OPEN ACCESS

External Validation as Multi-Environment Falsification for Clinically Transportable Prediction under Adaptive Hospital Systems

Rizky Saputra,[1] Dimas Hidayat[2] Arief Nugroho[3]

[1] Faculty of Public Health, Universitas Teuku Umar, Jalan Alue Peunyareng, Meureubo, Aceh Barat, Aceh, Indonesia;
[2] Department of Health Informatics, Universitas Nahdlatul Ulama Surabaya, Jalan Raya Jemursari No. 51–57, Surabaya, Jawa Timur, Indonesia;
[3] Department of Biomedical Engineering, Universitas Widya Husada Semarang, Jalan Subali Raya No. 12, Pedurungan, Semarang, Jawa Tengah, Indonesia;

Abstract

Hospitals now generate dense digital traces from perioperative care, ward monitoring, laboratory information systems, imaging workflows, discharge coding, and longitudinal follow-up platforms. These resources have made it possible to develop increasingly sophisticated models for diagnosis, prognosis, triage, and complication surveillance. At the same time, healthcare delivery has become more heterogeneous rather than less so. Institutions differ in admission pathways, documentation intensity, measurement cadence, intervention timing, escalation norms, and outcome ascertainment. This paper examines the consequences of that heterogeneity for translation. The central claim is that generalizability becomes a translational bottleneck because retrospective model development rarely identifies whether the learned risk surface reflects durable clinical regularity or context-bound structure generated by local care processes. The analysis develops a framework in which transportable prediction is treated as an identifiability problem under adaptive clinical environments. In this setting, internal validation is shown to be non-falsifying with respect to contextual overfitting, because repeated sampling within one environment does not test whether the score survives altered observation, altered treatment response, or altered labeling. External validation is therefore reinterpreted as multi-environment falsification rather than a routine add-on. The paper further introduces shadow deployment as an intermediate evidentiary stage between retrospective modeling and action-linked implementation, and it argues that translationally credible prediction requires environment-specific calibration analysis, threshold stress testing, and sequential monitoring under live data streams. The result is a stricter but more operationally meaningful account of what it takes for a clinical model to become usable outside the site that created it.

1. Introduction

The contemporary landscape of clinical machine learning presents an unusual contrast between local technical success and uncertain translational value [1]. On the one hand, there is no shortage of models that perform well in retrospective studies. Electronic health records, waveform repositories, surgical registries, and integrated imaging databases make it increasingly straightforward to extract complex predictor sets and to train flexible estimators that achieve appealing levels of discrimination in development cohorts. On the other hand, only a small fraction of those models appear to survive the journey from a retrospective report to a stable instrument in routine care. The usual explanation appeals to implementation friction, clinician skepticism, information-technology constraints, or the conservatism of regulatory pathways. Those explanations are not wholly wrong, but they are incomplete. Much of the difficulty lies earlier, at the level of evidence. A model that seems statistically persuasive in the environment that produced it may not yet answer the question that a new environment would ask of it.

That question is not merely whether the score continues to sort higher-risk from lower-risk cases. In a hospital, a model acquires meaning only through use. A risk estimate may guide who receives additional review, who is monitored more closely, who is sent for imaging, who is transferred, or how clinicians interpret signs that are already borderline. Each of these uses assumes that a given numerical output retains approximately the same practi-

© Copyright by the Authors.
Licensed as an open access article using a CC BY 4.0 license.

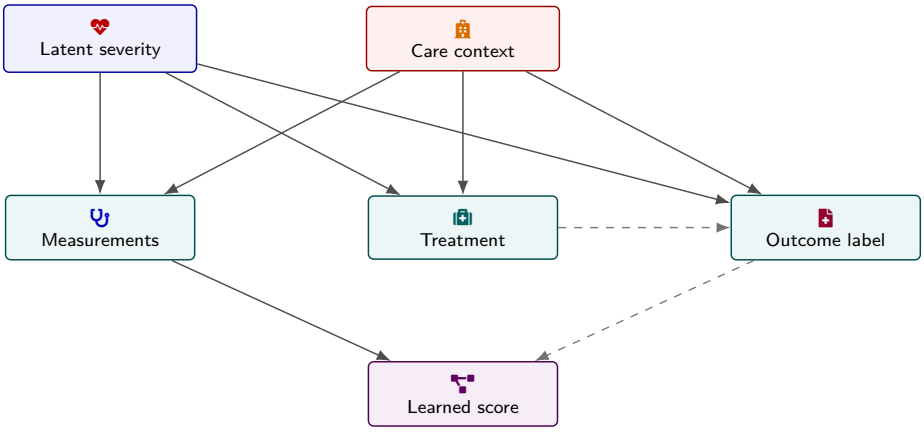


Figure 1: Transportable prediction can be viewed as an identifiability problem in an adaptive clinical environment. The observed score is learned from measurements and labels that are jointly shaped by latent illness, treatment response, and context-dependent labeling, so retrospective fitting alone cannot separate durable signal from care-process artifacts.

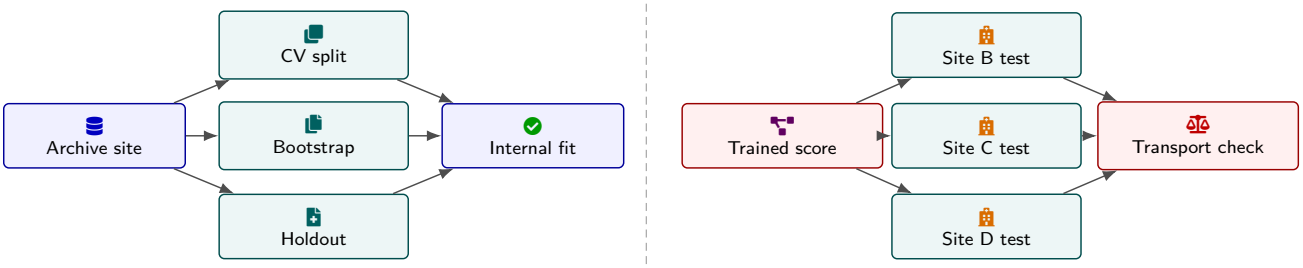


Figure 2: Internal validation repeatedly resamples a single environment and therefore cannot falsify contextual overfitting. Multi-site external validation plays a different role: it probes whether the score survives changes in observation process, treatment regime, and outcome ascertainment, making transport assessment a falsification exercise rather than a routine appendage.

cal interpretation when conditions change. Yet those conditions are rarely stable across institutions or over time. Laboratory tests are ordered under different thresholds of concern. Documentation density differs by service line and staffing pattern. Rescue interventions occur at different times. Outcome labels are assembled from different combinations of review, coding, and surveillance [2]. The same variables can therefore participate in different empirical relations depending on where they are measured. A predictor trained in one setting learns from those relations as they exist there, not from some context-free clinical reality suspended above local practice.

This is why the conventional development narrative can be misleading. It begins from a dataset, splits or resamples that dataset, optimizes model parameters, reports internally validated metrics, and concludes that the resulting score is promising for future use. The logic appears rigorous because it controls optimism relative to the source data-generating process. Yet the core translational question concerns another process entirely. A clinician contemplating adoption is not asking how the model behaves under repeated partitions of the original cohort. The clinician is asking whether the score means the same thing when the monitoring cadence differs, when patient flow differs, when intervention timing differs, and when the target event is documented under another institutional routine. Internal validation does not address those conditions, because it never leaves the jurisdiction of the source process. It tests stability under repetition, not stability under relocation.

