Retrospective Monoculture and Synthetic Hospital Perturbations for Stress Testing ICU Deterioration Benchmarks

Ahmad Firdaus[1]

[11]Faculty of Computing and Informatics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

ABSTRACT

Benchmark-driven progress in intensive care forecasting has produced increasingly capable deterioration models, yet most reported gains remain tied to retrospective corpora constructed within a narrow observational ecology. This creates a recurrent scientific problem. A model may improve substantially within one benchmark while still depending on regularities that arise from a single hospital's measurement cadence, documentation habits, intervention thresholds, and endpoint timing conventions. The result is not merely ordinary overfitting. It is overfitting to a data-generation culture. This paper develops a framework for understanding such benchmark monoculture as a hidden regularizer on model development. Instead of treating a retrospective ICU corpus as a transparent sample of an underlying universal task, the framework treats it as one realization of a hospital-specific observational grammar. The paper then introduces synthetic institutional perturbation as a method for stress testing whether reported improvements survive controlled shifts in recording policy, note density, escalation timing, and label boundary formation. The central argument is that many current evaluation pipelines mistake within-benchmark competence for general scientific progress because they fail to interrogate how much of the learned signal is conditional on one retrospective environment. By combining perturbation-based benchmark generation, invariance-aware optimization, and deployment-facing model selection criteria, the proposed methodology aims to distinguish models that learn clinically stable regularities from those that primarily exploit the idiosyncrasies of one dataset. The broader claim is that the next major improvement in ICU warning systems may come less from larger encoders and more from redesigning the benchmark itself as an object to be stress tested rather than trusted by default.

I. Introduction

The intellectual culture of machine learning has been shaped by benchmarks [1]. A benchmark offers a coherent task, a shared metric space, and a reproducible basis for comparison. Intensive care deterioration forecasting has understandably followed this pattern. Researchers choose a retrospective electronic health record corpus, define a prediction horizon, specify an endpoint such as mortality or treatment escalation, and then train models to assign risk to repeated windows drawn from patient stays. Under this regime, progress becomes legible. A model with stronger discrimination, better ranking, or better calibration than its predecessor appears to have learned something more about impending decline. Yet the benchmark framework also imposes a quiet simplification. It treats the retrospective dataset as though it were a sufficiently transparent sample of a general clinical task [2]. In reality, a retrospective ICU corpus is never just a set of patients with outcomes. It is a hospital-specific observational world.

This matters because an ICU record is not a neutral shadow of physiology. It is assembled through a locally organized mixture of measurement practice, documentation style, staffing rhythm, intervention thresholding, and institutional custom. The resulting dataset therefore contains far more than latent severity. It contains an entire ecology of how one system sees, writes, and reacts to deterioration. When a model is trained and tuned inside that ecology, some fraction of its apparent skill can arise from features that would not survive even modest movement into another hospital or another workflow configuration [3]. The danger is not simply that the model may fail later. The danger is that the field may misread benchmark gains as advances in the underlying science of deterioration when they are partly advances in exploiting one retrospective environment.

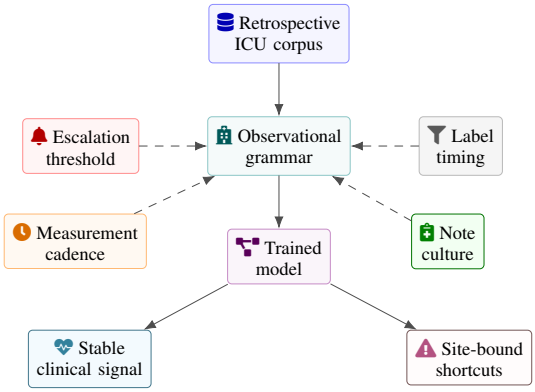

FIGURE 1: Retrospective monoculture can be viewed as an observational grammar wrapped around one ICU corpus. Benchmark gains may therefore mix portable physiology-relevant structure with hospital-specific regularities induced by cadence, notes, escalation policy, and label timing.

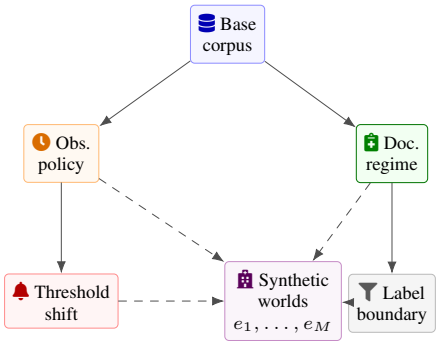

FIGURE 2: Synthetic institutional perturbation turns one retrospective benchmark into a neighborhood of alternate hospital worlds. The perturbations target observation policy, documentation regime, intervention thresholding, and label-boundary formation rather than generic random noise.

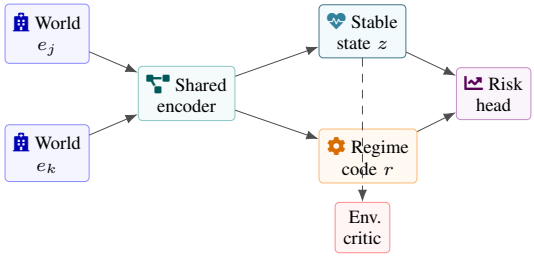

FIGURE 3: Invariance-aware training uses paired synthetic worlds to split the representation into a portable patient-state component and a regime-specific component. A separate environment critic discourages hospital-world information from collapsing into the stable state channel.

That concern is not a speculative worry attached from outside. It is already implicit in how many studies describe their own limitations. Sadanandan (2025), for instance, is careful to frame the reported results as retrospective evidence generated inside one public critical-care corpus rather than as a final demonstration of universal clinical validity [4]. That framing deserves to be taken more literally than it usually is. If a result is retrospective in one observational ecology, then the scientific object that has been demonstrated is not "the model predicts ICU deterioration" in the broad sense. The demonstrated object is narrower: the model predicts deterioration as represented by that corpus, under that corpus's note behavior, endpoint conventions, measurement rhythms, and treatment timing [5]. The claim is meaningful, but it is also conditional. Too often, the conditionality is acknowledged formally and then ignored substantively.

