Deployment Envelopes, Label Instability, and Measurement Operators in Clinical Prediction Translation

Pham Quang Huy[1], Doan Minh Tri[2], and Bui Thanh Nam[3]

[1]Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên, Khoa Công nghệ Thông tin, Đường Z115, Phường Quyết Thắng, Thành phố Thái Nguyên 250000, Việt Nam
[2]Trường Đại học Sư phạm Kỹ thuật Hưng Yên, Khoa Công nghệ Thông tin và Truyền thông, Khoái Châu, Xã Dân Tiến, Huyện Khoái Châu, Tỉnh Hưng Yên 160000, Việt Nam
[3]Trường Đại học Đồng Tháp, Khoa Toán – Tin học, 783 Phạm Hữu Lầu, Phường 6, Thành phố Cao Lãnh 810000, Việt Nam

Abstract

Clinical prediction has advanced into an era of abundant data, flexible modeling, and increasingly ambitious claims about decision support. Yet the main obstacle to translation remains unresolved. Models that perform convincingly in retrospective development often fail to preserve meaning when moved into new hospitals, new periods of practice, or new operational contexts. This paper examines that failure through a framework centered on deployment envelopes, measurement operators, and label instability. The argument is that a clinical risk model never learns risk in an abstract and context-free sense. It learns a mapping from institutionally produced records to institutionally produced outcomes, and both sides of that mapping vary across settings. The translational barrier therefore arises because development studies often compress heterogeneous observational regimes into a single statistical object and then treat the resulting model as though it were portable by default. A transport-centered framework is developed in which clinical prediction is represented as inference under uncertainty about how patient state is measured, acted upon, and labeled. This perspective clarifies why internally strong models can remain weakly interpretable for external use, why calibration is often more fragile than ranking, and why threshold policies drift even when discrimination appears preserved. The paper proposes an approach to validation that estimates deployment envelopes rather than isolated test-set scores, emphasizing domain-wise calibration, uncertainty bands on clinically relevant operating points, and explicit accounting for institutional variation in observation and action. The broader conclusion is that external validation is not simply a stronger version of internal testing. It is the principal empirical procedure through which a predictive score acquires stable meaning beyond the local system that generated it.

1 Introduction

The contemporary literature on clinical prediction often presents progress as a steady accumulation of algorithmic competence [1]. Larger datasets, more flexible architectures, and increasingly refined tuning procedures appear to move the field closer to reliable bedside decision support. This view captures part of the story, but it obscures the main difficulty. The central challenge is not merely discovering signal in clinical records. It is determining whether the relation extracted from one clinical environment remains interpretable and useful when the model enters another. A hospital is not only a location in which patients are observed. It is an organized process that determines which patients are seen, which variables are measured, how often they are measured, what interventions follow evolving signs of risk, and how outcomes are ultimately recorded. The data used to train a model are therefore not transparent reflections of patient state. They are products of a local observational and therapeutic regime. When the model is transported, the regime changes, and the score may no longer mean what its numerical form seems to imply.

This difficulty is often underappreciated because predictive studies are commonly organized around retrospective success. A cohort is defined, variables are selected, models are trained, and internal validation is used to estimate performance. The resulting metrics can be methodologically careful and statistically respectable. Even

so, they speak mainly to a narrow question: how well does the model behave under repeated sampling from the same underlying environment that produced the training data? Translation poses a broader question. What happens when the pattern of measurement changes, when an event label is operationalized differently, when rescue interventions occur earlier, or when case mix shifts because referral pathways are not the same? These are not secondary complications that appear after the main scientific task has been completed. They define the scientific task itself whenever a model is intended for use outside its source environment [2].

A common response is to describe this problem under the general heading of generalizability. That label is useful but sometimes too broad. It can suggest that the model either generalizes or does not, as though portability were a global trait. In reality, portability has dimensions. A model may preserve ranking while losing calibration. It may preserve average calibration while failing in clinically important subgroups. It may preserve both ranking and calibration and still cease to support the same intervention threshold because the cost of action is different or because local workflow cannot absorb the same alert burden. The translational question is therefore not only whether performance declines. It is whether the score remains attached to the same clinical interpretation, the same operating logic, and the same practical action. That is a more demanding requirement than retrospective success, and it cannot be inferred from internal testing alone.

The analysis in this paper takes a deliberately different route from standard summaries of transport failure. It does not begin from a familiar contrast between internal and external validation and then proceed to recommend more of the latter. Instead, it reconstructs the translational problem from the ground up by asking what exactly a clinical prediction model has learned. The answer advanced here is that the model has learned a mapping shaped by three institutional objects: a measurement operator translating patient state into recorded variables, an action operator translating emerging evidence into treatment and surveillance, and a label operator translating the downstream course of care into an outcome definition. These operators are rarely modeled explicitly, yet they determine what the predictor can mean. Once that structure is recognized, external validation emerges not as a procedural embellishment but as the only direct empirical test of whether the score survives a change in the operators that generated it [3].

The paper is organized accordingly. The next section reframes clinical prediction as an inverse problem under institution-specific measurement operators and explains why local success is compatible with nonportable inference. The following section studies label instability and shows how outcome definitions drift with surveillance intensity, coding practice, and intervention timing. A later section redefines external validation as estimation of deployment envelopes rather than isolated performance values, emphasizing the range of conditions under which a model remains clinically interpretable. The paper then develops a threshold-transfer geometry that explains why action policies drift even when ranking seems robust. The final analytic section treats deployment as closed-loop system identification, where model output changes clinician behavior and therefore changes the environment the model aims to describe. The argument throughout is restrained but consequential. It does not deny that local predictive systems can be useful. It argues that the decisive barrier to broader translation is the absence of stable evidence that a score trained under one set of operators preserves its meaning under another.

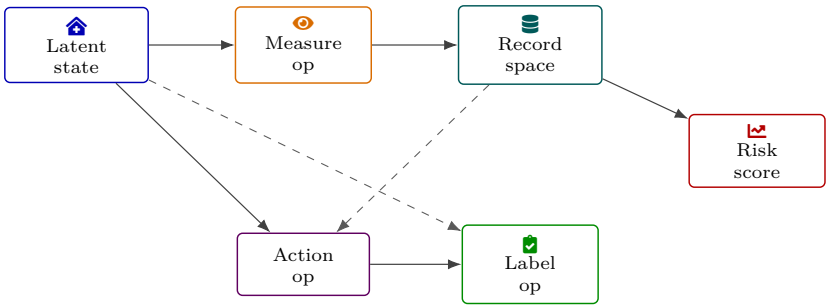

Figure 1: Clinical prediction is framed as inference from institutionally rendered records rather than from patient state directly. The score inherits meaning from how state is measured, how care responds, and how outcomes are labeled.

Table 1: Key Components of the Translational Framework

Component	Definition	Role in Prediction	Stability
Measurement Operator	Maps latent state to data	Shapes predictors	Variable
Label Operator	Maps outcomes to labels	Defines target	Variable
Action Operator	Governs interventions	Alters outcomes	Dynamic
Domain	Institutional context	Bundles practices	Heterogeneous

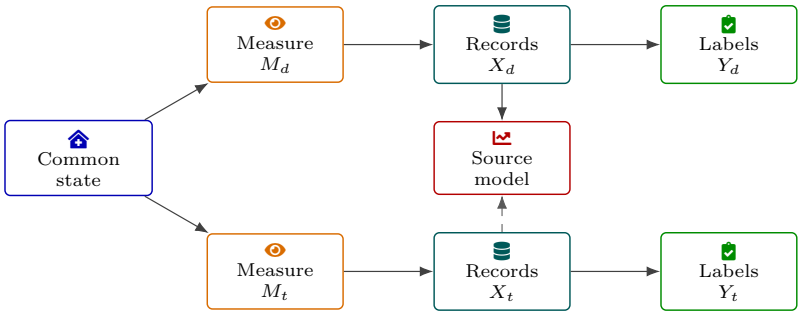

Figure 2: Transport is not a simple replay of the source task. Even with the same latent clinical process, changes in measurement and label operators alter the effective meaning of predictor space and the validity of the learned score.