A more precise account must therefore begin by asking what a clinical prediction model is actually estimating. The familiar notation $\Pr(Y = 1 \mid X)$ suggests that a model maps observed predictors to an observed outcome. In practice, the observed predictors are themselves institutional products. They arise through measurement systems,

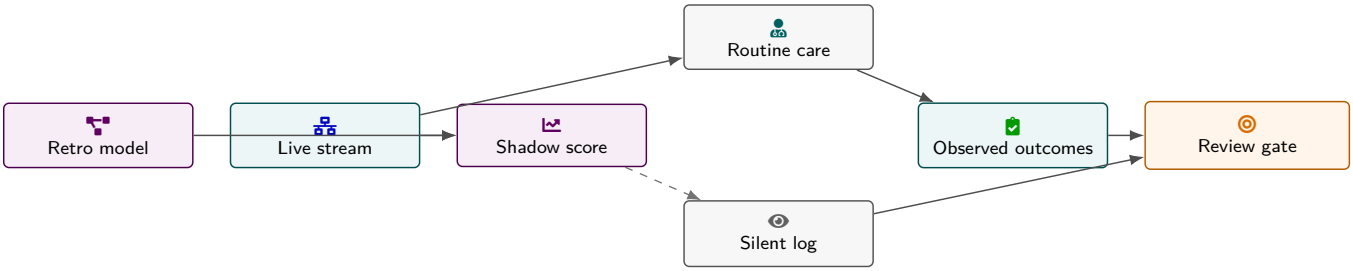


Figure 3: Shadow deployment creates an evidentiary stage between retrospective development and action-linked use. The model runs on live data, but care remains unchanged; silent scores are logged and paired with subsequent outcomes to evaluate calibration, threshold behavior, and operational plausibility before any implementation decision is made.

Table 1: Environment Components in Clinical Prediction Systems

Component	Description	Effect on Data	Variability
Observation O_D	Measurement process	Alters feature availability	High
Action A_D	Treatment policy	Changes outcome pathways	High
Labeling L_D	Outcome construction	Affects target definition	Moderate
Latent state H	True physiology	Underlying signal	Low

ordering behavior, documentation conventions, and care intensity. The observed outcome is likewise a composite generated by surveillance, adjudication, intervention success or failure, and coding practice. A model fitted to these variables is not directly estimating a universal physiological law [3]. It is estimating a relation embedded in a local care ecology. This does not make the model unscientific. It does mean that broader use demands broader evidence.

Once the problem is stated this way, generalizability stops looking like a secondary desirable property and starts looking like the core translational unknown. A model that cannot be shown to preserve interpretation under changed conditions has not yet demonstrated that it is more than a local artifact, even if its local performance is excellent. That unknown is amplified by the adaptive character of healthcare. Hospitals are not static observational settings. They respond to prediction. If a score is deployed and clinicians act on it, then the empirical relation between predictors and recorded outcomes can change. The model is inserted into the very system from which its future data will be drawn. Prediction in clinical care is therefore unlike prediction in a passive environment. It is entangled with intervention, and that entanglement makes naïve notions of transport especially fragile.

The aim of this paper is to develop a technical and conceptual framework adequate to that problem. Rather than describing external validation as a stronger version of internal testing, the paper treats it as multi-environment falsification. Under this view, the main danger in retrospective model development is not simply variance or conventional overfitting. The main danger is contextual overfitting: the model's ability to exploit structural regularities of one care environment that look like durable signal because they are stable within that environment. Internal validation cannot falsify such overfitting, because all of its resamples inherit the same environment [4]. A model becomes translationally credible only when it survives contact with other environments whose observation patterns, treatment responses, and labeling practices differ in clinically relevant ways.

The discussion proceeds accordingly. The next section frames prediction as an identifiability problem under adaptive care processes. A later section explains why internal validation is non-falsifying with respect to contextual overfitting and why high internal performance can coexist with deep translational ambiguity. The paper then analyzes how institutional policy and workflow distort the geometry of observed risk surfaces. After that, external validation is reconstructed as a multi-environment falsification strategy, not a ceremonial external test. The subsequent section introduces shadow deployment as an evidentiary bridge between retrospective modeling and live, action-linked use. The final analytic section turns to the broader structure of translational evidence, arguing that current benchmark cultures reward local fit more efficiently than they reward portable meaning. The position taken is intentionally moderate. It does not imply that broad transport is impossible. It implies that without external falsification and live-environment evidence, many apparently successful prediction models have not yet iden-

Table 2: Sources of Contextual Overfitting

Source	Mechanism	Internal Validation Effect	Transport Risk
Workflow patterns	Stable routines	Reinforced	High
Measurement bias	Selective testing	Invisible	High
Coding practices	Documentation rules	Preserved	Moderate
Care intensity	Monitoring density	Stabilized	High

Table 3: Validation Types and Their Limits

Validation Type	Data Source	What It Tests	Limitation
Split-sample	Same dataset	Sample optimism	No env change
Cross-validation	Same environment	Stability	No falsification
Temporal	Same site later	Drift over time	Limited shift
External	New site	Environment change	Depends on diversity

tified the quantity that bedside use would require them to estimate.

2. Predictive Estimands Under Adaptive Care Processes

A clinically deployed model is asked to estimate future risk from present records, but those records are neither direct readings of latent patient state nor neutral measurements unaffected by care. A more accurate formulation begins with an unobserved state vector H, representing physiology, anatomical status, disease burden, prior exposure, and other determinants of future clinical events. What the model sees is a record vector X, and the relation between H and X is produced by a care process rather than a fixed sensor. Clinical staff decide which tests to order, how often to chart, when to escalate, and what to document. As a result, the observation of state is selective, time-varying, and institutionally mediated [5].

Let D index an environment, where an environment comprises a hospital, service configuration, documentation system, staffing pattern, and associated care routines. Let O_D denote the observation mechanism, A_D the treatment and response policy, and L_D the labeling mechanism through which future outcomes become recorded events. A minimal probabilistic representation is

$$\begin{aligned} X &= O_D(H, U_O) \\ A &= A_D(H, X, U_A) \\ Y &= L_D(H, A, U_L), \end{aligned} \tag{1}$$

where U_O, U_A, and U_L summarize stochastic variation not captured explicitly. The crucial point is that Y depends on A, and A depends on X, so the recorded outcome is already entangled with the observed record and with the local policy that interprets that record. The model does not simply recover a hidden latent function from a noisy snapshot. It learns from a care-mediated process in which observation and intervention are part of the data-generating system.

The estimand commonly implied by predictive studies is the environment-specific conditional law

$$\eta_D(x) = \Pr_D(Y = 1 \mid X = x). \tag{2}$$

This object is entirely legitimate, but it is narrower than translational language often suggests. A score estimated from environment d is informative about $\eta_d(x)$, not automatically about $\eta_t(x)$ for some target environment t. Whether those two quantities are similar enough for practical transport depends on how observation, action, and labeling mechanisms differ across d and t. If all three mechanisms were invariant, the transport problem would be mild. In healthcare they rarely are.

The distinction matters because the same observed vector can encode different latent states in different environments. Suppose an elevated biomarker is measured selectively in one hospital only when concern is already high, but almost routinely in another. A particular numeric value of that biomarker is then embedded in different observation contexts [6]. In the first environment its presence already signals clinician suspicion; in the second it carries less additional information. The conditional law of

Table 4: Types of Risk Surface Drift

Drift Type	Cause	Effect on Model	Detectability
Policy drift	Intervention timing	Calibration shift	Moderate
Observation drift	Measurement changes	Feature distortion	High
Label drift	Outcome definition	Target mismatch	Low
Population drift	Case-mix change	Distribution shift	High

Table 5: External Validation as Falsification

Aspect	Traditional View	Falsification View	Implication
Purpose	Performance check	Hypothesis testing	Stronger evidence
Dataset	Independent	Challenging	matters
Outcome	Single metric	Multi-metric	Rich insight
Interpretation	Pass/fail	Failure modes	Informative

the outcome given the observed vector therefore changes even if the underlying pathophysiology has not. This is not simply a nuisance of missing data or marginal imbalance. It reflects a genuine change in the mapping from latent state to observed evidence.