This paper is built around the idea that retrospective monoculture should be treated as a central methodological problem in ICU prediction. By retrospective monoculture, I mean the tendency for model development, tuning, and interpretation to be dominated by one data-generation world whose internal regularities become silently naturalized. A benchmark becomes monocultural when its recording conventions, note regimes, action thresholds, and label construction choices

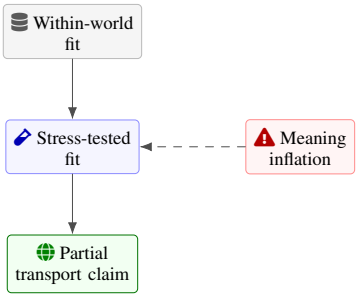

FIGURE 4: Benchmark evidence should be interpreted in layers. A held-out score supports a within-world claim first, a stronger claim only after perturbation stress testing, and transport-oriented language only after that narrower evidence chain is respected.

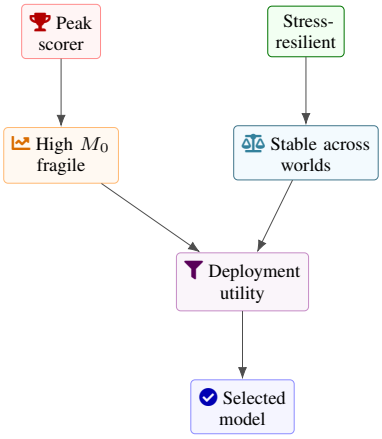

FIGURE 5: Operational selection changes once benchmark dependence is treated as a cost. The chosen model is not necessarily the one with the highest original-world score, but the one with better stability under perturbation, queue behavior, and threshold robustness.

TABLE 1: Benchmark monoculture concept

Component	Role	Source	Effect
Measurement cadence	Sampling rhythm	Hospital workflow	Temporal bias
Note density	Documentation rate	Staffing style	Text signal skew
Intervention timing	Escalation point	Local protocols	Label alignment shift
Endpoint definition	Outcome boundary	Dataset design	Task framing

begin to stand in for the task itself. Once that happens, model comparison becomes a race to fit one ecology more tightly. A model can therefore improve by learning properties that are entirely real inside the dataset and only weakly related to the clinically portable concept the field believes it is studying.

The solution proposed here is not merely more external validation in the usual sense, though that remains valuable [6]. The central proposal is to convert the benchmark from a fixed world into a stress-test generator. Instead of asking only whether a model performs well on one retrospective corpus, the evaluation should ask whether the model's gains survive a family of synthetic institutional perturbations that alter observation density, note timing, escalation boundaries, and label boundary semantics while preserving plausible patient trajectories. This idea shifts the benchmark's role. It ceases to be a trusted approximation of the task and becomes an instrument for probing what kind of regularities the model actually uses.

The paper develops this argument in a sequence intended to make the methodological stakes explicit. The next section defines retrospective monoculture as a hidden regularizer that shapes what models find easy to learn. After that, the discussion turns to synthetic institutional perturbation, a framework for constructing alternate plausible hospital worlds out of one retrospective dataset. The following section introduces invariance objectives and adversarial benchmarking as ways to separate robust clinical structure from environment-bound shortcuts [7]. A later section considers how scientific claims should be reinterpreted once benchmark dependence is acknowledged directly. The paper then examines deployment-

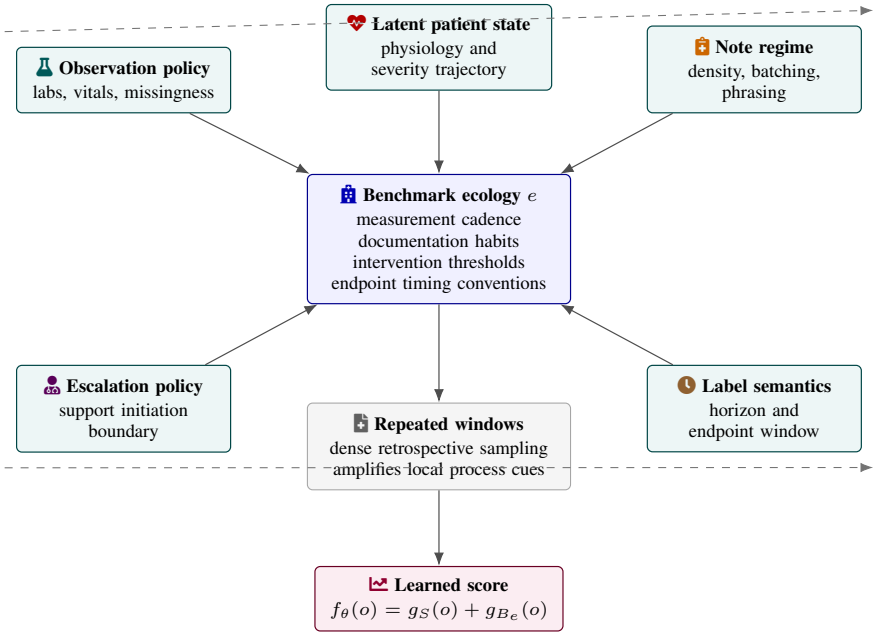

FIGURE 6: Retrospective monoculture can be viewed as an environment parameter that shapes not only the observation stream but also the operational meaning of labels. The benchmark ecology therefore acts as a hidden regularizer: repeated windows and dense reuse of one institutional recording grammar reward models for mixing clinically stable structure with hospital-bound shortcuts.

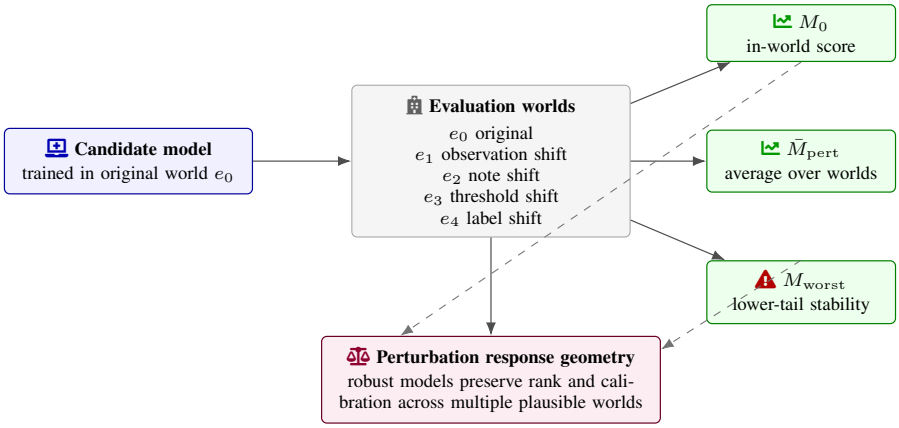

FIGURE 7: Performance becomes a vector over environments rather than a single held-out number. This view emphasizes response geometry: the important question is not only whether a model is strong in the original benchmark world, but also how sharply its utility deforms under alternate plausible institutional realizations.

facing selection rules that prefer stress-resilient models over within-benchmark champions. The conclusion returns to the larger point: in ICU deterioration forecasting, a benchmark is not just an evaluation instrument. It is an environment that actively shapes what is called knowledge.