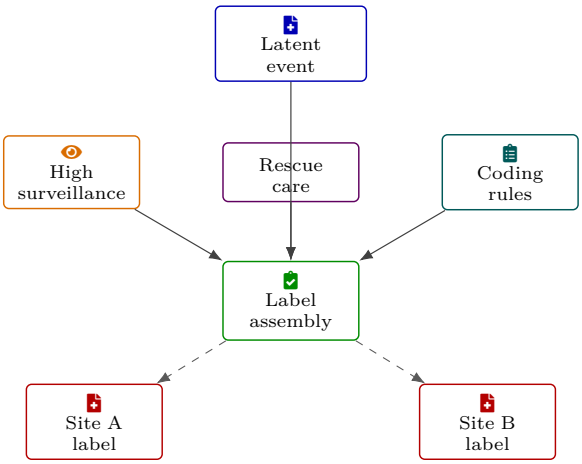

Figure 3: Observed outcomes are assembled through surveillance, intervention, and adjudication. Two institutions can face similar latent events while still produce labels with different operational meaning, making calibration drift partly a label problem rather than only a model problem.

2 Clinical Prediction as an Inverse Problem Under Measurement Operators

The usual formalization of supervised prediction begins with observed predictors X and an observed outcome Y. A model is then trained to approximate the conditional law of Y given X. This formulation is convenient, but it hides the essential translational fact that X is not primary. In clinical settings, what is primary is a latent patient state, denoted here by H, representing physiology, disease progression, anatomical context, prior treatment exposure, and other partially unobserved determinants of future events. The predictor vector X is a rendered image of that state. It is produced by an institutional measurement operator that determines which aspects of H become visible, at what times, with what fidelity, and under which patterns of missingness [4]. A clinical model therefore does not learn directly from patient state. It learns from an institutionally transformed signal.

Let D index a clinical domain. A domain is not merely a hospital identifier. It is a shorthand for a bundle of practices including referral pathways, documentation norms, ordering conventions, surveillance density, and treatment culture. The observed data generating process may be expressed as

$$\begin{aligned} X &= \mathcal{M}_D(H) + \varepsilon_x, \\ Y &= \mathcal{L}_D(H, A) + \varepsilon_y, \end{aligned} \tag{1}$$

where $\mathcal{M}_D$ is a measurement operator and $\mathcal{L}_D$ is a label operator whose output also depends on actions A taken during care. The first equation emphasizes that recorded predictors depend on how the institution observes the patient. The second emphasizes that recorded outcomes depend on what happened after observation, including institution-specific actions. A predictor trained on (X, Y) is therefore approximating a relation already shaped by local observation and intervention. The model does not recover a universal physiological law unless the relevant parts of $\mathcal{M}_D$ and $\mathcal{L}_D$ happen to be stable across deployment settings.

This interpretation becomes more concrete when one considers routine variables. A lactate value, a postoperative white cell count, a nursing note frequency, or an ICU admission flag may appear to be straightforward measurements, but each is filtered through local policy. A laboratory value is observed only if ordered. A charted

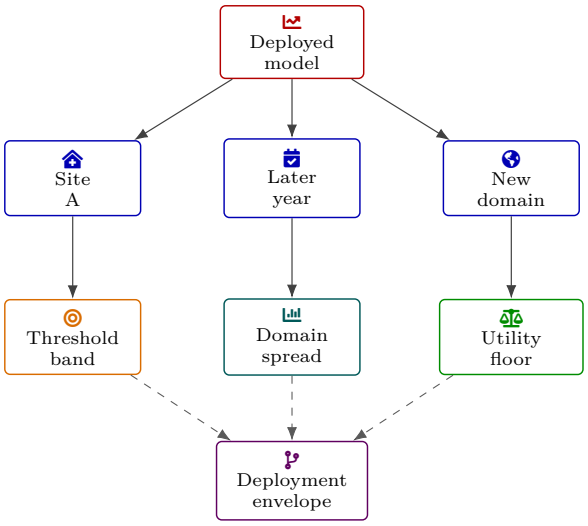

Figure 4: External validation is represented as mapping a deployment envelope rather than reporting a single external score. The relevant object is the region of domains where threshold behavior, calibration, and minimum acceptable utility remain stable enough for real use.

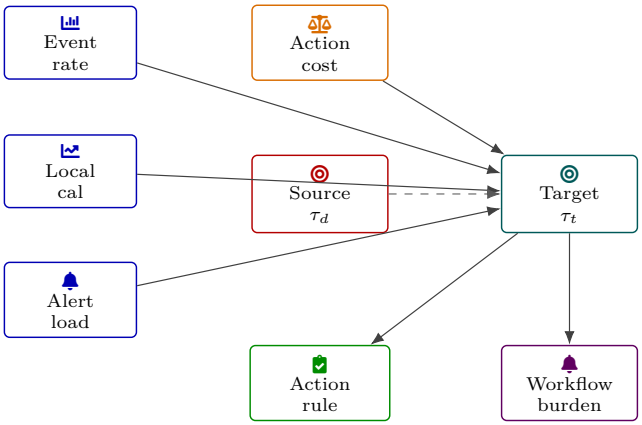

Figure 5: Threshold transfer depends on more than discrimination. Prevalence, local calibration, action cost, and alert capacity can shift the useful threshold enough that the same numerical score no longer implies the same operational policy across domains.

vital sign is recorded only if monitoring frequency and workflow permit. A transfer to intensive care reflects not only patient state but threshold practices and bed availability. Missingness can itself carry meaning because the absence of a measurement is often a decision. The observed predictor space is therefore a mixture of latent biology and local measurement choices [5]. Formally, one can decompose

$$\begin{aligned} X &= X^{\text{bio}} + X^{\text{obs}}, \\ X^{\text{obs}} &= \mathcal{O}_D(H, U), \end{aligned} \tag{2}$$

where X^{bio} denotes components that primarily reflect patient state and X^{obs} denotes components that primarily reflect observation processes, with U representing additional institutional factors. The decomposition is conceptual rather than exact, but it clarifies why a model optimized only for source-domain risk may rely heavily on observation-process features if they happen to be predictive there.

Prediction in this setting is an inverse problem because the model is asked to reconstruct future risk from a transformed and partial observation of latent state. Inverse problems are notoriously sensitive to the operator through which data are observed. If the operator changes, the same observed vector may correspond to a different region of latent state, and the same latent state may produce a different observed vector. This immediately undermines naive transport. Let the development domain be d and the target domain be t. Even if the latent disease process were stable, a change from $\mathcal{M}_d$ to $\mathcal{M}_t$ would alter the effective predictor map. A model f_d fit under domain d estimates

$$f_d \approx \Pr_d(Y = 1 \mid X) = \Pr_d\big(Y = 1 \mid \mathcal{M}_d(H)\big). \tag{3}$$

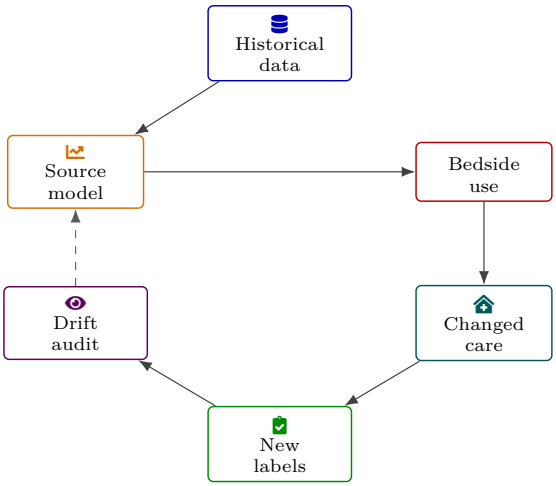

Figure 6: Deployment is a closed loop rather than a one-time test. Once the score changes observation or intervention behavior, the realized labels and calibration landscape shift, so auditing and updating become part of the scientific object rather than a post hoc maintenance step.