One way to make this visible is to factor the environment-specific risk through latent state:

$$\eta_D(x) = \int \Pr_D(Y = 1 \mid H = h, X = x)\, dF_D(h \mid x). \tag{3}$$

Transport becomes difficult because both the conditional event term and the posterior state distribution can vary by environment. The posterior term changes when observation policies differ, since the same observed record is compatible with different latent states depending on how and why the record was produced. The conditional event term changes when interventions triggered by the record differ, since the same latent state can lead to a different future outcome under another care policy. A model that fits $\eta_d(x)$ well can therefore fail to estimate $\eta_t(x)$ for two separate reasons that are rarely disentangled in routine reporting.

This adaptive-process perspective also clarifies why some features that look clinically sensible are nonetheless transport-fragile. A variable may be closely tied to patient state and yet owe much of its predictive value to how its measurement is triggered. Time-to-test, repetition frequency, and missingness pattern are obvious examples, but the problem extends further. Even a directly measured laboratory value can inherit transport fragility if the timing, normalization, or patient subset on which it is observed differs enough across environments. What looks like stable signal from the perspective of the source environment may partly encode local clinical attention.

An explicit decomposition helps. Let the observed record be written as

$$X = X^{\text{state}} + X^{\text{care}}, \tag{4}$$

where X^{state} carries direct information about latent patient condition and X^{care} carries information about the local process by which the condition became observed. The decomposition need not be exact or identifiable from one environment alone [7]. It is a conceptual reminder that empirical risk minimization does not know which part will travel. It simply uses whichever part reduces loss in the development setting. If X^{care} is highly stable and highly predictive within that setting, the model can lean on it heavily without suffering any detectable penalty in internal validation.

There is a further complication. The recorded outcome may not represent the same clinical quantity as the latent event of real interest. Let E denote a latent adverse process such as deterioration, complication, or disease progression. The model, however, is trained on Y, the recorded label. These are connected through the labeling mechanism:

$$Y = L_D(E, S_D, A_D), \tag{5}$$

where S_D denotes surveillance intensity or ascertainment behavior. If surveillance differs, then the same latent event burden can produce different recorded labels. If treatment differs, then some latent events may be mitigated before they manifest as recorded complications. The recorded target is therefore policy-sensitive. The model may be excellent at predicting how one system records the event while remaining weakly informative about how another system would.

Table 6: Shadow Deployment Analyses

Analysis	Objective	Metric	Benefit
Feature availability	Real-time inputs	Missing rate	Feasibility
Shadow calibration	Probability accuracy	$m(s)$	Reliability
Workload estimation	Alert burden	$W(\tau)$	Usability
Temporal monitoring	Stability	Rolling stats	Drift detection

Table 7: Performance Dimensions Across Environments

Dimension	Metric	Importance	Stability Need
Discrimination	AUC	Ranking	Moderate
Calibration	Slope/intercept	Probability meaning	High
Utility	Net benefit	Clinical action	High
Workload	Alert rate	Operational cost	High

In this setting the very notion of a universal clinically useful score depends on hidden invariance assumptions. The implicit hope is that there exists a representation $\phi(X)$ for which

$$\Pr_{D}(Y = 1 \mid \phi(X) = z) \approx \Pr_{D'}(Y = 1 \mid \phi(X) = z) \quad (6)$$

across relevant environments D and D'. That is a strong claim [8]. It requires that the representation strip away environment-specific observational and policy artifacts while retaining clinically durable structure. Standard model development does not test this claim. It simply assumes, often without naming the assumption, that the learned relation will not break when the environment changes.

The translational problem can therefore be restated as follows. A retrospective study yields an estimate of an environment-bound conditional law. Bedside use, however, requires a conditionally stable law across environments sufficiently similar to the intended deployment context. The gap between those two objects is not a technicality. It is the reason that external validation must do more than quantify general optimism. It must probe whether the model has identified a clinically transportable estimand or merely a locally serviceable one. Without that probe, retrospective performance remains evidence of fit, not evidence of portability.

3. Why Internal Validation Cannot Falsify Contextual Overfitting

Internal validation is designed to estimate how much a model benefits from accidental fit to peculiarities of the development sample. That is a useful objective. It can reveal excessive parameter tuning, instability from limited events, and poor optimism control. Yet internal validation addresses only one family of errors: those that vary under resampling from the same environment. Contextual overfitting belongs to another family. It arises when the model captures durable regularities of the source environment that are not durable regularities of the broader clinical task [9]. Because those regularities are reproduced across train-test splits within the same environment, internal validation treats them as legitimate signal. In consequence, a model can receive impeccable internal support while remaining deeply ambiguous for translation.

A convenient way to formalize this is to distinguish sample variation from environment variation. Let $\mathcal{R}_d(f)$ denote the predictive risk of model f in development environment d under loss ℓ:

$$\mathcal{R}_d(f) = \mathbb{E}_{(X,Y)\sim P_d}\big[\ell(f(X), Y)\big]. \quad (7)$$

Internal cross-validation targets an estimate of $\mathcal{R}_d(f)$ or of the optimism-corrected counterpart associated with models trained repeatedly under the same environment. The translational object of interest is instead $\mathcal{R}_t(f)$ for some target environment t. The difference

$$\Delta_{d,t}(f) = \mathcal{R}_t(f) - \mathcal{R}_d(f) \quad (8)$$

cannot be estimated by internal validation because no internal resampling scheme changes P_d into P_t. This is obvious in one sense, but its consequences are often understated. A methodology can be exceptionally careful about within-environment optimism and still remain silent about the quantity that deployment decisions care about most.

The problem grows more acute when one notices that contextual overfitting can produce highly stable source-domain gains. Suppose the true transportable signal is en-

Table 8: Benchmark Regimes Comparison

Feature	Monoculture	Transport-aware	Outcome
Datasets	Single-site	Multi-site	Better coverage
Metrics	Discrimination	Multi-dimensional	Richer eval
Reporting	Aggregated	Heterogeneous	Transparency
Goal	Local fit	Portability	Translation

Table 9: Stages of Translational Evidence

Stage	Setting	Evidence Type	Risk
Development	Retrospective	Internal metrics	Context bias
External validation	New datasets	Falsification	Partial shift
Shadow deployment	Live passive	Operational fit	Pre-policy
Active deployment	Live active	Sequential monitoring	Feedback effects

coded in a low-dimensional stable component Z, but the development environment contains an additional structure R_d tied to local workflow, documentation behavior, or surveillance intensity. A model with enough flexibility can learn

$$f_d(X) = g(Z, R_d), \tag{9}$$

where R_d improves source performance even though it will not be available, or will mean something different, elsewhere. Because R_d is stable across the entire development environment, its predictive contribution does not vanish under internal splits. In fact, the more reproducible the local structure is, the more internal validation will endorse it. What looks like reassuring stability is therefore compatible with a deeper form of fragility.

This can be framed through a bias decomposition. Let $f_t^\star$ be the target-optimal predictor for environment t within a function class $\mathcal{F}$. Then

$$\begin{aligned}\mathcal{R}_t(f_d) - \mathcal{R}_t(f_t^\star) = &\Big(\mathcal{R}_d(f_d) - \mathcal{R}_d(f_d^\star)\Big) \\ &+ \Big(\mathcal{R}_t(f_d) - \mathcal{R}_d(f_d)\Big) \\ &- \Big(\mathcal{R}_t(f_t^\star) - \mathcal{R}_d(f_d^\star)\Big),\end{aligned} \tag{10}$$

where the first term is the source-environment excess risk and the remaining terms capture environment mismatch. Internal validation chiefly addresses the first term [10]. Contextual overfitting lives in the mismatch terms. A model can have negligible internal excess risk and still incur large target excess risk because the learned function leans on source-environment structure that the target does not reproduce.