II. Retrospective Monoculture as a Hidden Regularizer

When a model is trained repeatedly in one retrospective corpus, the corpus begins to function like a regularizer. Not in the narrow mathematical sense of an explicit penalty term, but in the broader scientific sense of constraining the space of solutions that are attractive. The model does not merely learn from labels and predictors. It also learns from the particular ways in which labels were instantiated, predictors were recorded, notes were written, and windows were sampled [8]. If a benchmark is used often enough and cited often enough, its environment-specific regularities become almost invisible. They are no longer treated as contextual properties of the dataset. They are treated as properties of the task.

To see why this happens, let x_t denote latent patient state at time t, o_t the observed record, and y_t the benchmark label. In an abstract predictive formulation, the model approximates $p(y_t \mid o_{\le t})$. Yet in a retrospective ICU benchmark, the conditional $p(y_t \mid o_{\le t})$ is only one marginal of a larger

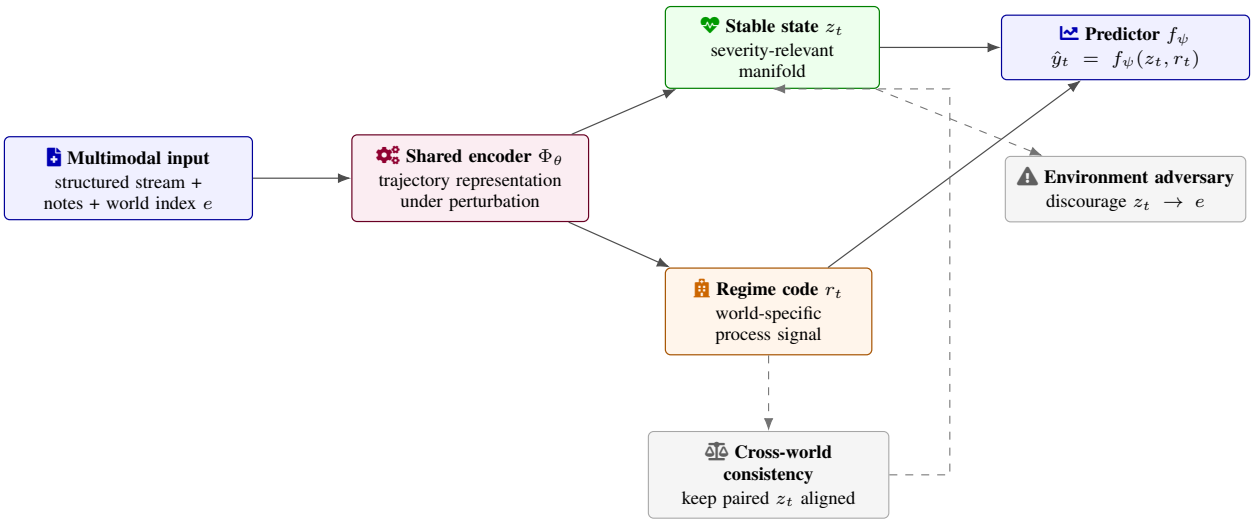

FIGURE 8: An invariance-aware training pipeline separates a patient-state representation from a regime representation instead of forcing the entire model to ignore environment. The adversarial branch pushes clinically meaningful structure into the stable channel, while paired-world consistency discourages unnecessary movement of that channel across perturbations that should preserve the underlying clinical story.

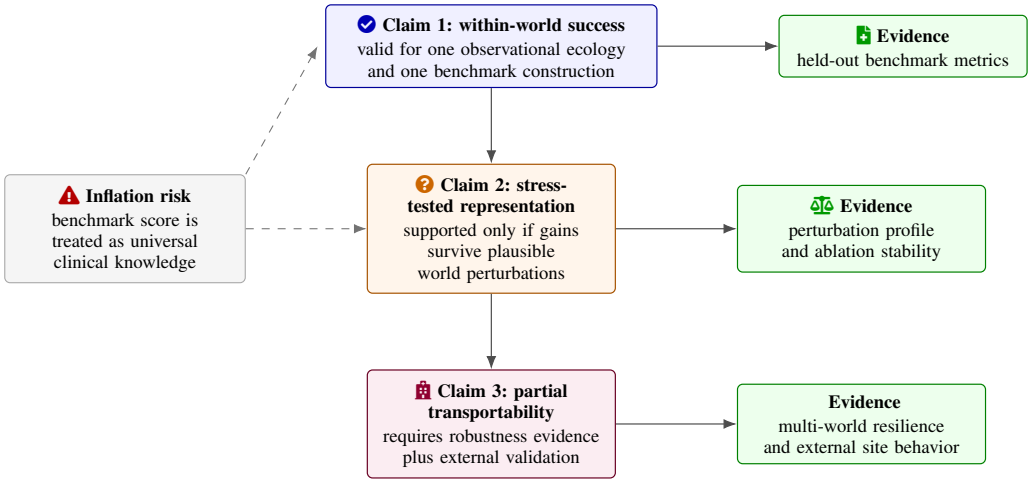

FIGURE 9: Benchmark results support different levels of interpretation, and those levels should not be collapsed into one another. Held-out performance supports a within-world predictive claim; perturbation resilience strengthens the representational claim; and only after additional robustness and external evidence does a partial transport claim become methodologically credible.

environment-bound process. One should more honestly think in terms of

$$p_e(x_{1:T}, o_{1:T}, y_{1:T}) = p_e(x_{1:T})\, p_e(o_{1:T} \mid x_{1:T})\, p_e(y_{1:T} \mid x_{1:T}, o_{1:T}) \tag{1}$$

where e indexes the benchmark environment. The environment parameter does not merely modify covariates slightly. It enters the observation mechanism, the label mechanism, and often the temporal alignment of both [9]. If a model is optimized only under one value of e, then whatever inductive path best fits that environment will look like model competence.