Table 2: Sources of Predictor Instability

Source	Mechanism	Impact
Missingness	Order-dependent measurement	Bias in features
Workflow	Recording frequency	Temporal distortion
Policy Variation	Test thresholds	Feature shift
Documentation	Coding practices	Noise injection

The transported task is

$$\Pr_t(Y = 1 \mid X) = \Pr_t\big(Y = 1 \mid \mathcal{M}_t(H)\big). \tag{4}$$

There is no reason to expect these conditional laws to coincide merely because both are written as functions of a vector called X. The conditional law itself changes when the observation operator changes, because the meaning of coordinates in predictor space shifts.

This account also explains why purely data-driven feature selection can be translationally hazardous. Suppose there exists a stable low-dimensional representation $Z = \Phi(H)$ that would support reasonable transport, but the source domain contains additional observational artifacts R_d that correlate with Y because of workflow or documentation patterns. A flexible model minimizing empirical risk will generally exploit both. One may write [6]

$$f_d(X) = g\big(Z, R_d\big), \tag{5}$$

with no built-in incentive to separate stable from unstable sources of predictive content. Internal validation will not expose the problem because R_d is genuinely reproducible inside the same domain. A cross-validated score can therefore be highly persuasive while resting on components that vanish or reverse under transport. This is not overfitting in the colloquial sense of memorizing random noise. It is overcommitment to domain-specific operator structure.

The operator perspective also recasts interpretability. Feature importance methods and coefficient summaries are often taken as evidence that a model can be understood. Yet an apparently interpretable association may still be operator-bound. A variable can be clinically plausible and still have a domain-specific predictive meaning because the conditions under which it is observed differ across sites. Interpretability in the translational sense therefore requires more than identifying influential variables. It requires understanding whether the variable's informational role is stable under changes in measurement practice. A model may be transparent and still not be portable.

One implication is that transport-oriented development should treat operator stability as a design objective. If multiple domains are available, one can seek representations that are predictive across measurement regimes rather than merely within one. Let $\phi_\theta(X)$ be a learned representation and let g_ψ map that representation to risk. A

Table 3: Forms of Label Instability

Factor	Description	Effect
Surveillance Intensity	Detection frequency	Event inflation
Coding Practice	Administrative variation	Label drift
Intervention Timing	Early treatment	Event suppression
Adjudication Rules	Chart review logic	Inconsistency

Table 4: Performance Dimensions Across Domains

Metric	Meaning	Sensitivity	Stability
Discrimination	Ranking ability	Low	Moderate
Calibration	Probability accuracy	High	Fragile
Threshold Utility	Decision value	High	Variable
Workload	Alert burden	Medium	Contextual

transport-oriented training criterion may take the form [7]

$$\begin{aligned}\min_{\theta,\psi} \quad & \sum_{j=1}^{m} \pi_j \, \hat{\mathcal{R}}_j\big(g_\psi \circ \phi_\theta\big) \\ & + \lambda \sum_{j=1}^{m} \left\| \Sigma_j^{\phi} - \bar{\Sigma}^{\phi} \right\|_F^2 \\ & + \gamma \, \Omega(\theta, \psi),\end{aligned} \tag{6}$$

where Σ_j^{ϕ} denotes the representation covariance in domain j, $\bar{\Sigma}^{\phi}$ is the average covariance across domains, and Ω is a complexity penalty. The middle term is one example of a stability regularizer discouraging representations whose geometry depends strongly on domain. The point is not that second-order alignment is sufficient, but that source-domain fit alone is inadequate when measurement operators vary.

The translational consequence is straightforward. A clinical model can be scientifically impressive in a narrow sense and still fail to estimate a transportable object. Its apparent success may derive from how one institution renders latent state into records. External validation is indispensable because it changes the operator. It asks whether the model has captured something that survives the shift from $\mathcal{M}_d$ to $\mathcal{M}_t$, or whether its success depended on an observational geometry that existed only where it was trained.

3 Label Instability and the Problem of Institutionally Induced Outcomes

Measurement instability on the predictor side is only half of the translational problem. The outcome itself is often less fixed than the notation Y suggests. Many clinically important labels are not direct observations of a physiological event. They are composites assembled through chart review, administrative coding, reintervention records, laboratory thresholds, imaging findings, or clinician adjudication. Each component depends on local surveillance intensity and local treatment pathways. The outcome that enters model training is therefore not simply an event in nature. It is an institutionally induced functional of patient state and care processes. Once this is recognized, the instability of transported prediction appears deeper than ordinary prevalence shift. The target of inference itself may be changing [8].

A useful starting point is to distinguish latent event state from observed label. Let $E \in \{0, 1\}$ denote whether the underlying clinical event has truly occurred and let $Y \in \{0, 1\}$ denote the recorded label used for modeling. The relation between them can be written as

$$Y = \mathcal{A}_D(E, S, A) + \varepsilon_\ell, \tag{7}$$

where $\mathcal{A}_D$ is an adjudication operator depending on domain D, surveillance pattern S, and actions A. Surveillance matters because some events are revealed only if looked for. Action matters because rescue interventions can prevent the event from evolving into a recorded endpoint or can create documentation pathways that make detection more likely. Under this formulation, even if the latent event process E were stable across domains, the observed label Y could still differ materially because adjudication and surveillance differ.

Table 5: Deployment Envelope Characteristics

Aspect	Description	Importance
Domain Range	Set of valid settings	Core scope
Performance Spread	Variability across sites	Risk indicator
Calibration Bounds	Acceptable deviation	Clinical meaning
Adaptation Need	Recalibration extent	Governance load

Table 6: Threshold Drift Drivers

Driver	Mechanism	Consequence
Prevalence Shift	Base rate change	Threshold shift
Calibration Error	Risk misalignment	Action error
Cost Variation	Utility differences	Policy change
Capacity Limits	Resource constraints	Alert adjustment

This issue has direct empirical relevance. In surgical prediction, a broad review noted that external validation was uncommon and that reproducibility gaps remained substantial, making it difficult to determine whether reported performance reflected durable signal or local conventions in data generation and event definition; Sadanandan (2024) treated this as a central limitation for clinical deployment rather than as a minor methodological omission [9]. The conceptual importance of that observation extends well beyond one subfield. When labels are assembled differently across institutions or time periods, external validation is not simply a harder test of the same target. It is a test of whether the target was ever sufficiently stable to support broader claims.

To see why, consider two institutions with identical latent event incidence but different surveillance intensity. In the first, postoperative imaging is ordered early and often, chart review is meticulous, and diagnostic coding is comprehensive. In the second, imaging is more selective, chart review is thinner, and coding practices are less consistent. The recorded event rate can differ sharply even if the underlying biology does not. A model trained in the first setting can appear miscalibrated in the second, not because its latent ranking of event propensity is entirely wrong, but because the label operator changed. Yet from the standpoint of deployment this distinction does not rescue the model [10]. If the meaning of a predicted 20% event risk depends on local adjudication intensity, then the score does not carry stable clinical meaning across the two settings.

Mathematically, one may write the development and target label operators as $\mathcal{A}_d$ and $\mathcal{A}_t$. The observed conditional law in each domain becomes

$$\begin{aligned} \eta_d(x) &= \Pr_d\big(Y = 1 \mid X = x\big), \\ \eta_t(x) &= \Pr_t\big(Y = 1 \mid X = x\big). \end{aligned} \tag{8}$$

Even if the latent event law $\Pr(E = 1 \mid H)$ were invariant, the functions η_d and η_t may differ through domain-specific mappings from latent event to observed label. A model that approximates η_d can therefore transport poorly because it is estimating the wrong observed target for domain t. The usual language of calibration drift may understate this. What has shifted is not only baseline incidence or slope. The very object whose probability is being estimated has changed its operational content.