The same point can be stated in more qualitative language. Internal validation asks whether the model has learned the development environment well in a way that is robust to sample perturbation. It does not ask whether the environment itself is a misleading tutor. A hospital with unusually thorough charting, unusually frequent laboratory repetition, or an unusually stable response protocol can teach a model relations that generalize perfectly inside that hospital and poorly outside it. Internal validation is blind to this because it never changes the tutor. It audits the student under repeated versions of the same exam.

This blindness is particularly problematic when model selection is performed adaptively. Development pipelines often compare many feature sets, preprocessing options, architectures, and hyperparameter combinations, even when they are packaged under a single final model report. Such pipelines are already vulnerable to selection-induced optimism, which careful internal validation tries to control. But even if that control is done correctly, the winning model may be the one that best exploits context-specific structure. Source-domain competition among models rewards any exploitable regularity that is stable within that environment. There is nothing in the selection criterion that privileges transportable signal over local surrogate signal unless the developer explicitly designs for it.

The consequence is a non-falsifiability problem. A scientific claim is weak when the available procedure cannot reveal its failure [11]. If the substantive claim is that the model captures clinically durable structure relevant outside the development site, internal validation is weak evidence because the procedure is not positioned to falsify environment-bound success. The claim might be true, but the method does not really challenge it. This matters

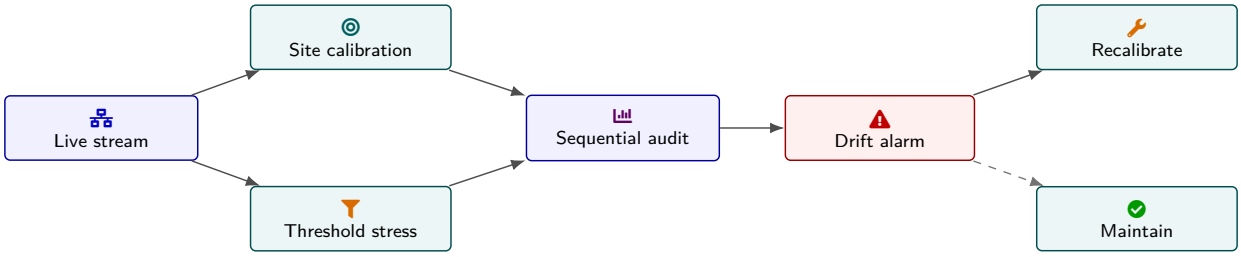


Figure 4: After deployment, translational credibility depends on ongoing surveillance rather than a one-time validation result. Environment-specific calibration checks, threshold stress testing, and sequential monitoring convert live performance into a governed process with explicit paths for maintenance or recalibration as practice patterns drift.

not only philosophically but operationally. A field can accumulate many internally convincing papers without accumulating corresponding evidence about portability, because the central translational failure mode is not being probed.

It is useful to contrast ordinary noise fitting with contextual overfitting. Noise fitting attaches the model to fluctuations that would disappear under another sample from the same environment. Contextual overfitting attaches the model to regularities that persist under every sample from the source environment but not under environment change. The second form is more dangerous because it survives all the usual internal checks and therefore presents itself as genuine model competence. One could even say that contextual overfitting is the natural outcome of empirical risk minimization when the environment itself contains highly structured shortcuts. The model is doing exactly what it is asked to do.

This helps explain why performance deterioration across hospitals is often surprising to investigators but not mysterious once one looks at the data-generating process. A development study may have used common-sense features, careful regularization, and proper internal validation. Yet the model's success may still depend on patterns of test ordering, coding sequence, or postprocedure surveillance that carry different meanings elsewhere. Since those patterns are part of the empirical record, the model has no reason not to use them. It does not know which correlations are civic locals and which are international travelers.

A stricter validation logic is therefore needed [12]. To falsify contextual overfitting, one must change the environment. More precisely, one must expose the model to environments that perturb the observational, treatment, or labeling structures in ways relevant to intended use. If the score remains interpretable, confidence in transportable signal increases. If it degrades, then the failure reveals not simply that the model was bad, but that the source evidence never isolated the transportable component strongly enough. External validation is central because it provides that perturbation. Without it, internally polished success may still be little more than local literacy.

4. Risk Surface Drift from Institutional Policy and Workflow

The difficulty of transport does not arise only from changed patient populations. It also arises because institutions reshape the very geometry of the risk surface that a model learns. The mapping from records to future events is influenced by when clinicians intervene, which thresholds trigger workup, how monitoring intensity escalates, and how outcomes become visible. These institutional policies change the observed relation between current predictors and future labels. A score trained in one policy regime may therefore face a different risk landscape when moved to another, even if the underlying pathology is unchanged.

To see this, let $P_D(A \mid X, H)$ denote the response policy in environment D. Because action affects future outcomes, the recorded event risk satisfies

$$\begin{aligned}\eta_D(x) &= \Pr_D(Y = 1 \mid X = x) \\ &= \int \sum_a \Pr_D(Y = 1 \mid H = h, X = x, A = a) \\ &\quad \times P_D(A = a \mid H = h, X = x)\, dF_D(h \mid x).\end{aligned} \tag{11}$$

The dependence on policy enters twice: through the action distribution and through the posterior over latent state given the observed record. As a result, two environments can share similar latent disease processes and still produce materially different observed risk surfaces. The model does not see those surfaces as policy artifacts. It sees them as the target function [13].

Consider a simple example. In one hospital, borderline postoperative signs trigger immediate imaging and early source control. In another, the same signs prompt watchful waiting unless further deterioration occurs. A model trained on the first hospital learns from data in which high-risk trajectories are partially interrupted by early action. A model trained on the second learns from a risk surface closer to untreated progression. Neither model is intrinsically wrong. Each is adapted to a different policy-coupled reality. Transporting one to the other changes both the meaning of early signs and the outcome frequencies attached to them.

This phenomenon can be written as a policy perturbation of the risk surface:

$$\eta_t(x) = \eta_d(x) + \delta_{\text{pol}}(x) + \delta_{\text{obs}}(x) + \delta_{\text{lab}}(x), \qquad (12)$$

where δ_{pol} captures changes induced by response policy, δ_{obs} captures differences in observation and measurement, and δ_{lab} captures differences in label construction. The decomposition is schematic, but it is analytically useful because it shows that risk drift is not a single monolithic process. Some drift may be mostly intercept-like, some may distort slopes, and some may concentrate in restricted parts of the state space. In practice, the distribution of these distortions matters more than any average metric. A model can look externally respectable by global discrimination while failing badly in the regions where clinical thresholds are set.

Workflow effects can also create topological distortions in the observed feature space. Suppose that high-concern cases receive repeated measures and richer notes, while low-concern cases remain sparsely documented. The resulting records may cluster not only by latent state but by care intensity [14]. A flexible model can exploit these clusters because they predict outcomes within the source environment. When moved to a site with flatter documentation practices, the same clusters disappear or reorganize. What changes is not simply the average value of a feature, but the local neighborhood structure that the model uses to interpolate risk.

A graph-based representation is helpful here. Let patients in environment D form a similarity graph G_D in observed feature space, with edge weights based on clinical similarity as represented by the recorded predictors. The model's behavior can often be interpreted locally through neighborhoods on this graph. If the environment changes, the graph changes. Nodes that were close because of one site's documentation cadence may not be close in another. A score that depends on local neighborhoods therefore experiences a kind of graph transport problem. The stability of performance then depends on whether clinically meaningful neighborhoods, rather than workflow-induced neighborhoods, are preserved.