The phrase hidden regularizer is appropriate because the benchmark world narrows the hypothesis space without explicitly announcing itself as a constraint. Suppose one hospital writes abundant nursing notes, uses particular section templates, escalates respiratory support according to one local standard, and tends to measure certain labs more frequently when concern rises. A model trained there can profitably use note density, section-order patterns, intervention timing, and observation bursts as proxies for future deterioration. These proxies are not fictional. They are genuinely predictive inside that world. But they are only partly about latent disease dynamics. They are also about how that hospital operationalizes concern [10]. The benchmark therefore regularizes the model toward explanations in which local process features appear indispensable.

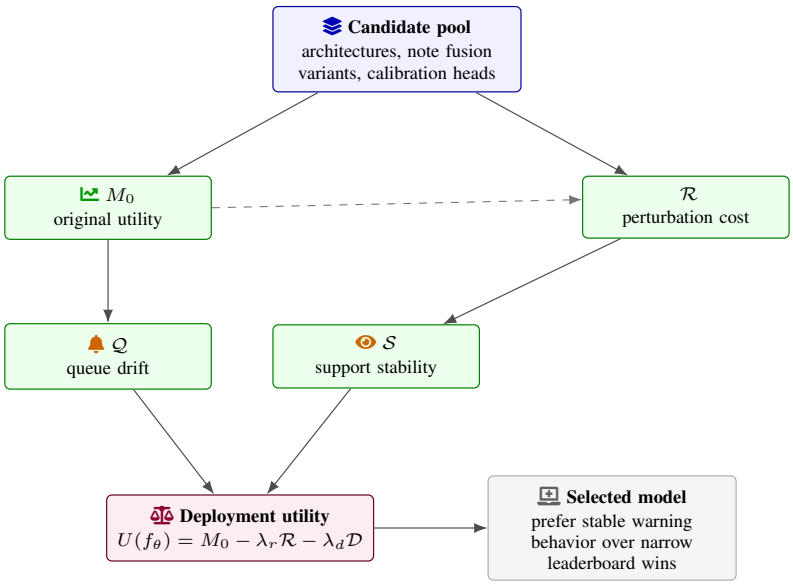

FIGURE 10: Operational model selection should penalize monoculture dependence directly rather than treating within-benchmark utility as the whole objective. A deployment-facing selector can combine original-world performance with perturbation sensitivity, queue instability, and support fragility so that the chosen warning system remains usable when the recording environment shifts.

TABLE 2: ICU deterioration prediction pipeline

Stage	Input	Process	Output
Data collection	EHR streams	Aggregation	Patient timeline
Window sampling	Timeline	Sliding windows	Training examples
Model training	Windows	Optimization	Risk model
Evaluation	Test set	Metrics	Benchmark score

TABLE 3: Sources of environment-specific signal

Signal type	Example	Location	Risk
Process artifacts	Lab bursts	Structured data	Shortcut learning
Documentation habits	Template notes	Text stream	Style bias
Escalation policy	Ventilation start	Interventions	Threshold leakage
Sampling rhythm	Hourly vitals	Monitoring policy	Temporal cueing

This regularization becomes stronger through repeated-window construction. A single patient trajectory can contribute dozens of highly similar examples. If a site's practice style induces certain observation and note patterns around positive episodes, then those patterns are replicated across many windows and therefore receive disproportionate gradient. The model is not only encouraged to learn them. It is rewarded for treating them as stable features of the task. In this way, temporal dependence interacts with monoculture. One hospital's way of surrounding deterioration with chart behavior can become embedded in the score surface as though it were universal [11].

The same issue applies to preprocessing. Every benchmark bakes in decisions about imputation, note truncation, tokenization, sampling cadence, endpoint exclusion, and admissibility of late-arriving information. These are often treated as practical necessities rather than scientific choices. But each choice modifies the relation between latent state and training signal. A corpus whose note embeddings are updated at one cadence rather than another is not the same task. A cohort that excludes certain intervention sequences is not the same target. When one benchmark pipeline dominates the literature, these decisions become part of the hidden

TABLE 4: Synthetic perturbation categories

Class	Target	Mechanism	Goal
Observation policy	Measurements	Density shift	Test monitoring bias
Documentation regime	Notes	Timing change	Test text reliance
Intervention threshold	Treatments	Onset offset	Test escalation cues
Label boundary	Targets	Horizon change	Test task definition

TABLE 5: Observation-policy perturbations

Policy	Variable	Change	Expected effect
Dense labs	Lactate	+frequency	Early signal shift
Sparse nights	Chemistry	-night checks	Missingness change
Focused monitoring	SpO2	event-driven	Burst pattern
Uniform sampling	Vitals	fixed cadence	Reduced irregularity

TABLE 6: Documentation perturbations

Regime	Transformation	Parameter	Effect
Delayed notes	Time shift	+hours	Lagged evidence
Compressed text	Template merge	ratio	Lower density
Boilerplate removal	Repetition filter	threshold	Content emphasis
Sparse reporting	Note drop	probability	Reduced cues

TABLE 7: Model robustness metrics

Metric	Definition	Scope	Insight
Average score	Mean across worlds	Global	Overall stability
Worst-case score	Minimum score	Safety	Failure exposure
Variance	Score dispersion	Reliability	Sensitivity level
Queue overlap	Top-K intersection	Operational	Alert stability

regularizer too. They tell the field which structures are visible and which are not [12].

It is tempting to argue that such regularization is harmless so long as models are ultimately tested externally. That response misses the deeper point. If the monoculture shapes architecture selection, hyperparameter tuning, and interpretive narratives during development, then it has already influenced what kinds of models are taken seriously. External validation may reveal the cost later, but it does not undo the years spent treating one observational ecology as the default scientific world. The problem is therefore not only poor transport at deployment. It is conceptual narrowing during research.

