tag{19}$$

This structure allows one to estimate which predictors behave consistently and which require domain-specific adjustment. A small posterior variance for a coefficient across domains suggests a stable feature. A large variance suggests context sensitivity. Such models are not only predictive devices but also transport diagnostics [22]. They identify the components of the score most likely to fail under unseen domain shift.

Representation learning can be adapted in the same spirit. Let $\phi_\theta(X)$ be a learned representation and g_ψ the predictor head. One may optimize

$$\begin{aligned}\min_{\theta,\psi} \quad & \sum_{d=1}^{m} \pi_d \,\hat{\mathcal{R}}_d\big(g_\psi(\phi_\theta(X))\big) \\ & + \eta \sum_{d=1}^{m} \left\| \mu_d^{\phi} - \bar{\mu}^{\phi} \right\|_2^2 \\ & + \zeta \sum_{d=1}^{m} \left\| \Sigma_d^{\phi} - \bar{\Sigma}^{\phi} \right\|_F^2,\end{aligned} \tag{20}$$

where μ_d^ϕ and Σ_d^ϕ are the mean and covariance of the learned representation in domain d. The aim is not to erase all domain differences, which could suppress clinically meaningful signal, but to discourage representations whose

predictive value relies heavily on domain-specific geometry. Such regularization is only defensible when paired with substantive scrutiny, since enforcing invariance without regard to clinical mechanism can remove useful structure. The relevant principle is not blind homogenization but selective stabilization.

Robust development also requires explicit treatment of unsupported regions. A model trained across several domains may still face patients whose feature combinations are weakly represented anywhere in the training data. Instead of forcing precise-looking outputs in these regions, a transport-first system should estimate support and permit abstention or deferral. Let $\rho_d(x)$ denote a domain support function, and define

$$s(x) = \sum_{d=1}^{m} \pi_d \, \rho_d(x). \tag{21}$$

Predictions for points with low $s(x)$ can be flagged as low-support cases, routed to clinician review, or accompanied by wider uncertainty intervals. This is not merely a defensive technical device. It acknowledges that generalization is uneven across the predictor space and that the model's confidence should reflect that unevenness.

Missingness should be modeled as part of the data-generating process rather than treated solely as a preprocessing nuisance. Multiple imputation, missing indicators, and learned embeddings each have roles, but none is sufficient without examining whether the missingness mechanism is stable across sites. In some settings, the most transportable strategy may be to rely on features available under routine workflows in most target domains, even if more exotic variables offer higher source-domain performance [23]. In others, the missingness pattern itself may be valuable, but only after confirming that its clinical meaning is not highly institution-dependent. A transport-first design therefore demands exploratory work that many development papers omit: assessing how observation processes vary across domains and whether the model is learning from those differences.

Interpretability also takes on a new role in this framework. Explanatory methods are often invoked to improve clinician trust, but they are equally valuable as transport audits. If site-specific variables or implausible proxies dominate local feature attributions, the model may be exploiting unstable structure. Conversely, a model whose predictive burden rests on clinically coherent features across domains is not automatically transportable, but it offers a more credible basis for expecting transport. Interpretability should therefore be used less as a post hoc reassurance device and more as a diagnostic instrument during development.

The result of these strategies is not a guarantee of external validity. No method can ensure portability across arbitrarily different institutions. What transport-first development can do is reduce the mismatch between the objective optimized during training and the claim implied at deployment. Instead of asking the model to fit one historical environment as closely as possible and then hoping that this fit generalizes, the development process asks the model to remain useful under known forms of heterogeneity. That change in objective does not eliminate translational risk, but it makes the evidence generated during development more relevant to the problem translation actually poses.

VII. Decision-Theoretic and Operational Consequences of Failed Generalization

The practical harm of poor generalizability is not exhausted by degraded statistical performance. A transported model affects work allocation, clinical attention, intervention timing, and the distribution of false alarms and missed cases. These are decision-theoretic consequences. A model may suffer only modest average calibration error and yet generate serious operational inefficiency if that error is concentrated near an action threshold or within a subgroup that consumes substantial resources when flagged. Translational assessment must therefore move from performance description to consequence analysis.