This can be linked to spectral analysis. Let L_D be the graph Laplacian for environment D, and let its low-frequency eigenvectors capture large-scale geometry of the observed cohort. If a model relies heavily on directions aligned with environment-specific low-frequency structure, then transfer to another graph can distort the score even if many marginal distributions look similar. One may measure this through an eigenbasis alignment criterion,

$$\Xi(d,t) = \sum_{k=1}^{K} \left\| u_k^{(d)} - u_k^{(t)} \right\|_2^2, \qquad (13)$$

where $u_k^{(d)}$ and $u_k^{(t)}$ are low-frequency eigenvectors of L_d and L_t. A large value suggests that the manifold on which the source score was learned has changed. While this is a stylized abstraction, it captures an important translational fact: the model's local geometry can be environment-specific even when its variables look clinically orthodox.

Institutional policy also affects calibration more directly through prevalence modulation and selective prevention [15]. If a model is trained where aggressive prophylaxis reduces event rates in the very patients who look highest risk, then the apparent mapping from predictor pattern to outcome probability will differ from that in a less interventionist site. When the model is deployed in the second site, it may underpredict realized events among high-risk patients. The reverse can also occur. This is why external calibration often degrades even when variable definitions and patient demographics appear broadly comparable. The hidden difference lies in what happens after the predictors are observed.

Such policy-coupled drift has an especially important implication for explanatory interpretation. Investigators sometimes inspect model coefficients or feature importances and conclude that the model is clinically sensible because it depends on familiar predictors. Yet a clinically familiar predictor can still play a different predictive role under another policy regime. The variable itself may be stable, while the intervention pathways it triggers are not. Clinical plausibility of individual features is therefore a weak guarantee of transport if the surrounding care ecology differs.

This is also the setting in which one sees the limits of purely statistical domain adaptation strategies. Reweighting, alignment, or representation matching can help when

the primary problem is shifted sampling over largely stable conditional relations. They are less decisive when the target relation itself has been altered by different policies or different event construction. In those cases, one is not merely trying to map one covariate cloud onto another. One is trying to transport a predictor across environments in which records participate differently in care. That requires empirical evidence, not only algorithmic cleverness [16].

The broader lesson is that institutional policy is not a peripheral context variable that can be ignored once a model is fit. It is part of the mechanism that determines what the model has learned. Any serious account of clinical translation must therefore treat policy and workflow as generators of risk surface drift. External validation is important not only because it uses different data, but because it changes these generators. If the score continues to behave coherently, then one gains evidence that the learned relation is not simply a policy-bound local shortcut. If it fails, then the failure reveals that the model's apparent competence was inseparable from the way one institution organized care.

5. External Validation as Multi-Environment Falsification

The usual presentation of external validation is procedural. A trained model is applied to an independent dataset, one or more metrics are calculated, and the result is interpreted as support or lack of support for generalizability. That presentation is too thin for translational work. What external validation actually contributes is a form of falsification. It tests whether the model's apparent success survives environments that perturb the structures most likely to make local success misleading. Once framed in this way, the quality of external validation depends not merely on whether the dataset is independent, but on which aspects of the development environment it challenges.

Let $\mathcal{E}$ denote a family of clinically relevant environments. For each environment $e \in \mathcal{E}$, define a risk functional $\mathcal{R}_e(f)$ and possibly additional functionals for calibration, threshold utility, or workload. The translational hypothesis of interest is not that the model performs well in one holdout sample, but that its clinically relevant behavior remains acceptable over a meaningful subset of $\mathcal{E}$. External validation thus becomes a sequence of attempted refutations. Each new environment asks whether a distinct kind of contextual overfitting has been exposed.

A formal way to express this is to define an acceptance region $\mathcal{A}$ in the space of performance summaries and to consider

$$\mathcal{V}(f) = \Big\{ e \in \mathcal{E} : G_e(f) \in \mathcal{A} \Big\}, \tag{14}$$

where $G_e(f)$ is a vector of performance quantities. A broad translational claim implicitly asserts that $\mathcal{V}(f)$ contains the environments relevant to intended deployment. The scientific value of external validation lies in reducing uncertainty about that set [17]. One external cohort is a single probe. Multiple heterogeneous cohorts provide a more informative falsification sequence because they challenge different components of the contextual-overfitting story.

A surgical review provides a concrete illustration of why this matters. Sadanandan (2024) observed that only 4 of 22 reviewed predictive studies had external validation and that reproducibility and implementation readiness remained limited [18]. Interpreted through the present framework, the issue is not merely that an extra methodological box was left unticked. The issue is that most models were not exposed to enough distinct environments to test whether their retrospective success represented durable structure or locally stable artifacts. The evidentiary deficit is therefore a deficit in falsification, not only in reporting completeness.

This interpretation also changes how one should design validation studies. Not all external environments are equally informative. A temporal holdout within the same hospital tests drift over time but may leave observation and policy mechanisms largely intact. A geographic validation in a similar tertiary center tests some institutional movement but may still remain within a narrow operational niche. A validation in a lower-resource site, a community hospital, or a service line with different monitoring density may be much more informative if the intended deployment spans such contexts. The goal is not simply independence. It is challenge. A weak challenge can produce a favorable external result without telling us much about translation.

One can make this explicit by defining an environment challenge functional $\Gamma(e, d)$ measuring how much environment e perturbs the aspects of development environment d most relevant to transport. Such a functional might incorporate differences in measurement frequency, missingness structure, intervention timing, case-mix composition, or outcome ascertainment [19]. Then a valida-

tion program could be summarized not only by the number of cohorts but by the weighted challenge it imposes:

$$\mathcal{Q}(f) = \sum_{j=1}^{m} w_j \Gamma(e_j, d), \tag{15}$$

where $e_1, \ldots, e_m$ are validation environments. A model that survives a high-challenge validation program has passed a more meaningful test of portability than a model validated only in weakly perturbed environments, even if both report comparable external c-statistics.

External validation should also estimate heterogeneity rather than merely pool across sites. Suppose the model is evaluated in environments $e_1, \ldots, e_m$, yielding performance summaries $g_1, \ldots, g_m$. A random-effects representation

$$\begin{aligned} g_j &= \mu_g + b_j + \varepsilon_j, \\ b_j &\sim \mathcal{N}(0, \tau_g^2), \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma_j^2) \end{aligned} \tag{16}$$

decomposes average external performance, between-environment instability, and within-environment estimation noise. The translationally important quantity is often not μ_g alone but the combination of μ_g and τ_g^2. A model with moderate average performance and low heterogeneity may be more operationally dependable than one with stronger average performance but high heterogeneity. Falsification is informative precisely because it reveals this dispersion.

Calibration deserves special attention in a falsification framework because it is the most direct test of whether the score retains the same practical semantics. If the model output is $s = f(X)$, one may examine environment-specific calibration maps

$$m_e(s) = \Pr_e(Y = 1 \mid f(X) = s). \tag{17}$$

A model that maintains ranking but requires large and inconsistent recalibration across environments has not been falsified in the weak sense of usefulness for ordering, but it may be falsified for uses requiring literal probability interpretation. This is why external validation should be purpose-sensitive. The same score can pass one translational test and fail another, depending on whether the intended action relies on ranking, calibration, or thresholded utility.