One can formalize the narrowing effect with a decomposition of predictive performance. Let f_θ be a model trained in environment e [13]. Suppose that performance in a new environment e' depends on two kinds of learned structure: environment-stable signal S and environment-bound signal B_e. Then the score function can be decomposed conceptually as

$$f_\theta(o) = g_S(o) + g_{B_e}(o). \tag{2}$$

Inside environment e, both terms help. Across environments, only g_S is guaranteed to remain useful. The problem is that ordinary benchmark optimization has no built-in incentive to keep these components separate. If g_{B_e} is easier to learn and strongly predictive, the model will absorb it. Retrospective monoculture is therefore the condition under which the benchmark actively rewards failure of decomposition.

This is why the benchmark cannot simply be treated as a neutral proving ground. It is a source of inductive pressure

TABLE 8: Representation decomposition

Variable	Meaning	Source	Purpose
z_t	Patient state	Encoder	Stable signal
r_t	Regime context	Environment	Local adaptation
e	Benchmark world	Dataset/perturbation	Domain index
$\hat{y}_t$	Risk prediction	Predictor	Output score

TABLE 9: Training objectives

Objective	Target	Mechanism	Benefit
Task loss	Prediction	Supervision	Accuracy
Adversarial loss	Environment	Classifier	Invariance pressure
Consistency loss	Representation	Cross-world match	Stability
Regime encoding	Context features	Auxiliary head	Adaptation

TABLE 10: Deployment-oriented selection criteria

Criterion	Measure	Motivation	Outcome
Baseline utility	Original score	Benchmark fit	Initial ranking
Robustness cost	Perturbation drop	Stability check	Safer model
Queue drift	Top-K change	Alert reliability	Clinical trust
Support stability	Confidence shift	Epistemic safety	Abstention policy

[14]. It rewards some feature grammars over others, some temporal explanations over others, and some uses of notes or interventions over others. Once one accepts that, the obvious methodological response is not to discard benchmarks, but to make the benchmark speak in more than one institutional voice. That is the purpose of synthetic institutional perturbation.

III. Synthetic Institutional Perturbations as Alternate Hospital Worlds

A single retrospective ICU corpus cannot become a true multicenter dataset merely by being perturbed. Yet it can be transformed into a family of alternate plausible hospital worlds that expose how dependent a model is on one particular observational ecology. The central idea of synthetic institutional perturbation is to modify the benchmark in ways that mimic real inter-hospital differences without destroying the underlying patient trajectories entirely. Rather than treating the original corpus as one immutable test of generalization, the perturbation framework treats it as a base world from which a neighborhood of related worlds can be generated.

These alternate worlds are not arbitrary corruptions [15]. Their purpose is not to test generic robustness to noise. Their purpose is to alter clinically and operationally meaningful aspects of the data-generation process while preserving plausibility. Such perturbations can be organized into at least four classes: observation-policy perturbations, documentation-regime perturbations, intervention-threshold perturbations, and label-boundary perturbations.

Observation-policy perturbations alter how often variables appear, how dense measurements are in different phases of a stay, and how missingness correlates with local severity. Let $o_t^{(s)}$ denote structured observations and m_t a vector of observation indicators. A perturbation operator Π_{obs} can produce a new record

$$(o_t^{(s)\prime}, m_t') = \Pi_{\text{obs}}\big(o_t^{(s)}, m_t; \alpha\big), \tag{3}$$

where α encodes a synthetic hospital policy such as more frequent lactate checks under broad hemodynamic concern, less frequent overnight labs, or more aggressive respiratory monitoring. The goal is not to delete data randomly, but to reshape the conditional distribution of observation given latent trajectory proxies. If a model's performance depends heavily on one site's observation density patterns, these perturbations will reveal it [16].

Documentation-regime perturbations operate differently. Notes can be delayed, made sparser, rebatched into fewer larger documents, stripped of boilerplate repetition, or softened in the relationship between note density and acute

concern. Let d_t denote the note stream. A perturbation Π_{doc} may be defined as

$$d'_t = \Pi_{\text{doc}}\big(d_{1:T};\, \beta\big), \tag{4}$$

where β governs timing shifts, template compression, lexical homogenization, or density changes. A model that uses notes as contextual clinical evidence should degrade differently under these transformations than a model that exploits one documentation culture's surface habits. The distinction is important and often invisible under ordinary in-sample evaluation.

Intervention-threshold perturbations mimic differences in when hospitals initiate supports or record escalation. These perturbations are more delicate, because interventions are not freely movable without violating clinical plausibility [17]. Still, one can alter the observed onset time of an intervention-sensitive endpoint within a clinically constrained window or redefine positivity according to stricter or looser latent severity proxies. If t_k is the observed onset of endpoint channel k, a perturbation may create

$$t'_k = t_k + \delta_k, \tag{5}$$

where δ_k is sampled from an environment-specific distribution restricted by plausibility constraints. This produces alternate hospital worlds in which the same latent trajectory crosses the operational boundary earlier or later. Models overfitted to one site's threshold geometry will often show abrupt instability under such perturbation.

Label-boundary perturbations operate at the benchmark definition level. The same patient-time windows can be relabeled under different horizons, different exclusion rules, or different treatment of post-intervention periods. Let Y_t be the benchmark label [18]. A perturbation family Π_{lab} creates

$$Y'_t = \Pi_{\text{lab}}\big(Y_t;\, \gamma\big), \tag{6}$$

where γ may specify alternate horizons, narrower endpoint sets, or alternate composite compositions. These perturbations do not simulate a new hospital directly. They simulate a change in how the task itself is operationalized. That is important because model robustness to benchmark-definition shift is part of scientific credibility, not only deployment readiness.

Once a set of alternate worlds $e_1, \ldots, e_M$ has been generated, evaluation can be reframed. A model is no longer summarized by one performance score on one benchmark. It is characterized by a response surface over perturbation space. Let $M_j(f_\theta)$ denote performance in synthetic world e_j [19]. Then the benchmark no longer asks only for $\max M_0(f_\theta)$ in the original world. It also asks how $M_j(f_\theta)$ deforms as j changes. Robustness is therefore not a yes-or-no property. It is a geometry.

One can summarize that geometry in several ways. The simplest is the average perturbed performance

$$\bar{M}_{\text{pert}} = \frac{1}{M}\sum_{j=1}^{M} M_j(f_\theta), \tag{7}$$

but this loses structure. A more informative quantity is the worst-world or lower-tail performance, [20]

$$M_{\text{worst}} = \min_{1\le j\le M} M_j(f_\theta), \tag{8}$$

or a risk-sensitive aggregate that weights especially plausible perturbations more heavily. Even more informative is the perturbation profile itself: which kinds of world changes break the model, and which do not.