This instability is particularly severe when intervention changes observed outcomes. Suppose a score is trained to predict an adverse event under historical care patterns. During deployment, clinicians respond to high-risk predictions with earlier imaging, preventive treatment, or closer monitoring. The consequence is that realized event rates among high-risk cases may fall. From one perspective the score now appears to overpredict. From another perspective it remains useful because it correctly identifies patients who would have been high risk under historical care. The trouble is that these are not the same estimand. One is the probability of the observed event under usual care [11]. The other is the probability of the event under a policy that reacts to the score. A model cannot be evaluated coherently unless it is clear which label regime it is intended to estimate.

This can be formalized through counterfactual notation. Let $Y(a)$ denote the observed label that would result under action policy a. Historical training data often identify a score for

$$\Pr\big(Y(A^{\mathrm{hist}}) = 1 \mid X\big), \tag{9}$$

where A^{hist} is the historical action pattern. After deployment, however, the policy may become A^{mdl}, inducing the target

$$\Pr\big(Y(A^{\mathrm{mdl}}) = 1 \mid X\big). \tag{10}$$

Table 7: Validation Strategies

Type	Description	Limitation
Internal	Same-domain resampling	No transport insight
External	New domain testing	Limited coverage
Silent Prospective	Real-time no action	Pre-deployment only
Live Deployment	Closed-loop use	Feedback effects

Table 8: Closed-Loop Deployment Effects

Effect	Mechanism	Outcome
Behavior Change	Clinician response	Data shift
Outcome Shift	Intervention success	Label change
Monitoring Drift	Input variation	Calibration loss
Policy Feedback	Score-action loop	System evolution

These quantities need not be equal. If clinicians intervene effectively on alerted patients, then the model's own use changes the event process. Standard retrospective validation cannot settle the resulting tension because the training labels were generated under a different policy. This is one reason why postdeployment recalibration must be interpreted carefully. A discrepancy between predicted and realized risk may reflect target drift, label drift, or beneficial policy response.

Label instability also undermines the comparability of published studies. Two models developed for nominally the same endpoint may in fact be learning different observed functionals because inclusion windows, diagnostic codes, chart review rules, and reintervention criteria differ. Reported discrimination values then do not live on a common scale. One model may look stronger not because it is closer to an invariant physiological process, but because the label in its source study was easier to predict under that institution's surveillance regime. Without explicit standardization of outcome construction, the literature accumulates performance claims that are difficult to synthesize. This is not a superficial reporting issue [12]. It is a structural reason why apparent progress in model development may fail to translate into cumulative knowledge.

The problem can be expressed through a label operator discrepancy. Let $\mathcal{L}_d$ and $\mathcal{L}_t$ map latent trajectories into recorded outcomes. Then one may define

$$\Delta_{\mathcal{L}} = \sup_{h,a} \left| \mathcal{L}_d(h,a) - \mathcal{L}_t(h,a) \right|. \tag{11}$$

If $\Delta_{\mathcal{L}}$ is nontrivial, transport of predicted probabilities becomes intrinsically difficult because the model is being asked to estimate different functions in the two domains. External validation under such conditions is still crucial, but its interpretation changes. A poor result can reveal either weakness of the learned physiological relation or instability of the label regime. This is not an argument against validation. It is an argument for designing validation studies that document outcome definitions, surveillance pathways, and intervention structures with enough detail to interpret the result.

A practical consequence follows for model design. Features closely linked to local treatment decisions should be handled cautiously when the label is also treatment-sensitive. Otherwise the predictor may learn a closed local loop in which observations and labels are mutually shaped by one institution's response policy. This can produce impressive internal fit and poor transport. A more conservative model may deliberately prefer variables whose association with latent event propensity is less entangled with downstream action. Such a model may sacrifice some source-domain discrimination while gaining stability in the meaning of the event it predicts.

In this light, external validation is not merely a test of whether a score remains numerically accurate. It is a test of whether the modeled outcome retains the same operational identity across domains [13]. If labels are institutionally induced, then portability requires evidence about label construction as much as about algorithm architecture. A model can only be clinically interpretable elsewhere if the event whose risk it estimates is recognized as the same event in a sufficiently strong empirical sense.

4 External Validation as Estimation of Deployment Envelopes

Once both predictors and outcomes are recognized as products of local operators, the standard image of external validation as a single independent test becomes inadequate. A model intended for clinical use rarely faces only one future environment. It may be deployed across multiple hospitals, across future periods within the same hospital, or across services with different workflow intensity and monitoring density. The relevant translational question is therefore not whether the model works in one external cohort, but under what range of conditions the model

Table 9: Governance Elements for Portability

Element	Purpose	Benefit
Benchmark Design	Multi-domain testing	Stability insight
Provenance Reporting	Data transparency	Reproducibility
Uncertainty Modeling	Variance estimation	Risk control
Scope Definition	Usage boundaries	Safe deployment

Table 10: Cost Structure of Translation

Cost Type	Description	Implication
Development	Model training	Local focus
External Validation	Multi-site testing	Portability evidence
Local Adaptation	Site-specific tuning	Duplication risk
Monitoring	Ongoing evaluation	Sustained reliability

remains clinically interpretable. That range may be called the deployment envelope. External validation should be understood as estimation of that envelope rather than as the computation of a solitary external metric.

Let $\mathcal{T}$ denote a family of plausible target domains. Each domain $t \in \mathcal{T}$ is characterized by a measurement operator $\mathcal{M}_t$, a label operator $\mathcal{L}_t$, and an action environment governing how predictions would be used. For a trained model f, define a domain-specific performance functional $G_t(f)$. The functional might be a vector containing calibration slope, calibration intercept, threshold-specific sensitivity, action rate, and a utility summary. The deployment envelope is the set

$$\mathcal{E}_\alpha(f) = \Big\{ t \in \mathcal{T} : G_t(f) \succeq \alpha \Big\}, \tag{12}$$

where α is a clinically acceptable lower bound interpreted componentwise. The translational problem becomes estimation of $\mathcal{E}_\alpha(f)$ or of its measure under a weighting distribution over plausible targets. A single external cohort can provide one point in this space, but it cannot establish the envelope except under unrealistically strong assumptions.

This formulation reveals why mean external performance is not enough. A model may have favorable average behavior and still possess a dangerously irregular envelope. Suppose two models have similar expected target utility, but one exhibits wide dispersion across sites and the other is more stable [14]. The second may be preferable because its operating characteristics are easier to govern. If $t_1, \dots, t_m$ are observed external domains, one may summarize mean and dispersion by

$$\begin{aligned} \bar{G}(f) &= \frac{1}{m} \sum_{j=1}^{m} G_{t_j}(f), \\ V_G(f) &= \frac{1}{m-1} \sum_{j=1}^{m} \big\| G_{t_j}(f) - \bar{G}(f) \big\|_2^2 . \end{aligned} \tag{13}$$

The quantity $V_G(f)$ is not a mere secondary statistic. It describes how sharply the envelope bends across clinically relevant domains. A model with slightly weaker average performance but a flatter envelope may be more suitable for broad deployment than a model whose external behavior is highly variable.

Calibration is a central component of the envelope because clinical interpretation often depends more on probability meaning than on ranking alone. Let $S = f(X)$ be the model score. In domain t, the calibration map can be represented as

$$m_t(s) = \Pr_t\big(Y = 1 \mid S = s\big). \tag{14}$$

For deployment purposes one is usually interested not in the full function alone but in whether it remains close to an acceptable family. A simple parametric approximation writes

$$\operatorname{logit} m_t(s) = \alpha_t + \beta_t \operatorname{logit}(s) + r_t(s), \tag{15}$$

where α_t is an intercept displacement, β_t a scale distortion, and $r_t(s)$ a residual nonlinearity. The deployment envelope can then be characterized through the set of domains where $|\alpha_t|$, $|\beta_t - 1|$, and $\|r_t\|$ remain small enough for the intended clinical use. This description is already more informative than an external c-statistic because it speaks directly to whether the same numeric score carries approximately the same meaning in different environments.