Suppose a policy acts on patients with score $f(X)$ exceeding threshold c [24]. Let the benefit of appropriate action for a true case be b, the burden of unnecessary action for a noncase be k, and the harm of missed opportunity for a case below threshold be h. Then the expected utility in domain d can be written as

$$\begin{aligned} U_d(c) = {} & b \Pr_d\big(f(X) \geq c, Y = 1\big) \\ & - k \Pr_d\big(f(X) \geq c, Y = 0\big) \\ & - h \Pr_d\big(f(X) < c, Y = 1\big). \end{aligned} \tag{22}$$

A threshold chosen in the source domain,

$$c_d^* = \arg \max_{c \in [0,1]} U_d(c), \tag{23}$$

need not be close to optimal in a target domain. Prevalence may differ, resource costs may differ, and calibration may shift so that the same numeric score represents a different underlying risk. Transporting the threshold without reestimation can therefore lead to systematic underaction or overaction even if global discrimination remains moderate to strong. This is the operational face of failed generalization.

Workload instability is one of the most immediate manifestations. Let

$$W_d(c) = \Pr_d\big(f(X) \geq c\big) \tag{24}$$

denote the proportion of patients triggering action. Two domains can share similar c-statistics while having very different $W_d(c)$ because their score distributions differ or because their calibration curves displace risk mass differently. A hospital with limited specialist coverage may be unable to absorb the alert burden generated by a threshold tuned elsewhere. In practice, teams often respond by informally raising the threshold until workload is tolerable, thereby changing sensitivity and policy utility without a clear evidentiary basis. External validation should anticipate

this problem by reporting action-rate curves and capacity-constrained operating points, not only abstract discrimination measures.

The regret induced by generalization error is often concentrated near clinically meaningful thresholds. Let $r_d(x)$ denote the true target-domain risk and $f(x)$ the deployed estimate. If action should be taken whenever $r_d(x) \geq c$, then the error most relevant to policy is the set of cases for which $f(x)$ and $r_d(x)$ lie on opposite sides of c. This can be expressed as

$$\mathcal{L}_{\mathrm{thr},d}(f) = \mathbb{E}_{P_d}\Big[\Delta u(X,Y)\,\mathbf{1}\big\{(r_d(X)-c)(f(X)-c)<0\big\}\Big], \tag{25}$$

where $\Delta u(X,Y)$ is the utility gap between the correct and incorrect action. A model can have acceptable global calibration and yet perform poorly on $\mathcal{L}_{\mathrm{thr},d}(f)$ if local distortions cluster around c. This is one reason clinical utility cannot be inferred from average loss alone [25]. Translationally, the relevant question is often not whether the model predicts well on average, but whether it predicts safely where decisions change.

Subgroup heterogeneity complicates matters further. If the consequences of false positives and false negatives differ across patient groups or service lines, then identical calibration error can produce unequal practical effects. A threshold optimized for overall utility may overburden one unit, underdetect risk in another, or generate systematically different error trade-offs across demographic groups. This is not reducible to a generic fairness slogan. It reflects the fact that deployment contexts attach different costs to the same predictive errors. External validation should therefore estimate not only domain-average performance but group-conditional utility and workload profiles. Otherwise, a model may appear acceptable in aggregate while behaving poorly for precisely those populations in which reliable support is most needed.

Human factors amplify these concerns. Clinicians do not respond mechanically to risk estimates. They weigh them against their own assessment, prior experience, and current workload. A model with unstable calibration can erode trust even when its ranking is decent, because clinicians experience repeated mismatches between stated probability and observed reality. Conversely, a model with stable calibration but poorly timed alerts may be ignored because it arrives after decisions are effectively made. Generalization therefore includes the transport of meaning into workflow. If the score's semantics shift across sites, or if operational constraints alter how the score can be acted upon, the model's real-world function changes even if its numerical outputs remain similar.