A further implication is that adaptation and falsification should not be conflated [20]. If a model is recalibrated or partially refit in the target environment, the result answers a different question from raw transport. It asks whether the source model provides a useful scaffold for local adaptation. That is a legitimate and often valuable question, but it should not be mistaken for evidence that the original model transported unchanged. A strong validation program should therefore separate at least three stages: raw evaluation, limited recalibration, and structural updating. Each stage corresponds to a different practical burden of deployment.

The falsification view also treats failure differently. An unfavorable external result is not merely a negative verdict on a particular model. It is information about what kind of signal the development study failed to isolate. If a model fails in a site with sparser measurement but similar case mix, the problem may lie more in reliance on observation-process cues than in physiology. If it fails under temporal drift after a protocol change, the problem may lie in policy sensitivity. If it fails only for probability interpretation but retains ranking, the issue may be score semantics rather than ranking geometry. The value of external validation lies partly in exposing these distinctions.

Finally, multi-environment falsification explains why translation remains difficult even when the field is methodologically sophisticated. Many studies are optimized to minimize source-domain loss and to report carefully optimism-corrected source metrics. Falsification requires another level of discipline: a deliberate search for environments that can prove the model more local than it appears. This is less glamorous than model invention but more central to bedside credibility [21]. A predictor that has not been confronted with meaningful environmental challenge may be an interesting local approximation, but it has not yet earned the right to travel.

6. Shadow Deployment and Sequential Counterfactual Approximation

Even a strong external validation program does not settle the entire translational problem, because retrospective environments do not fully reproduce the conditions of live use. Once clinicians see a score and begin to react to it, the empirical process changes. High-risk cases may receive earlier intervention, additional review, or intensified monitoring. These actions can reduce event rates among those the score identifies, alter what gets measured next, and modify the meaning of future labels. A model that looked well calibrated retrospectively may therefore appear miscalibrated once it starts succeeding operationally. Conversely, a model that seemed acceptable retrospectively may prove difficult to use because the required inputs are

unavailable in real time or because the alert burden overwhelms workflow. This is why an intermediate stage is needed between retrospective validation and live clinical action. That stage may be called shadow deployment.

In shadow deployment the model runs on the live data stream but does not yet change care. Its outputs are logged, time-stamped, and compared against future outcomes and operational realities without being shown to clinicians or without being permitted to alter decisions. The purpose is not merely caution. It is identification. Shadow mode reveals whether the score that seemed coherent in research extracts remains coherent when fed production data with real-time latency, missingness, interface constraints, and uncurated feed behavior. It also allows estimation of how the score would behave in the target environment before the environment starts reacting to it.

Formally, let X_t^{live} denote live-stream inputs at time t, and let $S_t = f(X_t^{\text{live}})$ be the shadow score. Since the score does not yet influence action, observed future outcomes remain generated under the predeployment policy π^0. The shadow setting therefore approximates the estimand [22]

$$\Pr\left(Y^{\pi^0} = 1 \mid X_t^{\text{live}}\right), \tag{18}$$

where Y^{π^0} is the outcome under baseline care. This is often closer to the retrospective training target than the fully deployed environment will be. The shadow stage is thus the last opportunity to study transport under the operational data stream before policy feedback changes the process.

Shadow deployment supports several crucial forms of analysis. The first is operational feature fidelity. Research datasets frequently include variables that are available only after cleaning, backfilling, or delayed integration. A model may depend on such variables without the issue being obvious during development. In shadow mode, one can estimate the real-time availability process

$$M_t^{\text{avail}} = \mathbf{1}\{X_t^{\text{live}} \text{ contains required inputs at decision time}\}, \tag{19}$$

and assess how often the model would abstain, impute, or degrade. A high-performing retrospective model may turn out to be a weak operational candidate if real-time feature availability is poor.

The second analysis is shadow calibration and shadow workload. Because the live environment may differ from the retrospective extract even before clinicians react to the score, one should estimate environment-specific calibration maps and action-rate curves under the production feed. Let $m_{\text{sh}}(s)$ denote the shadow calibration function and let

$$W_{\text{sh}}(\tau) = \Pr\left(S_t \geq \tau\right) \tag{20}$$

denote the shadow alert fraction for threshold τ. These quantities can differ materially from their retrospective analogues because timestamps, coding latency, and live documentation patterns alter score distribution. Shadow mode therefore serves as an empirical bridge from research reality to operational reality.

The third analysis concerns sequential uncertainty. Clinical environments drift [23]. A fixed external validation cohort can only offer a static picture, whereas shadow deployment provides a time-indexed sequence of performance summaries. Let g_t denote any monitored quantity such as calibration slope, positive predictive value at a candidate threshold, or missing-input rate. One can then track a rolling estimate

$$\bar{g}_{t,h} = \frac{1}{h} \sum_{u=t-h+1}^{t} g_u \tag{21}$$

over a window of length h, allowing investigators to examine whether the model's operational behavior is stable before live adoption. This is especially important when deployment is contemplated during periods of protocol change, seasonal case-mix shifts, or software migration.

A subtler role of shadow deployment is sequential counterfactual approximation. Once the score is activated, the realized outcome process becomes policy-dependent. Shadow mode preserves the predeployment policy long enough to approximate what the model would have predicted and how it would have calibrated under baseline care. This matters because it creates a reference against which postdeployment changes can be interpreted. If outcomes later change among alerted patients, one can compare against the shadow-period baseline to ask whether the change likely reflects beneficial intervention, drift, or both. Without such a reference, postdeployment recalibration can become conceptually confused.

The transition from shadow mode to active deployment can be expressed in terms of policies. Let π^0 be the baseline clinical policy and π^1 the policy once clinicians can react to the score. The postdeployment outcome corresponds to

$$Y^{\pi^1} = L_D(H, A^{\pi^1}), \tag{22}$$

not to the predeployment quantity Y^{π^0}. The model may continue to estimate the latter more faithfully than the former, especially immediately after activation. This is not necessarily a defect [24]. Some clinical uses genuinely want a score approximating baseline risk before incremental intervention. But it does mean that the operational meaning of calibration must be specified. Is the score supposed to predict risk under usual care, risk under score-informed care, or something closer to latent untreated vulnerability? Shadow deployment helps disentangle these possibilities before live feedback blurs them.

Once the model becomes active, sequential monitoring must continue. A simple framework uses monitored summaries g_t and a cumulative deviation statistic

$$Z_T = \sum_{t=1}^{T} w_t \, (g_t - g_{\text{ref}}), \tag{23}$$

where g_{ref} is established during shadow mode and w_t emphasizes recent data. Exceeding a predefined tolerance would trigger review, recalibration, or withdrawal. In a translational framework, such monitoring is not an implementation luxury. It is part of the evidentiary path by which a model remains clinically interpretable over time.

Shadow deployment also changes how one thinks about threshold selection. Thresholds chosen from retrospective cohorts may not produce acceptable workload or capture in live operation. During shadow mode, one can estimate the threshold-specific workload, case yield, and timing behavior under production conditions. This allows thresholds to be chosen with knowledge of actual operational burden rather than idealized retrospective burden. The point is not to sneak in target tuning under the guise of validation, but to recognize that a threshold is part of the deployed instrument. An untested threshold is an untested policy.

The larger significance of shadow deployment is that it acknowledges a missing evidentiary layer in much of current prediction research. Retrospective validation and prospective use are often treated as adjacent stages. In reality, there is a substantial gap between them [25]. Production data streams, real-time feature availability, live timestamp semantics, and the impending policy response all stand in that gap. Shadow mode makes the gap visible and measurable. A model that clears this stage has not solved the full problem of translation, but it has crossed from retrospective plausibility into a more demanding form of operational reality.