A second virtue of the envelope view is that it makes explicit the distinction between pure transport and transport after adaptation. Let $f^{(0)}$ be the source model. In a target domain t, one may define an adapted family [15]

$$f_t^{(1)} = \text{logit}^{-1}\big(\alpha_t + \text{logit}(f^{(0)})\big),$$
$$f_t^{(2)} = \text{logit}^{-1}\big(\alpha_t + \beta_t \,\text{logit}(f^{(0)})\big), \tag{16}$$

corresponding to intercept-only and intercept-plus-slope recalibration. The deployment envelope for $f^{(0)}$ without adaptation is different from the envelope for the class of locally adapted models. A model whose raw envelope is narrow but whose recalibrated envelope is broad may still be useful, but it imposes a different governance burden from a model that travels with little correction. Explicitly estimating both envelopes avoids the common confusion whereby a model is described as externally validated even though substantial local updating was required.

Estimating deployment envelopes also changes sample size logic. In a conventional view, an external dataset is adequate if it yields a sufficiently precise point estimate for a chosen metric. Under an envelope view, one needs enough information to estimate not only the center of target behavior but also its spread and its threshold-specific stability. This often implies larger event counts and more diverse sites than are typical in the current literature. A small external cohort may be enough to suggest that gross failure has not occurred, but not enough to delimit where the score remains dependable. When deployment decisions are made on such thin evidence, the envelope is implicitly guessed rather than estimated.

A formal risk envelope can be defined through acceptable loss. Let $\mathcal{R}_t(f)$ be a target-domain loss. Then for tolerance τ one may define

$$\mathcal{E}_\tau^{\mathcal{R}}(f) = \Big\{t \in \mathcal{T} : \mathcal{R}_t(f) \le \tau\Big\}. \tag{17}$$

For multiple relevant losses, including calibration loss and decision regret, the envelope becomes the intersection of corresponding acceptable sets. The translational question is then whether the intended deployment region lies inside this intersection. External validation studies can be designed to sample representative points from the anticipated deployment region, thereby yielding empirical evidence about the envelope's extent rather than merely producing an isolated performance certificate.

This perspective also makes room for negative results as substantive findings. If a model validates in tertiary referral centers but not in lower-acuity community settings, the observed envelope is not empty [16]. It is simply narrower than perhaps hoped. Such a finding can still be operationally useful because it specifies where the score should and should not be trusted. A literature focused on deployment envelopes would therefore treat heterogeneity in external behavior as a first-class scientific output, not as an inconvenience to be averaged away.

The broader implication is that external validation should be designed as a mapping exercise. The goal is to chart the region of clinical space in which a score preserves interpretability, calibration, and acceptable utility. A single external cohort can contribute to that map, but it cannot by itself stand in for the map. Translation requires the map because bedside use always occurs in some domain, and domains differ in precisely the features that current retrospective studies too often treat as background conditions.

5 Threshold Transfer Geometry and the Drift of Clinical Action

A risk model becomes clinically consequential only when connected to an action rule. That rule may be explicit, as when a score above a threshold triggers additional imaging, consultation, or preventive treatment, or implicit, as when clinicians give more attention to high-risk cases. The translational problem is therefore not exhausted by whether risk estimates remain statistically respectable. The decisive issue is whether the same scores support the same actions in new environments. Threshold transfer is often assumed to be easy once external performance is acceptable, but this assumption is fragile. Small calibration shifts, prevalence differences, or workload constraints can move the optimal threshold enough to change policy meaning substantially.

Let $f(X) \in [0,1]$ be a score and let action occur when $f(X) \ge \tau$. In domain t, define the action rate [17]

$$W_t(\tau) = \Pr_t\big(f(X) \ge \tau\big), \tag{18}$$

and define threshold-specific true-positive and false-positive rates by

$$\text{TPR}_t(\tau) = \Pr_t\big(f(X) \ge \tau \mid Y = 1\big),$$
$$\text{FPR}_t(\tau) = \Pr_t\big(f(X) \ge \tau \mid Y = 0\big). \tag{19}$$

Suppose the local value of intervention is summarized by benefit b_t for a correct action on a case, burden c_t for unnecessary action on a noncase, and loss h_t for missing an actionable case. Then expected utility at threshold τ may be written as

$$\begin{aligned} U_t(\tau) = b_t \Pr_t\big(f(X) \geq \tau, Y = 1\big) \\ - c_t \Pr_t\big(f(X) \geq \tau, Y = 0\big) \\ - h_t \Pr_t\big(f(X) < \tau, Y = 1\big). \end{aligned} \tag{20}$$

The source-domain optimal threshold,

$$\tau_d^* = \arg\max_{\tau \in [0,1]} U_d(\tau), \tag{21}$$

has no intrinsic reason to remain optimal in a new domain. Calibration may shift, prevalence may change, and the cost of action may differ because resources, staffing, or downstream workflow differ. When the same numeric threshold is transplanted without re-estimation, what travels is an assumption about score meaning, not an empirically justified policy.

The geometry of this drift can be made more precise. Consider a target domain with calibration map $m_t(s)$. The expected utility may be rewritten as

$$\begin{aligned} U_t(\tau) = \int_\tau^1 \Big(b_t\, m_t(s) - c_t\,\big(1 - m_t(s)\big)\Big)\, dF_t(s) \\ - h_t \int_0^\tau m_t(s)\, dF_t(s), \end{aligned} \tag{22}$$

where F_t is the score distribution. Both m_t and F_t can differ across domains. Thus threshold drift is driven by at least two mechanisms: a shift in the clinical meaning of a given score and a shift in how often that score occurs. A threshold that looked reasonable in the development environment may generate an unmanageable alert burden elsewhere even if its discrimination remains intact. Conversely, a threshold tuned to maintain a fixed alert burden may operate at a very different clinical risk level across sites. These are not marginal implementation details [18]. They determine whether the model is usable.

A useful local approximation arises from the first-order condition for optimality. If U_t is differentiable, then an interior optimum τ_t^* satisfies

$$\begin{aligned} 0 &= \frac{d}{d\tau} U_t(\tau_t^*) \\ &= -\Big(b_t\, m_t(\tau_t^*) - c_t\,\big(1 - m_t(\tau_t^*)\big)\Big) f_t(\tau_t^*) \\ &\quad + h_t\, m_t(\tau_t^*) f_t(\tau_t^*), \end{aligned} \tag{23}$$

where f_t is the score density. Assuming $f_t(\tau_t^*) > 0$, the effective risk threshold satisfies

$$m_t(\tau_t^*) = \frac{c_t}{b_t + c_t + h_t}. \tag{24}$$

This expression is revealing. The clinically appropriate decision threshold is defined in terms of target-domain calibrated risk and target-domain utility parameters, not in terms of the raw score produced in development. If calibration changes, the same raw threshold no longer corresponds to the same effective risk threshold. If costs change, even a perfectly calibrated score requires a new raw threshold. Threshold transfer therefore fails whenever the score-to-risk map or the risk-to-action tradeoff changes. Both are common in cross-domain deployment.

The most consequential errors often concentrate near decision boundaries. Let $r_t(X)$ denote the target-domain calibrated risk and define a boundary band

$$\mathcal{B}_t(\tau, \epsilon) = \big\{x : |r_t(x) - r_t(\tau)| \leq \epsilon\big\}, \tag{25}$$

where $r_t(\tau)$ is shorthand for the risk level corresponding to raw threshold τ. Errors inside $\mathcal{B}_t(\tau, \epsilon)$ have disproportionate policy impact because they flip action. One may define threshold regret as

$$\Gamma_t(\tau) = \mathbb{E}_{P_t}\Big[\Delta u(X, Y)\, \mathbf{1}\big\{(r_t(X) - r_t(\tau))(f(X) - \tau) < 0\big\}\Big], \tag{26}$$

where $\Delta u(X, Y)$ is the utility gap between correct and incorrect action. A model can preserve respectable global ranking and still incur large $\Gamma_t(\tau)$ if calibration error is concentrated near the action boundary [19]. This is why

external validation focused only on global metrics is incomplete. What matters clinically is the stability of the action geometry around the thresholds that institutions can plausibly use.