One implication is that silent prospective deployment is often the only responsible bridge between retrospective validation and action-linked use. Running the model live without exposing outputs to clinicians allows estimation of score distributions, workload fractions, missing-input rates, and calibration under production data pipelines. It also reveals whether variables available in the research extract are reliably available in practice at the intended decision time [26]. These are not peripheral concerns. Many retrospective models quietly depend on inputs that arrive too late, are backfilled, or are systematically missing in real-time operation. A model that cannot reproduce its own feature set in live workflow is not merely inconvenient. It is not the same model in the operational sense.

The decision-theoretic perspective also clarifies why modestly lower but more stable performance may be preferable. Suppose Model A has slightly better mean discrimination than Model B across several validation sets, but Model A's calibration and workload vary widely across institutions. Model B is somewhat less sharp yet more consistent in threshold behavior. For bedside deployment, Model B may be the rational choice because its induced policy is easier to govern, monitor, and communicate. The relevant optimization target is not leaderboard dominance but acceptable regret under realistic institutional variation.

Ultimately, what external validation contributes in decision-theoretic terms is evidence about whether a score can carry the same action meaning in a new environment. A transported model is useful not because it remains mathematically elegant, but because the choices it supports do not become systematically distorted when case mix, workflow, and resource constraints change. Once prediction is treated as part of a decision system, generalizability becomes the central safeguard against policies that are locally trained and globally misapplied.

VIII. Conclusion

The central claim of this paper has been narrow in form but broad in implication. Clinical prediction models are frequently evaluated as if the main technical challenge were extracting signal from large and complicated datasets. In many translational settings, however, the more serious challenge is determining whether the extracted signal remains meaningful outside the environment that produced it. Hospitals are not passive backdrops against which a universal patient distribution is sampled. They are active components of the data-generating process, shaping what is measured, how quickly it is measured, which interventions follow emerging risk, and how outcomes are ultimately recorded. Under these conditions, a model can be statistically impressive in its source domain while remaining weakly justified for use elsewhere [27].

This reframing has consequences for both methodology and interpretation. Generalizability is not an intrinsic badge attached to a trained algorithm. It is a relation between a predictor, a target domain, an action policy, and an acceptable level of uncertainty. External validation matters because it is the first direct empirical procedure that interrogates this relation under changed institutional conditions. Internal validation still has a necessary role in estimating optimism

within a source process, but it cannot reveal whether the source process itself was unusually favorable, unusually narrow, or entangled with local workflows in ways that will not survive transport.

A transport-oriented perspective also changes what counts as a meaningful development strategy. It encourages multi-domain data assembly, explicit modeling of institutional heterogeneity, calibration-centered evaluation, and conservative treatment of features whose predictive content may depend on local observation practices. It favors validation schemes that estimate both average external performance and variability across settings. It treats recalibration burden as evidence rather than inconvenience, and it recognizes that temporal drift and intervention feedback can alter the target of prediction after deployment. None of these considerations eliminates the value of sophisticated modeling. They simply relocate sophistication from the pursuit of local fit toward the pursuit of stable usefulness.

The broader implication is that translational maturity should be judged less by how well a model performs in the environment that created it and more by how transparently its behavior has been characterized under environments that did not. A model that transports with modest adaptation, stable workload, and interpretable calibration may contribute more to care than a model with stronger retrospective metrics but poorly mapped external behavior. That judgment is not anti-innovative. It reflects the ordinary standards of evidence required whenever a system trained on historical data is asked to influence prospective clinical decisions.

For that reason, generalizability and external validation should not remain peripheral expectations appended to otherwise completed predictive studies. They should become the core around which translational modeling is organized. When the intended use is nonlocal, the decisive question is transport. Everything else, including internal discrimination gains and architectural novelty, matters only insofar as it supports a score whose meaning survives beyond the place where it was first learned [28].

REFERENCES

[1] S. Joglekar, A. Huynh, W. Zhou, L. Chong, A. Sayed-Hassen, M. Hii, and R. Cade, "398. the hybrid intra-thoracic oesophago-gastric anastomosis—contemporary performance with low benign anastomotic stricture formation," *Diseases of the Esophagus*, vol. 36, 8 2023.