7. Translational Evidence Beyond Benchmark Monoculture

A final reason generalizability remains a translational obstacle is that the prevailing culture of evidence does not reward portability as consistently as it rewards local optimization. Many benchmark practices, publication norms, and model-comparison habits are structured around source-domain performance. Random train-test splits within one dataset are common because they are convenient, reproducible, and statistically legible. Leaderboards emphasize discrimination because it is easy to compare. Even when external cohorts are available, results are often summarized into one global number that hides where transport is weak. The field therefore produces strong incentives to create locally high-performing models and comparatively weaker incentives to map the limits of portability in a disciplined way.

This can be called benchmark monoculture: the tendency to treat one style of evaluation as if it captured the relevant dimensions of scientific progress. In clinical prediction, the monoculture typically privileges source-domain ranking performance over environment heterogeneity, calibration stability, and operational feasibility. The result is not trivial. A benchmark regime shapes what developers optimize, which features they retain, and which tradeoffs they accept. If the benchmark values one-hospital internal performance, then models that exploit one-hospital structure will flourish. If the benchmark values cross-environment stability, different models will emerge.

A more translationally relevant benchmark structure would summarize performance across environments and uses [26]. Let $\mathcal{E}_{\text{test}}$ denote a set of external environments, and let each candidate model f be associated with a scorecard

$$S(f) = \Big(\bar{D}(f), V_D(f), \bar{C}(f), V_C(f), \bar{U}(f), V_U(f)\Big), \tag{24}$$

where $\bar{D}$ and V_D summarize discrimination and its heterogeneity, $\bar{C}$ and V_C summarize calibration or calibration loss and its heterogeneity, and $\bar{U}$ and V_U summarize clinically specified utility and its heterogeneity. Such a scorecard does not force all models into one scalar ranking. It instead acknowledges that translational value depends on both average behavior and stability across settings. A model that is modestly less sharp but much more stable could then be recognized as superior for deployment purposes.

This evidence structure would also change sample-size priorities. Under benchmark monoculture, an impressive development dataset from one source is often treated as sufficient foundation for model prominence. Under a transport-sensitive regime, investment would shift toward more diverse validation environments and better characterization of site-level differences. A smaller but genuinely heterogeneous consortium can be more informative about portability than a massive archive built under one operational style. The key resource is not only the number of rows. It is the number of ways in which clinically relevant environment change has been sampled.

Another aspect of monoculture is the treatment of negative findings. A model that fails in a certain setting may be quietly discarded or its failure underreported, because the culture prizes successful predictions more than boundary mapping. Yet for translation, knowing where a score fails is extremely valuable. Failure in low-resource settings, for instance, may reveal dependence on dense measurement. Failure after a protocol change may reveal policy sensitivity. Failure only in specific subgroups may reveal instability in label construction or feature support. A mature evidentiary culture would treat these results as contributions to cartography rather than embarrassments [27]. The field needs maps of where models travel, not only stories about where they shine.

The broader issue is cumulative knowledge. If studies use incompatible outcome definitions, vague reporting of preprocessing, and limited description of the care environment, then even an abundance of published models does not cumulate into reliable translational evidence. Different papers may appear to address the same clinical task while actually modeling different operational targets. Comparisons based on headline metrics then become misleading. The solution is not only more models or larger models. It is more structured evidence about environment, label construction, measurement practice, and intended action. Without such structure, external validation itself can be hard to interpret because one cannot tell whether a failure reflects model brittleness, target drift, or data-reconstruction mismatch.

A transport-sensitive evidence regime would also recognize that no single model is likely to dominate across all environments and uses. The goal is not to identify a universal winner, but to identify models whose performance is stable for specific scopes of use. This has practical consequences. A hospital might rationally choose a model with slightly lower average discrimination if its calibration is easier to maintain and its workload curve smoother under local constraints. That choice is invisible under benchmark monoculture because the benchmark does not represent those dimensions adequately.

There is, finally, a governance implication. If publication and benchmarking systems continue to privilege locally optimized retrospective performance, then downstream institutions must shoulder the burden of discovering portability limits themselves. Each adopting site becomes responsible for additional validation, recalibration, and operational testing. This decentralized burden slows translation and creates uneven standards [28]. A literature that invested more heavily in multi-environment evidence would lower that downstream burden by clarifying earlier which models are likely to travel and under what conditions.

The argument here is not that conventional benchmarks are useless. They are useful for certain stages of method development. The problem is allowing them to stand in for translational evidence. Once a model is discussed as a candidate for broad clinical use, the benchmark should change. It should ask whether the score survives challenge, whether its semantics remain stable, whether its thresholds behave coherently, and whether its operational burden is governable. Until those questions become ordinary rather than exceptional, generalizability will continue to appear as a stubborn translational obstacle because the field will keep rewarding the production of local success more efficiently than the production of portable meaning.

8. Conclusion

Clinical prediction becomes translationally difficult not because statistical learning has failed to find signal, but because the signal discovered in one care environment is often inseparable from the way that environment observes patients, responds to risk, and records outcomes. A model can therefore be internally robust and externally ambiguous at the same time. The central issue is not only whether performance decreases elsewhere. It is whether the quantity estimated by the model remains the quantity that bedside use requires. Once this distinction is made explicit, generalizability becomes less an optional embellishment and more the main unresolved scientific problem for transport beyond the source site.

The framework developed in this paper has approached that problem through identifiability, falsification, and live-environment evidence. Prediction in hospitals was treated as estimation under adaptive care processes rather than under passive data collection. Internal validation was shown to be strong for sample optimism and weak for contextual overfitting. Institutional policy and workflow were

treated as generators of risk surface drift, not background nuisances [29]. External validation was recast as multi-environment falsification that probes whether apparent success survives meaningful challenge. Shadow deployment was introduced as an intermediate evidentiary stage that reveals whether the model remains coherent under live data streams before clinician response changes the target process again. Finally, the paper argued that prevailing benchmark cultures underreward precisely the evidence that translation needs most.

This perspective has practical consequences. A retrospective model should not be treated as broadly promising merely because it has elegant architecture or excellent internal metrics. Its scope of use should be proportional to the environments in which it has survived challenge. Calibration should be evaluated as carefully as ranking when the score will be interpreted literally. Thresholds should be stress-tested under target workload constraints rather than copied from source studies. Live deployment should be preceded by shadow evaluation and followed by sequential monitoring because the act of using the model changes the system from which future evidence is drawn. These requirements are demanding, but they are not excessively so. They simply align the evidentiary standard with the complexity of the object being transported.

A useful clinical model does not need to be universal. It may be local, conditional, or valid only after controlled adaptation. What matters is clarity about what has been shown. When a score is carried beyond the setting that generated it, the decisive question is no longer how well it fit the past. The decisive question is whether it has survived enough environmental challenge to justify belief that its output still means approximately the same thing. External validation matters because it is the empirical method by which that belief becomes more than hopeful inference [30].

References

[1] I. Bartella, S. Brinkmann, H. F. Fuchs, J. M. Leers, H. A. Schlößer, C. J. Bruns, and W. Schröder, "Two-stage hybrid ivor-lewis esophagectomy as surgical strategy to reduce postoperative morbidity for high-risk patients," *Surgical endoscopy*, vol. 35, pp. 1182–1189, 3 2020.

[2] R. F. Payne and J. A. Pain, "Duct-to-mucosa pancreaticogastrostomy is a safe anastomosis following pancreaticoduodenectomy," *The British journal of surgery*, vol. 93, pp. 73–77, 1 2006.

[3] K. Xu, H. Xie, H. Ma, and B. Ni, "Diagnosis and treatment for intrathoracic gastroesophageal anastomotic leak: investigation of a new mode," *Zhonghua wai ke za zhi [Chinese journal of surgery]*, vol. 54, pp. 114–118, 2 2016.