Capacity constraints add another layer. Many deployments do not optimize unconstrained utility. They operate under resource limits, such as the number of charts that can be reviewed daily or the number of additional scans that can be performed. The relevant threshold then solves a constrained problem:

$$\tau_t^{\mathrm{cap}} = \arg\max_{\tau\in[0,1]} U_t(\tau) \quad \text{subject to} \quad W_t(\tau) \le \kappa_t, \tag{27}$$

where κ_t is local capacity. A threshold calibrated in the development environment can be badly misaligned with κ_t in the target environment. Hospitals often respond informally by moving the threshold until workload becomes tolerable. That operational improvisation changes the model's effective sensitivity and utility, often without documentation. From a translational standpoint, the model has not simply been transported. It has been redefined on the fly.

A practical way to stabilize threshold transfer is to validate not only the score but the score-action map. External studies should therefore report workload curves, case-capture curves, and utility estimates over threshold ranges tied to realistic capacities. More ambitiously, they should estimate threshold transport functions mapping development thresholds to target-domain equivalents. If τ_d is a source threshold associated with effective risk q, then its target-domain analogue solves [20]

$$m_t(\tau_t) = q. \tag{28}$$

Such mappings reveal whether clinical action meaning can be preserved by simple recalibration or whether deeper site-specific redesign is necessary.

The broader point is that external validation must interrogate the geometry of decisions, not only the geometry of prediction. A model is translationally successful when it supports approximately the same action logic under changed conditions. Without explicit threshold analysis, a score may appear externally respectable while quietly ceasing to guide practice in a coherent way.

6 Deployment as Closed-Loop System Identification

Retrospective prediction treats the environment as if it were fixed. Clinical deployment destroys that assumption. Once a model is introduced into practice, its outputs influence clinician attention, monitoring frequency, escalation patterns, and treatment timing. The model becomes part of the system that generates future data. Translation must therefore be understood as a closed-loop problem, more akin to system identification under intervention than to static function approximation. External validation is essential because it probes whether the model survives movement across environments, but even that is not the end of the matter. A model that transports well in silent mode may behave differently once it begins to shape care.

To formalize this, let H_k denote the latent clinical state at time step k, X_k the observed predictors, $S_k = f(X_k)$ the model output, and A_k the action taken in response to all available information including the score. The state dynamics may be written as

$$\begin{aligned} H_{k+1} &= F_k H_k + B_k A_k + \omega_k, \\ X_k &= C_k H_k + \nu_k, \\ A_k &= \pi_k\big(X_k, S_k, R_k\big), \end{aligned} \tag{29}$$

where F_k captures disease progression, B_k maps action into state change, C_k maps latent state into observations, and π_k is the clinician or institutional response policy, possibly depending on resource state R_k. Once the score enters π_k, it changes the future distribution of both latent state and observed data [21]. The data seen after deployment are therefore not generated by the same mechanism as the data used for development. A model can invalidate its own historical calibration precisely by being used successfully.

This is one reason silent prospective validation is often the necessary bridge between external validation and live deployment. In silent mode the model runs on the production data stream without influencing action. This reveals whether the data pipeline, variable timing, and missingness semantics in the operational environment match those assumed during development. It also allows estimation of target-domain calibration and action-rate implications under real-time constraints. If large discrepancies appear here, they are evidence of operator mismatch before the closed loop has even begun. Many retrospective systems fail at this stage because features that seemed available in research extracts prove delayed, partially backfilled, or differently encoded in live workflow.

Once action is linked to the score, monitoring must move beyond aggregate retrospective metrics. What should be tracked are dynamic summaries of calibration, workload, subgroup behavior, and input support. Let g_k denote

an operational statistic at time k, such as calibration intercept, calibration slope, alert fraction, or missing-feature rate. A monitoring process can be defined by

$$Z_K = \sum_{k=1}^{K} w_k \left(g_k - g_{\text{ref}}\right), \tag{30}$$

where g_{ref} is a reference estimate from silent validation and w_k allows recent periods to matter more. A persistent excursion of Z_K beyond tolerance suggests that the deployed system has drifted from the envelope within which the score was originally judged interpretable. Crucially, the monitored quantities should include subgroup-specific values, because closed-loop effects often appear unevenly across patient strata or service lines.

The closed-loop perspective also changes the meaning of postdeployment recalibration. In a static world, recalibration is a correction for mismatch between predicted and observed risk [22]. In a closed-loop world, mismatch may reflect beneficial intervention, adverse drift, changed label semantics, or some mixture. Suppose realized event rates drop among high-score patients after deployment. An intercept correction might restore apparent calibration, but it could also erase a meaningful signal if the drop was caused by successful response to the alert. The right interpretation depends on the estimand. Is the system intended to estimate untreated risk, realized risk under current policy, or risk conditional on no incremental action beyond baseline care? Without clarity on this point, recalibration becomes a statistical patch detached from the policy mechanism generating the data.

A more disciplined approach treats deployment as ongoing system identification. The objective is not merely to preserve numerical agreement between predictions and outcomes, but to understand how the score interacts with local response. One can represent the effect of deployment through a policy perturbation:

$$\pi_k = \pi_k^0 + \Delta\pi_k(S_k), \tag{31}$$

where π_k^0 is the predeployment response policy and $\Delta\pi_k$ is the additional behavior induced by seeing the score. If $\Delta\pi_k$ is large, then the outcome process under deployment may differ substantially from historical training labels. The model must then be evaluated as part of a new decision system, not simply as a transported predictor. This is why prospective implementation studies are qualitatively different from retrospective validation, even when both use the same nominal endpoint.

The system-identification frame also clarifies the role of updating. Suppose the deployed model is periodically recalibrated or partially refit. Each update changes the closed loop, since clinicians respond to the updated score distribution and the updated thresholding policy. Aggressive adaptive updating can improve short-term fit while making longitudinal interpretation and governance more difficult [23]. Conservative updating preserves continuity but may allow drift to accumulate. A principled compromise is to separate lightweight updates that preserve model form from structural updates that alter its meaning. Intercept adjustment and threshold revision can often be justified with relatively modest evidence. Changes to representation, feature set, or coefficient structure require stronger evidence because they produce a new instrument, not merely a better-calibrated version of the old one.

One may formalize a guarded update policy as

$$\begin{aligned} \theta_{k+1} &= \theta_k + \eta_k \, \Gamma_k, \\ \eta_k &= 0 \quad \text{if} \quad \mathcal{C}_k < c_{\min}, \end{aligned} \tag{32}$$

where θ_k are current model parameters, Γ_k is a candidate update direction, and $\mathcal{C}_k$ is a governance confidence measure that must exceed a minimum threshold before updating occurs. The exact form of $\mathcal{C}_k$ can vary, but the principle is consistent: model evolution in a closed loop should be evidence-driven and policy-aware rather than automatic.

Ultimately, deployment is not the endpoint of translation but its most demanding phase. A score that has passed external validation has shown that it can survive at least some operator changes while the loop remains open. Once live use begins, the loop closes and the environment becomes endogenous to the score. Translation succeeds only if the model's meaning remains controlled under this feedback. That requires silent validation, dynamic monitoring, careful interpretation of recalibration, and disciplined updating. These are not add-ons to external validation. They are the natural continuation of the same translational logic.

7 Governance, Benchmark Structure, and the Economics of Portable Evidence

The scientific problem of portability is inseparable from the institutional problem of how evidence is organized. A field can accumulate many models and still make little translational progress if each model is evaluated under incomparable definitions, incomplete reporting, and isolated local datasets. What appears as a sequence of technical

innovations may then amount to a weakly connected collection of local claims. Portability requires an evidentiary infrastructure capable of distinguishing stable gains from context-bound artifacts [24]. In that sense, governance is not external to the science of generalization. It determines what kinds of generalization can be observed and trusted.