[2] P. J. J. Herrod, H. Boyd-Carson, B. Doleman, J. Trotter, S. Schlichtemeier, G. Sathanapally, J. Somerville, J. P. Williams, and J. N. Lund, "Quick and simple; psoas density measurement is an independent predictor of anastomotic leak and other complications after colorectal resection," *Techniques in coloproctology*, vol. 23, pp. 129–134, 2 2019.

[3] D. Gupta, M. K. Arora, and M. Srinivas, "Azygos vein anomaly: the best predictor of a long gap in esophageal atresia and tracheoesophageal fistula.," *Pediatric surgery international*, vol. 17, pp. 101–103, 3 2001.

[4] M. Borgulya, C. Ell, and J. Pohl, "Transnasal endoscopy for direct visual control of esophageal stent placement without fluoroscopy.," *Endoscopy*, vol. 44, pp. 422–424, 3 2012.

[5] R. E. Schwarz, "Obesity and postgastrectomy outcomes: large risks, fat chances, or no big deal?," *Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, vol. 15, pp. 121–123, 2 2012.

[6] 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.

[7] F. Younis, M. Shnell, N. Gluck, S. Abu-Abeid, S. Eldar, and S. Fishman, "Endoscopic treatment of leaks and strictures after laparoscopic one anastomosis gastric bypass," 9 2019.

[8] M. Jagielski, J. Piatkowski, G. Jarczyk, and M. Jackowski, "Transrectal endoscopic drainage with vacuum-assisted therapy in patients with anastomotic leaks following rectal cancer resection," *Surgical endoscopy*, vol. 36, pp. 1–9, 3 2021.

[9] E. Rodrigues-Pinto, A. Repici, G. Donatelli, G. Macedo, J. Devière, J. E. van Hooft, J. M. Campos, M. G. Neto, M. Silva, P. Eisendrath, V. Kumbhari, and M. A. Khashab, "International multicenter expert survey on endoscopic treatment of upper gastrointestinal anastomotic leaks," *Endoscopy international open*, vol. 07, pp. E1671–E1682, 11 2019.

[10] M. Schootman, C. Li, J. Ying, S. T. Orcutt, and J. Laryea, "Maximizing readmission reduction in colon cancer patients.," *The Journal of surgical research*, vol. 295, pp. 587–596, 12 2023.

[11] L. Claassen, F. van Workum, and C. Rosman, "Learning curve and postoperative outcomes of minimally invasive esophagectomy.," *Journal of Thoracic Disease*, vol. 11, 12 2018.

[12] K. Safiejko, R. Tarkowski, T. P. Kozlowski, M. Koselak, M. Jachimiuk, A. Tarasik, M. Pruc, J. Smereka, and L. Szarpak, "Safety and efficacy of indocyanine green in colorectal cancer surgery: A systematic review and meta-analysis of 11,047 patients.," *Cancers*, vol. 14, pp. 1036–1036, 2 2022.

[13] W. Ye, Z. Zhu, X. Wu, S. Lin, J. Xu, L. Li, and Z. Huang, "Technical feasibility and oncological safety of low and high ligation of the inferior mesenteric artery in colorectal cancer surgery for asian populations: a systematic review and meta-analysis," 2 2021.

[14] D. Gowthaman, L. R. Gabor, K. Y. Lin, J. Yang, G. A. Dellacerra, S. S. Isani, and D. Y. Kuo, "Use

of endoluminal vacuum therapy after anastomotic leak in a gynecologic oncology patient with rectosigmoid resection: A case report.," 12 2022.

[15] S. Ubels, M. Lubbers, M. H. P. Verstegen, S. A. W. Bouwense, E. van Daele, L. Ferri, S. S. Gisbertz, E. A. Griffiths, P. Grimminger, G. Hanna, M. Hubka, S. Law, D. Low, M. Luyer, R. E. Merritt, C. Morse, C. L. Mueller, G. A. P. Nieuwenhuijzen, M. Nilsson, J. V. Reynolds, U. Ribeiro, R. Rosati, Y. Shen, B. P. L. Wijnhoven, B. R. Klarenbeek, F. van Workum, and C. Rosman, "Treatment of anastomotic leak after esophagectomy: insights of an international case vignette survey and expert discussions.," *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, vol. 35, 4 2022.