This synthetic-environment approach does not pretend to replace real external validation. It does something more immediate. It tells the researcher whether a proposed improvement survives alternations of the benchmark world that are clinically plausible and methodologically revealing. If the gain collapses as soon as note density is perturbed or intervention thresholds are shifted modestly, then the improvement was benchmark-bound in a scientifically important sense. If the gain survives across a wide family of worlds, then the model has earned a stronger interpretive claim even before a multicenter trial is available.

The challenge, of course, is that not every perturbation should matter equally, and some models may become robust by becoming uninformatively conservative [21]. This is why perturbation alone is not enough. It must be coupled to training principles that explicitly encourage the separation of hospital-specific regularities from clinically stable ones. That is the subject of the next section.

IV. Invariance Objectives and Adversarial Benchmarking

Synthetic hospital worlds are most useful when they influence not only evaluation but learning. If a model is trained solely to maximize performance in the original retrospective benchmark, then perturbation testing remains an external audit. Useful, but disconnected from optimization. A stronger methodology uses alternate worlds to shape the representation itself. The goal is not to force total invariance, because some environment-specific adaptation is legitimate [22]. The goal is to make the model distinguish between signal that should travel and signal that should be treated as local.

A useful starting point is to decompose the learned representation into a patient-state component z and an environment component r. Let e index the original benchmark world and its synthetic perturbations. The encoder produces

$$z_t, r_t = \Phi_\theta(o_{\le t}, e), \tag{9}$$

where z_t is meant to capture latent deterioration-relevant state that should remain meaningful across worlds, while r_t captures features of the current observational regime. The predictor can then combine these with controlled interaction:

$$\hat{y}_t = f_\psi(z_t, r_t). \tag{10}$$

This alone does not solve the problem, because the network may still pack everything into z_t. One needs training signals that force the intended split [23].

One such signal is adversarial environment prediction. If z_t truly encodes environment-stable structure, then a separate

classifier should have difficulty predicting the perturbation world e from z_t alone. One can therefore optimize

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_r \mathcal{L}_{\text{regime}} - \lambda_a \mathcal{L}_{\text{adv}}, \tag{11}$$

where $\mathcal{L}_{\text{task}}$ is the deterioration loss, $\mathcal{L}_{\text{regime}}$ encourages r_t to carry environment information, and $\mathcal{L}_{\text{adv}}$ is the loss of an adversary attempting to recover e from z_t. The minus sign indicates that the encoder is trained to make z_t less predictive of environment while still retaining task-relevant structure.

A complementary signal comes from consistency across worlds. If a patient trajectory is transformed by a perturbation that preserves the underlying clinical story, then the patient-state representation should not move arbitrarily. Let $o^{(j)}$ and $o^{(k)}$ be two versions of the same base trajectory under synthetic worlds e_j and e_k. One can impose

$$\mathcal{L}_{\text{cons}} = \sum_{(j,k)} \| z_t^{(j)} - z_t^{(k)} \|_2^2 \, \omega_{jk}, \tag{12}$$

where ω_{jk} weights how strongly one believes the two worlds should preserve latent clinical meaning. This does not forbid all movement [24]. It penalizes gratuitous movement of the patient-state manifold under environment changes that should not alter core severity semantics.

Another useful ingredient is adversarial benchmarking. Instead of fixing one perturbation family, one can search for transformations that maximally expose the model's dependence on environment-specific cues while preserving plausibility. Let Π_η be a parameterized perturbation generator constrained to produce clinically coherent alternate worlds. One can then solve a minimax problem

$$\min_{\theta} \max_{\eta} \quad \mathcal{L}_{\text{task}}\big(f_\theta(\Pi_\eta(o)), y\big) - \lambda_p \mathcal{C}(\eta), \tag{13}$$

where $\mathcal{C}(\eta)$ penalizes implausible perturbations. The inner maximization searches for the environment shift most damaging to the model among those that remain clinically defensible. The outer optimization then seeks representations that withstand such shifts. This is more demanding than standard data augmentation because the perturbations are not generic [25]. They are targeted probes of benchmark dependence.

There is an important subtlety here. Invariance should not be pursued blindly. Some environment-specific information is clinically appropriate. A hospital's intervention threshold may matter for predicting whether that hospital will initiate support soon. Note density may genuinely correlate with downstream action in a way that should influence a local warning system. The purpose of invariance training is therefore not to erase all environment effects. It is to separate them from the part of the representation that the field tends to mistake for universal deterioration knowledge [26]. By modeling environment explicitly, one gains the option to retain local adaptation while still knowing what is local.

One can also integrate perturbation robustness into model selection more directly. Rather than selecting the model with the best original-world validation performance, one can choose

$$\theta^\star = \arg\max_{\theta} \Big(M_0(f_\theta) - \lambda_{\text{var}} \text{Var}_j[M_j(f_\theta)] \Big), \tag{14}$$

or a risk-averse analogue that penalizes collapse under alternate worlds. Such a criterion will often prefer models that sacrifice a little in-sample peak performance for much greater stability across institutional perturbations. This is a scientifically defensible trade when the benchmark world is known to be only one realization of a broader task.

The conceptual consequence is important. Once invariance and adversarial benchmarking enter the training pipeline, the benchmark is no longer treated as a static arbiter [27]. It becomes a training environment with explicit alternate worlds. This changes the meaning of model improvement. A gain that survives under hospital-world perturbation is stronger evidence of learning something clinically portable than a larger gain confined to the original retrospective ecology. The next section takes that thought further and asks what kinds of scientific claims should be considered legitimate under monoculture-aware evaluation.

V. What a Benchmark Result Is Actually Allowed to Mean

Much of the confusion around model performance arises from overly broad interpretation. A paper trains a system on one retrospective corpus, reports strong held-out metrics, perhaps compares to a few baselines, and then the result begins to circulate as evidence that a certain architecture "predicts ICU deterioration" well. The statement is not exactly false, but it is often too broad for what the experiment has really established. Once retrospective monoculture is acknowledged, benchmark results require narrower and more structured interpretation [28].