A first requirement is benchmark structure aligned to domain heterogeneity rather than to leaderboard simplicity. Many published benchmarks implicitly reward source-domain discrimination because they rely on random partitions within a single institutional dataset. Such benchmarks are useful for local model comparison but reveal little about the size or shape of a deployment envelope. A transport-oriented benchmark should instead index performance by domain. Let $\mathcal{D}_{\text{ext}} = \{t_1, \dots, t_m\}$ be a set of external domains and let each candidate model f be summarized by

$$B(f) = \Big(\bar{C}(f), V_C(f), \bar{K}(f), V_K(f), \bar{\Gamma}(f), V_\Gamma(f)\Big), \tag{33}$$

where $\bar{C}$ and V_C are mean and variance of calibration-oriented performance, $\bar{K}$ and V_K summarize workload or capacity burden, and $\bar{\Gamma}$ and V_Γ summarize threshold regret across domains. A benchmark designed around $B(f)$ changes incentives. It rewards models that are not merely sharp in one place, but stable under the kinds of heterogeneity that actual deployment entails.

A second requirement is provenance-rich reporting. Because both predictors and outcomes are operator-bound, external performance is only interpretable when data provenance and preprocessing are documented in detail. This includes timing definitions, feature construction windows, missingness handling, outcome logic, and any site-specific mapping rules. Without such detail, a failed external validation may reflect genuine nontransportability or merely inconsistent reconstruction of inputs and labels. Likewise, a successful external validation may owe part of its success to hidden harmonization choices that cannot be reproduced elsewhere. Governance documents should therefore treat cohort construction and variable semantics as part of the model, not as mundane preprocessing details.

A third requirement concerns statistical uncertainty [25]. Portable evidence is not generated by point estimates alone. Confidence intervals around discrimination are not enough when calibration, threshold behavior, and workload variation across sites are central. Domain-wise uncertainty should therefore be estimated and carried into any pooled summary. If g_j denotes a domain-specific operational metric, a hierarchical structure such as

$$\begin{aligned} g_j &= \mu_g + b_j + \varepsilon_j, \\ b_j &\sim \mathcal{N}(0, \tau_g^2), \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma_j^2), \end{aligned} \tag{34}$$

is often more informative than a single pooled estimate because it separates average behavior from between-domain instability. A model with modest μ_g and small τ_g^2 may be governance-friendly in ways that a slightly stronger but highly variable model is not. In high-stakes deployment, uncertainty about where performance will land is itself part of the risk.

The economics of portability also deserve explicit treatment. External validation, multi-site benchmarking, and prospective monitoring are expensive relative to another round of retrospective training on a local dataset. This cost asymmetry partly explains why the literature contains many more models than validated deployment candidates. Yet the apparent economy of local development is misleading. If portable evidence is not generated early, the cost is merely deferred to the adoption stage, where each institution must perform its own recalibration, threshold tuning, workflow testing, and governance review. The system-level consequence is duplication of effort and uneven quality. One can express total translational cost for a model family as

$$\mathcal{K}_{\text{tot}} = \mathcal{K}_{\text{dev}} + \mathcal{K}_{\text{ext}} + \sum_{j=1}^{m} \mathcal{K}_{\text{local},j}, \tag{35}$$

where $\mathcal{K}_{\text{dev}}$ is development cost, $\mathcal{K}_{\text{ext}}$ is coordinated external validation cost, and $\mathcal{K}_{\text{local},j}$ is site-specific adaptation and governance cost. Underinvestment in $\mathcal{K}_{\text{ext}}$ often causes the local terms to balloon. A field that values translation should therefore shift some effort from local model proliferation toward shared infrastructure for portable evidence.

Governance must also define scope of use explicitly [26]. A model validated only in tertiary centers should not be described as a general clinical predictor without qualification. A model requiring local recalibration before use should be labeled accordingly. A model whose threshold policy was evaluated only under generous resource assumptions should not be promoted for settings where capacity is constrained. These are not public-relations niceties. They are necessary boundaries on the meaning of the score. In this sense, deployment documentation should function less like marketing copy and more like an instrument specification. It should state where the model has been evaluated, what adaptations are permitted, what level of drift triggers review, and what forms of action the score was designed to support.

An additional governance issue concerns subgroup performance. Domain heterogeneity is not only geographic or temporal. It also includes clinically and socially meaningful subpopulations. A model may appear externally stable on average while producing very different workload or threshold regret across groups. Because these differences can interact with existing care disparities, portability claims should include subgroup-conditioned summaries wherever feasible. The purpose is not merely fairness signaling. It is to determine whether the model's deployment envelope has hidden fractures inside domains as well as between them.

Finally, evidence governance should acknowledge that nondeployment is sometimes the correct outcome of validation [27]. A score that cannot maintain stable meaning across the settings in which it is intended to operate is not a near-miss awaiting minor improvement. It may be estimating a fundamentally local relation. Recognizing this early can save substantial downstream cost and prevent misleading clinical uptake. A benchmark and governance structure that permits clear negative conclusions is therefore a scientific asset, not a barrier to innovation. Portable evidence advances when the field becomes better at ruling out unjustified transport claims, not only at generating new local predictors.

The broader lesson is that the translational barrier posed by poor generalizability is partly methodological and partly institutional. Better modeling alone cannot solve it if benchmark design, reporting norms, and deployment governance remain oriented toward source-domain success. A field that wants portable clinical models must build a portable evidence economy in which external validation is rewarded, operator provenance is explicit, and the cost of weak transport claims is made visible rather than absorbed downstream.

8 Conclusion

Clinical prediction models do not emerge from neutral data landscapes. They are trained on records generated by specific measurement practices, intervention pathways, and outcome-construction rules. This paper has argued that the central translational barrier follows directly from that fact. A model that performs well in retrospective development may still be estimating a relation tied to one institutional measurement operator and one label regime. When such a model is moved elsewhere, its score can lose calibration, its action thresholds can drift, and even the event whose risk it purports to estimate can shift in meaning. The main problem is therefore not only that external performance may be lower. It is that the clinical interpretation of the score may no longer be stable.

Reframing the problem through measurement operators, label instability, deployment envelopes, and closed-loop deployment clarifies why external validation occupies a unique role [28]. Internal validation can estimate optimism inside a fixed data-generating regime. External validation changes the regime. It tests whether the model has captured something that remains interpretable under altered observation, altered labeling, and altered care structure. When designed as estimation of deployment envelopes rather than as a single external score, validation reveals not only average transport but the range of domains within which the model can be used responsibly. When linked to threshold analysis, it shows whether the model supports the same decisions across settings rather than merely producing similar rankings.

The implications are practical as well as conceptual. Development should favor features and representations that are less entangled with local measurement artifacts. Outcome construction should be treated as a central modeling issue rather than a background technicality. Benchmarking should be organized around domain-wise calibration, threshold regret, and workload stability rather than source-domain discrimination alone. Deployment should proceed through silent validation, dynamic monitoring, and disciplined updating because the act of using a model changes the environment that generated its training data. Governance should state clearly where the score has been shown to preserve meaning and what local adaptation is permitted before it becomes a different instrument.

None of these points deny the utility of local prediction. A model can be clinically helpful inside the system that produced it even if its broader portability remains unproven. The stronger claim is narrower and more consequential. Whenever the intended use of a model extends beyond its source environment, generalizability becomes the central translational problem, and external validation becomes the empirical procedure through which that problem is confronted rather than assumed away. Without that step, a predictive score may remain technically impressive yet clinically provincial [29].

References

[1] C. Coco, V. Tondolo, L. E. Amodio, D. P. Pafundi, F. Marzi, and G. Rizzo, "Role and morbidity of protective ileostomy after anterior resection for rectal cancer: One centre experience and review of literature.," *Journal of clinical medicine*, vol. 12, no. 23, pp. 7229–7229, Nov. 22, 2023. DOI: 10.3390/jcm12237229

[2] S. L. Due, D. Wattchow, J. L. Sweeney, L. Milliken, and C. Luke, "Colorectal cancer surgery 2000–2008: Evaluation of a prospective database," *ANZ journal of surgery*, vol. 82, no. 6, pp. 412–419, Apr. 26, 2012. DOI: 10.1111/j.1445-2197.2012.06078.x

[3] F. Jin, X. Xie, X. Li, and G. Zhu, "Meta-analysis on transanal tube drainage for the prevention of anastomotic leakage after low anterior resection for rectal cancer," *Chinese Journal of General Surgery*, vol. 31, no. 11, pp. 947–950, Nov. 25, 2016.