[16] M. K. Schilling, M. Eichenberger, C. A. Maurer, R. H. Greiner, P. Zbären, and M. W. Büchler, "Long-term survival of patients with stage iv hypopharyngeal cancer: Impact of fundus rotation gastroplasty," *World journal of surgery*, vol. 26, pp. 561–565, 2 2002.

[17] J. Ye, Y. Tian, L. Wang, Y. Ye, H. Zheng, Y. Xiang, and W. Tong, "Analysis of factors related to anastomotic leakage after transanal total mesorectal excision," *International Journal of Surgery*, vol. 46, pp. 232–237, 4 2019.

[18] C. Fischer, H. F. Lingsma, R. H. Hardwick, D. A. Cromwell, E. W. Steyerberg, and O. Groene, "Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer," *The British journal of surgery*, vol. 103, pp. 105–116, 1 2016.

[19] R. Morse, T. Mouw, M. Moreno, J. Erwin, P. DiPasco, M. Al-Kasspooles, and A. Hoover, "Is minimally invasive surgery preferred? peri-operative outcomes of varying esophagectomy approaches in the treatment of esophageal malignancy," 5 2021.

[20] H. Kikuchi, H. Endo, H. Yamamoto, S. Ozawa, H. Miyata, Y. Kakeji, H. Matsubara, Y. Doki, Y. Kitagawa, and H. Takeuchi, "Impact of reconstruction route on postoperative morbidity after esophagectomy: Analysis of esophagectomies in the japanese national clinical database.," *Annals of gastroenterological surgery*, vol. 6, pp. 46–53, 9 2021.

[21] A. C. Gordon, A. J. Cross, E. W. Foo, and R. H. Roberts, "C-reactive protein is a useful negative predictor of anastomotic leak in oesophago-gastric resection," *ANZ journal of surgery*, vol. 88, pp. 223–227, 7 2016.

[22] G. Nandakumar, B. G. Richards, K. Trencheva, and G. Dakin, "Surgical adhesive increases burst pressure and seals leaks in stapled gastrojejunostomy," *Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery*, vol. 6, pp. 498–501, 1 2010.

[23] J. P. Grantham, A. Hii, and J. Shenfine, "Preoperative risk modelling for oesophagectomy: A systematic review.," *World journal of gastrointestinal surgery*, vol. 15, pp. 450–470, 3 2023.

[24] 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.

[25] S. Awad, A. I. A. El-Rahman, A. Abbas, W. Althobaiti, S. Alfaran, S. Alghamdi, S. Alharthi, K. Alsubaie, S. Ghedan, R. Alharthi, M. Asiri, A. Alzahrani, N. Alotaibi, A. Shoma, and M. S. A. Sheishaa, "The assessment of perioperative risk factors of anastomotic leakage after intestinal surgeries; a prospective study," *BMC surgery*, vol. 21, pp. 1–9, 1 2021.

[26] A. I. Amin and I. A. Shaikh, "Topical negative pressure in managing severe peritonitis: A positive contribution?," *World journal of gastroenterology*, vol. 15, pp. 3394–3397, 7 2009.

[27] M. Bassiony, A. Ramadan, A. Anwar, A. Said, and R. Sedik, "Robotics in bariatric surgery: Benefits, limitations, and challenges; an umbrella review of systematic reviews and meta-analyses," 9 2023.

[28] J. C. Woodfield, K. Clifford, B. Schmidt, G. A. Turner, M. A. Amer, and J. L. McCall, "Strategies for antibiotic administration for bowel preparation among patients undergoing elective colorectal surgery: A network meta-analysis.," *JAMA surgery*, vol. 157, pp. 34–, 1 2022.