A benchmark result licenses at least three different kinds of claims, and they should not be conflated. The first is a within-world predictive claim: under this dataset's observational regime, endpoint construction, and temporal sampling, the model achieves a certain level of performance. This claim is usually well supported. The second is a representational claim: the model has learned regularities that plausibly correspond to latent clinical dynamics rather than only to benchmark-specific artifacts. This claim is much harder to justify and requires slice analysis, perturbation stress, and some evidence of environment-stable structure. The third is a transport claim: the system is likely to remain useful when the hospital world changes. This claim is stronger still and cannot be inferred directly from the first without additional evidence.

The problem in the literature is that these claims are often allowed to bleed into one another [29]. A system with strong within-world metrics is described in language that implies representation of the clinical phenomenon itself. Later it is discussed as though external deployment were mainly an engineering hurdle rather than a shift in the task's observational ecology. This inflation of meaning is understandable, but it is methodologically costly. It encourages

the field to overinterpret retrospective gains and underinvest in understanding what exactly was learned.

A monoculture-aware methodology suggests a stricter hierarchy. The weakest claim is pure benchmark success. The next claim is stress-tested benchmark success under a family of plausible alternate worlds [30]. Only after that should one move toward claims of partial transportability. In more formal terms, if M_0 denotes original-world performance and $M_1, \ldots, M_J$ denote performance under synthetic hospital perturbations, then a result should not be described as environment-stable unless some lower-bound property such as

$$\min_{1 \le j \le J} M_j(f_\theta) \ge \kappa \tag{15}$$

or a comparable robust criterion is satisfied for a meaningful threshold κ. Even this would not prove transport to a real external hospital, but it would justify a stronger claim than ordinary held-out testing alone.

This way of thinking also changes the meaning of ablation results. Suppose adding notes improves performance in the original benchmark. Under ordinary interpretation one concludes that notes improve ICU deterioration prediction. Under a monoculture-aware interpretation one asks a narrower question: do notes improve performance in a way that survives documentation-regime perturbation, or is the gain mostly carried by one note culture's timing and density structure? The same logic applies to interventions, missingness features, or calibration methods [31]. Every gain should be treated as conditional until it has been tested against alternate worlds likely to break benchmark-specific dependence.

There is a related issue concerning null results. A model that does not improve under the original benchmark but is much more stable under hospital perturbation may be scientifically more valuable than a benchmark winner whose gain disappears immediately outside the original ecology. Current publication norms often underreward such models because the benchmark scoreboard remains the dominant object. If the field reorients around environment-stable claims, then models optimized for robustness rather than narrow benchmark fit would become easier to recognize as progress.

Scientific humility is especially important when composite endpoints and retrospective note corpora are involved, because both are rich sources of environment-specific signal. A model may learn to use them brilliantly within one corpus. That success is real and worth studying [32]. But without stress testing, the appropriate interpretation is not that the model has solved deterioration forecasting in the abstract. It is that the model has solved a particular predictive problem situated inside one observational and linguistic world. That is a meaningful result. It is also a more limited one than the field sometimes implies.

The practical purpose of narrowing claims is not rhetorical caution for its own sake. It is to align research incentives with what future deployment will require. If papers are rewarded for making stronger claims than their experiments justify, then the fastest route to publication may remain overfitting one benchmark more elegantly. If, instead, the field values results that survive alternate hospital worlds, then the benchmark stops being an arena for monocultural optimization and becomes a laboratory for discovering what is clinically stable [33].

This point is directly relevant to model selection under deployment constraints. A system that loses slightly on original-world AUROC but remains coherent under note-density shifts, threshold shifts, and alternate label boundaries may be the better system for any institution that is not the training corpus itself. The next section therefore turns from scientific interpretation to operational choice and asks how a model should be selected when benchmark dependence is treated as a first-class cost.

VI. Operational Selection Without the Illusion of Portability

Once benchmark dependence is made visible, the problem of model selection changes. The dominant habit in machine learning is to choose the model with the best validation metric or some calibrated variant of it. In ICU warning systems this often means selecting the architecture with the highest AUROC, the best AUPRC, or the strongest performance at one clinically chosen operating point. If the benchmark environment is monocultural, however, then such a rule can select models that are maximally tuned to one world while being fragile under even modest shifts. Operational selection should therefore be based on a broader utility notion that incorporates stress-tested resilience [34].

Let $M_0(f_\theta)$ be original-world utility and $M_j(f_\theta)$ utilities under synthetic hospital worlds. Let $\mathcal{R}(f_\theta)$ summarize robustness cost, for example the variance or downside risk across perturbations:

$$\mathcal{R}(f_\theta) = \frac{1}{J} \sum_{j=1}^{J} \Big(M_0(f_\theta) - M_j(f_\theta) \Big)^2, \tag{16}$$

or

$$\mathcal{R}_{\text{down}}(f_\theta) = \max_{1 \le j \le J} \Big(M_0(f_\theta) - M_j(f_\theta) \Big). \tag{17}$$

An operationally sensible selection rule might then maximize

$$U(f_\theta) = M_0(f_\theta) - \lambda_r \mathcal{R}(f_\theta) - \lambda_d \mathcal{D}(f_\theta), \tag{18}$$

where $\mathcal{D}(f_\theta)$ captures deployment-specific costs such as queue instability, support weakness in the upper tail, or explanation brittleness. This redefines "best model" in a way better aligned with the real risk of deployment.

Queue behavior deserves special emphasis because alert systems are consumed through constrained attention. A model with excellent original-world ranking but strong perturbation sensitivity may surface a very different top-K set when documentation density or observation cadence shifts. That can degrade trust faster than a modest drop in pooled discrimination. Operational selection should therefore

examine queue overlap and utility across synthetic worlds. Let $Q_K^{(0)}$ and $Q_K^{(j)}$ denote the top-K sets in the original and j-th world. One useful penalty is [35]

$$\mathcal{Q}(f_\theta) = \frac{1}{J}\sum_{j=1}^{J}\left(1 - \frac{|Q_K^{(0)} \cap Q_K^{(j)}|}{|Q_K^{(0)} \cup Q_K^{(j)}|}\right), \tag{19}$$

interpreted carefully. One does not want total invariance, because some world changes should appropriately shift the queue. But extreme queue drift under mild plausible perturbations is a warning sign that the ranking is benchmark-bound.