[4] Z.-y. Zhang, Z. Zhu, Y. Zhang, L. Ni, and B. Lu, *A nomogram for predicting feasibility of laparoscopic anterior resection with trans-rectal specimen extraction (noses) in patients with upper rectal cancer*, Mar. 2, 2021. DOI: 10.21203/rs.3.rs-256283/v1

[5] S. Giacopuzzi, J. Weindelmayer, and G. de Manzoni, "Ramie: Tradition drives innovation-feasibility of a robotic-assisted intra-thoracic anastomosis.," *Updates in surgery*, vol. 73, no. 3, pp. 1–6, Nov. 27, 2020. DOI: 10.1007/s13304-020-00932-1

[6] G. Branagan and N. Davies, "Early impact of centralization of oesophageal cancer surgery services.," *The British journal of surgery*, vol. 91, no. 12, pp. 1630–1632, Oct. 29, 2004. DOI: 10.1002/bjs.4753

[7] F. Delaunay et al., "Analysis of malpractice claims: The franco-belgian "cœlio club" experience.," *Journal of visceral surgery*, vol. 156, S33–S39, Jul. 11, 2019. DOI: 10.1016/j.jviscsurg.2019.04.011

[8] R. Makhija, P. Tsai, and A. N. Kingsnorth, "Pylorus-preserving pancreatoduodenectomy with billroth i type reconstruction: A viable option for pancreatic head resection," *Journal of hepato-biliary-pancreatic surgery*, vol. 9, no. 5, pp. 614–619, Nov. 1, 2002. DOI: 10.1007/s005340200083

[9] 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.

[10] X. Cui, Y. He, L. Chen, Y. Lin, J. Zhang, and C. Zhou, "The value of thoracic lavage in the treatment of anastomotic leakage after surgery for type iii esophageal atresia.," *Medical science monitor : international medical journal of experimental and clinical research*, vol. 26, e919962–, Mar. 11, 2020. DOI: 10.12659/msm.919962

[11] A. G. Casson, K. Madani, S. Mann, R. Zhao, B. Reeder, and H. J. Lim, "Does previous fundoplication alter the surgical approach to esophageal adenocarcinoma," *European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery*, vol. 34, no. 5, pp. 1097–1102, Sep. 6, 2008. DOI: 10.1016/j.ejcts.2008.07.059

[12] G. M. Jung et al., "Novel endoscopic stent for anastomotic leaks after total gastrectomy using an anchoring thread and fully covering thick membrane: Prevention of embedding and migration," *Journal of gastric cancer*, vol. 18, no. 1, pp. 37–47, Feb. 20, 2018. DOI: 10.5230/jgc.2018.18.e2

[13] J. W. T. Toh et al., "Association of mechanical bowel preparation and oral antibiotics before elective colorectal surgery with surgical site infection: A network meta-analysis," *JAMA network open*, vol. 1, no. 6, e183226–, Oct. 5, 2018. DOI: 10.1001/jamanetworkopen.2018.3226

[14] M. A. Abozeid, A. A. Abou-Zeid, H. M. Hussein, and M. A. Omar, "Totally versus planned conversion laparoscopic colectomy for colorectal cancers: The beginning of the learning curve," *SVU-International Journal of Medical Sciences*, vol. 4, no. 2, pp. 240–245, Aug. 1, 2021. DOI: 10.21608/svuijm.2021.83691.1189

[15] A. Bilal et al., "Six months follow up study of two hundred cases of transthoracic oesophagectomy," *Journal of Postgraduate Medical Institute*, vol. 19, no. 4, Jul. 25, 2011.

[16] T. A. Bowles, K. M. Sanders, M. E. Colson, and D. A. K. Watters, "Simplified risk stratification in elective colorectal surgery.," *ANZ journal of surgery*, vol. 78, no. 1-2, pp. 24–27, Jan. 7, 2008. DOI: 10.1111/j.1445-2197.2007.04351.x

[17] L. C. Gomes et al., "Endoscopic vacuum therapy for esophageal anastomotic leaks – data from a tertiary hospital," *Endoscopy*, vol. 55, no. S 02, S284–S284, Apr. 14, 2023. DOI: 10.1055/s-0043-1765784

[18] F. Puccetti, B. P. L. Wijnhoven, M. Kuppusamy, M. Hubka, and D. E. Low, "Impact of standardized clinical pathways on esophagectomy: A systematic review and meta-analysis.," *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, vol. 35, no. 2, Apr. 30, 2021. DOI: 10.1093/dote/doab027

[19] Y. W. Min et al., "Endoscopic vacuum therapy for postoperative esophageal leak.," *BMC surgery*, vol. 19, no. 1, pp. 37–37, Apr. 11, 2019. DOI: 10.1186/s12893-019-0497-5

[20] A. D'Hoore et al., "Compres : A prospective postmarketing evaluation of the compression anastomosis ring car 27/colonring," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 17, no. 6, pp. 522–529, May 20, 2015. DOI: 10.1111/codi.12884

[21] I. Pasternak, M. Dietrich, R. J. Woodman, U. Metzger, D. Wattchow, and U. Zingg, "Use of severity classification systems in the surgical decision-making process in emergency laparotomy for perforated diverticulitis.," *International journal of colorectal disease*, vol. 25, no. 4, pp. 463–470, Nov. 29, 2009. DOI: 10.1007/s00384-009-0852-6

[22] J. P. S. Guzman, L. L. Resurreccion, M. L. R. Suntay, and R. G. Bernaldez, "Comparison between hepaticojejunostomy and hepaticoduodenostomy after excision of choledochal cyst in children: A cohort study," *World Journal of Pediatric Surgery*, vol. 2, no. 2, e000029–, Jun. 4, 2019. DOI: 10.1136/wjps-2018-000029

[23] P. Albers, S. Schäfers, H. Löhmer, and P. de Geeter, "Seminal vesicle-sparing perineal radical prostatectomy improves early functional results in patients with low-risk prostate cancer.," *BJU international*, vol. 100, no. 5, pp. 1050–1054, Aug. 30, 2007. DOI: 10.1111/j.1464-410x.2007.07123.x

[24] J. Verhulst, "Hyperthermic intraperitoneal chemoperfusion with high dose oxaliplatin: Influence of perfusion temperature on postoperative outcome and survival," *F1000Research*, vol. 2, pp. 179–179, Sep. 3, 2013. DOI: 10.12688/f1000research.2-179.v1

[25] A. M. Brown et al., "A standardized comparison of peri-operative complications after minimally invasive esophagectomy: Ivor lewis versus mckeown," *Surgical endoscopy*, vol. 32, no. 1, pp. 204–211, Jun. 22, 2017. DOI: 10.1007/s00464-017-5660-4

[26] B. Yildiz and M. Tez, "Shift in paradigm of clinical management of anastomotic leak," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 19, no. 10, pp. 943–943, Oct. 3, 2017. DOI: 10.1111/codi.13734

[27] S. Kirmiz et al., "Regular vs. selective use of closed suction drains following robot-assisted radical prostatectomy: Results from a regional quality improvement collaborative.," *Prostate cancer and prostatic diseases*, vol. 23, no. 1, pp. 151–159, Aug. 29, 2019. DOI: 10.1038/s41391-019-0170-1

[28] K. Arunsenthilnathan, *A comparative study of efficacy of staplers versus hand sewn anastomosis in bowel surgeries*, May 1, 2019.

[29] B. Suter et al., "Surgical sealant with integrated shape-morphing dual modality ultrasound and computed tomography sensors for gastric leak detection.," *Advanced science (Weinheim, Baden-Wurttemberg, Germany)*, vol. 10, no. 23, e2301207–e2301207, Jun. 5, 2023. DOI: 10.1002/advs.202301207