[4] C. J. Bradshaw, K. Sloan, A. Morandi, B. Lakshminarayanan, S. Cox, A. J. W. Millar, A. Numanoglu, and K. Lakhoo, "Outcomes of esophageal replacement: Gastric pull-up and colonic interposition procedures.," *European journal of pediatric surgery : official journal of Austrian Association of Pediatric Surgery ... [et al] = Zeitschrift fur Kinderchirurgie*, vol. 28, pp. 22–29, 9 2017.

[5] J. Guay, M. Nishimori, and S. L. Kopp, "The cochrane library - epidural local anaesthetics versus opioid-based analgesic regimens for postoperative gastrointestinal paralysis, vomiting and pain after abdominal surgery," *The Cochrane database of systematic reviews*, vol. 7, pp. CD001893–, 7 2016.

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

[7] C. T. Nyabanga, J. C. Veniero, and B. Shen, "Computed tomography-guided endoscopic needle knife therapy for ileal pouch sinus," *Gastroenterology report*, vol. 4, pp. 334–336, 7 2015.

[8] T. Sammour, M. Lewis, M. L. Thomas, M. Lawrence, A. Hunter, and J. Moore, "A simple web-based risk calculator (www.anastomoticleak.com) is superior to the surgeon's estimate of anastomotic leak after colon cancer resection," *Techniques in coloproctology*, vol. 21, pp. 35–41, 12 2016.

[9] E. Kalogera, H. K. V. Houten, L. R. Sangaralingham, B. J. Borah, and S. C. Dowdy, "Use of bowel preparation does not reduce postoperative infectious morbidity following minimally invasive or open hysterectomies.," *American journal of obstetrics and gynecology*, vol. 223, pp. 231.e1–231.e12, 2 2020.

[10] J. Ma, H. Hu, J. Xiao, Y. Zhao, X. Chen, S. Wei, and H. Zhang, "Esophageal covered stent treatment for

gastroesophageal cervical anastomotic fistula," *Chinese journal of radiology*, vol. 53, pp. 385–388, 5 2019.

[11] W. Salgado, F. Q. Cunha, J. S. dos Santos, C. B. Nonino-Borges, A. K. Sankarankutty, O. de Castro e Silva, and R. Ceneviva, "Routine abdominal drains after roux-en-y gastric bypass: a prospective evaluation of the inflammatory response.," *Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery*, vol. 6, pp. 648–652, 10 2009.

[12] V. E. Strong, A. E. Russo, M. Nakauchi, M. A. Schattner, L. V. Selby, G. Herrera, L. H. Tang, and M. Gonen, "Robotic gastrectomy for gastric adenocarcinoma in the usa: Insights and oncologic outcomes in 220 patients," *Annals of surgical oncology*, vol. 28, pp. 742–750, 7 2020.

[13] M. A. Mederos, M. J. de Virgilio, R. Shenoy, L. Ye, P. A. Toste, S. S. Mak, M. Booth, M. M. Begashaw, M. W. Wilson, W. Gunnar, P. G. Shekelle, M. Maggard-Gibbons, and M. D. Girgis, "Comparison of clinical outcomes of robot-assisted, video-assisted, and open esophagectomy for esophageal cancer: A systematic review and meta-analysis.," *JAMA network open*, vol. 4, pp. e2129228–, 11 2021.

[14] S. R. Benz, I. S. Feder, S. Vollmer, Y. Tam, A. Reinacher-Schick, R. Denz, W. Hohenberger, H. Lippert, A. Tannapfel, and I. Stricker, "Complete mesocolic excision for right colonic cancer: prospective multicentre study.," *The British journal of surgery*, vol. 110, pp. 98–105, 11 2022.

[15] M. Ammendola, M. Ruggiero, C. Talarico, R. Memeo, G. Ammerata, A. Capomolla, R. Filippo, R. Romano, S. Pallio, G. Navarra, S. Montemurro, and G. Currò, "No coil® placement in patients undergoing left hemicolectomy and low anterior resection for colorectal cancer," *World journal of surgical oncology*, vol. 18, pp. 327–327, 12 2020.

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

[17] N. Hanna, Z. M. Mir, E. Williams, S. J. Merchant, B. Zevin, and W. Chung, "Laparoscopic ischaemic conditioning of the gastric conduit prior to a hybrid mckeown oesophagectomy may not decrease the risk of anastomotic leak.," *Wideochirurgia i inne techniki maloinwazyjne = Videosurgery and other miniinvasive techniques*, vol. 16, pp. 669–, 4 2021.

[18] 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.

[19] L. W. Schaheen, S. H. Blackmon, and K. S. Nason, "Optimal approach to the management of intrathoracic esophageal leak following esophagectomy: a systematic review," *American journal of surgery*, vol. 208, pp. 536–543, 7 2014.

[20] A. T. Cristaudo, K. Hitos, R. Gunnarsson, A. de Costa, and A. Yuen, "Re: Development and validation of a multivariable prediction model in open abdomen patients for entero-atmospheric fistula.," 12 2023.

[21] Z. Zheng, Y. Hu, J. Tang, W. Xu, W. Zhu, and W. Zhang, "The implication of gut microbiota in recovery from gastrointestinal surgery.," *Frontiers in cellular and infection microbiology*, vol. 13, pp. 1110787–1110787, 2 2023.

[22] B. K. Wölnerhanssen, I. Langer, U. Eriksson, and M. Schneider, "Sudden death after open gastric bypass surgery," *Langenbeck's archives of surgery*, vol. 394, pp. 573–577, 7 2008.

[23] R. Dávila-Pérez, E. Bracho-Blanchet, J. M. Tovilla-Mercado, J. A. Hernandez-Plata, A. Reyes-Lopez, and J. Nieto-Zermeño, "Unnecessary gastric decompression in distal elective bowel anastomoses in children: A randomized study," *World journal of surgery*, vol. 34, pp. 947–953, 2 2010.

[24] K. Arunsenthilnathan, "A comparative study of efficacy of staplers versus hand sewn anastomosis in bowel surgeries," 5 2019.

[25] N. Patel and V. D. Yagnik, "Laparoscopic colorectal cancer surgery - a prospective study of short-term outcomes of consecutive cases over 3 years," *Internet Journal of Medical Update - EJOURNAL*, vol. 14, pp. 25–31, 8 2021.

[26] D. A. Telem, E. H. Chin, S. Q. Nguyen, and C. M. Divino, "Risk factors for anastomotic leak following colorectal surgery: a case-control study.," *Archives*

of surgery (Chicago, Ill. : 1960), vol. 145, pp. 371–376, 4 2010.

[27] X. Wu, W. Li, X. Gan, and Q. Cao, "182. laparoscopic-assisted single-port mediastinoscopic esophagectomy can provide a safe and effective surgical option for patients with specific esophageal cancers," *Diseases of the Esophagus*, vol. 35, 9 2022.

[28] S. Eğin and S. Hot, "Improvement of esophageal anastomotic leak by vacuum-assisted closure in three cases with gastric cancer.," *Journal of the College of Physicians and Surgeons–Pakistan : JCPSP*, vol. 29, pp. S98–S100, 12 2019.

[29] T. W. Hakkarainen and D. R. Flum, "Selective vs nonselective nonsteroidal anti-inflammatory drugs and anastomotic leakage after colorectal surgery—reply," *JAMA surgery*, vol. 150, pp. 686–687, 7 2015.

[30] A. John, S. D. Chowdhury, R. T. Kurien, D. David, A. K. Dutta, E. G. Simon, V. Abraham, A. J. Joseph, and I. Samarasam, "Self-expanding metal stent in esophageal perforations and anastomotic leaks," *Indian journal of gastroenterology : official journal of the Indian Society of Gastroenterology*, vol. 39, pp. 445–449, 10 2020.