Threshold selection also changes under this perspective. A threshold that is optimal in the original world may be dangerously brittle under alternate worlds. Rather than choosing $\tau^\star$ to optimize one validation curve, one can solve

$$\tau^\star = \arg\max_{\tau} \min_{1\le j\le J} U_j(\tau), \tag{20}$$

where $U_j(\tau)$ is some thresholded utility in world j. This maximin policy will often produce more conservative thresholds, but the conservatism is justified if deployment into a shifted environment is genuinely expected [36]. A model whose threshold has to be retuned constantly as the note regime or endpoint semantics drift is not operationally elegant even if its pooled score remains high.

There is also an interaction with abstention. If the model estimates support, then cases lying in low-support regions under environment shift may be routed to softer actions rather than full alerts. This suggests that support should be stress-tested along with risk. A model whose top-risk cases retain high support under perturbation is operationally preferable to one whose high-risk tail becomes epistemically hollow outside the original world. Therefore a deployment utility should include not only score accuracy but support stability:

$$\mathcal{S}(f_\theta) = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\big[|q_i^{(0)} - q_i^{(j)}| \mid i \in Q_K^{(0)}\big]. \tag{21}$$

Large values indicate that the model's own confidence about important cases is highly environment dependent.

The practical result is that a model selected under benchmark monoculture and a model selected under stress-tested operational criteria may not be the same [37]. The latter may appear less triumphant in a leaderboard table. It may be exactly the model one would prefer to put in front of clinicians. This is not a compromise between science and practice. It is a correction to the habit of equating within-benchmark fit with the whole meaning of performance.

VII. Conclusion

The widespread use of retrospective benchmarks in ICU prediction has created a productive but incomplete scientific culture. It has made model comparison easier, encouraged methodological rigor, and helped the field accumulate a shared language of horizons, endpoints, calibration, and discrimination. At the same time, it has also encouraged a quiet overconfidence in what one benchmark world can stand for. A retrospective corpus is not simply a sample of patient trajectories [38]. It is a sample of one institution's way of observing, writing, escalating, and labeling those trajectories. When a model learns inside that world, it learns from all of it. Some of what it learns may be clinically stable. Some may be tied tightly to the local ecology. If the field treats the benchmark as transparent, then the difference between those two kinds of learning is easy to miss.

This paper has argued that the right response is not merely to add a warning sentence about generalizability at the end of otherwise benchmark-centered work. The right response is methodological redesign. Retrospective monoculture should be treated as a hidden regularizer that shapes both what models learn and how the field interprets their success [39]. Once that is recognized, the benchmark should no longer be trusted as a single fixed proving ground. It should be turned into a generator of alternate plausible hospital worlds. Under that regime, strong performance is no longer defined only by excellence in one ecology, but by resilience across controlled variations in measurement policy, documentation behavior, intervention thresholding, and label construction.

That shift changes the meaning of progress. A model that wins a leaderboard because it exploits the note density and escalation timing of one corpus may be less scientifically interesting than a model that does slightly worse in that world but remains coherent when the world is perturbed. A multimodal gain that survives timing shifts, note compression, and alternate label boundaries is more convincing than a larger gain that disappears under mild synthetic domain change. A calibration procedure that remains stable when intervention thresholds move is more valuable than one that is perfectly tuned to the original benchmark and nowhere else. These are not merely deployment concerns tacked onto modeling [40]. They are part of what it should mean to claim that a system has learned something durable about clinical deterioration.

There is also a deeper lesson about what the field should count as explanation. It is not enough to say that a model uses structured variables and notes in a clinically plausible way within one corpus. One must also ask whether that dependence remains sensible when the note regime changes, when density is perturbed, when endpoint boundaries shift, or when action timing moves. Explanation that is benchmark-bound is not no explanation at all, but it is a narrower scientific object than the literature often implies. A model that can explain itself only inside one documentation culture has not yet earned a broad interpretive claim.

The synthetic institutional perturbation framework proposed here is necessarily imperfect. It cannot reproduce the full richness of a real external hospital, and it cannot fully disentangle what is locally adaptive from what is spuriously benchmark-specific [41]. Yet it is still a substantial improvement over the default assumption that the benchmark environment may be ignored until some later stage of validation. It gives researchers a way to interrogate

what the model depends on before deployment. It gives them a richer basis for choosing among architectures. It also changes the standards by which claims are made. A paper that reports robust performance under a family of plausible alternate hospital worlds has shown something stronger than a paper that reports one held-out number from one retrospective ecology, even if the latter number is slightly larger.

A more mature ICU modeling culture would therefore likely look different from the current benchmark race. It would care more about environment-conditional claims, more about note-regime sensitivity, more about threshold dependence, and more about whether the queue generated by a model remains meaningful when the recording world shifts. It would value models that separate latent patient structure from environment-specific process features [42]. It would treat external validation not as the first moment at which transport becomes relevant, but as one stage in a broader transportability program that begins during benchmark construction itself. It would recognize that a dataset is not only a source of supervision. It is also a source of inductive pressure, and therefore a participant in the scientific claim.

In the long run, progress in ICU deterioration forecasting may depend less on building ever more powerful encoders atop unchanged retrospective pipelines and more on changing what those pipelines ask models to prove. A benchmark that cannot distinguish between hospital-specific pattern extraction and clinically portable learning invites the wrong race. A benchmark reimagined as a stress-tested environment family invites a better one. That race is harder to win and harder to summarize in a single number, but it is also closer to the reality in which these systems are expected to operate. For warning systems that will ultimately redistribute attention among real patients in real ICUs, that is the criterion that should matter most [43].

REFERENCES

[1] V. Sharma, J. McDermott, J. Keen, S. Foster, P. Whelan, and W. Newman, "Pharmacogenetics clinical decision support systems for primary care in england: Co-design study (preprint)," 5 2023.

[2] M. Z. I. Chowdhury, E. G. Cervantes, W.-Y. Chan, and D. Seitz, "Use of machine learning and artificial intelligence methods in geriatric mental health research involving electronic health record or administrative claims data: A systematic review.," *Frontiers in psychiatry*, vol. 12, pp. 738466–, 9 2021.

[3] C. S. Kruse, M. Mileski, V. Alaytsev, E. Carol, and A. Williams, "Adoption factors associated with electronic health record among long-term care facilities: a systematic review," *BMJ open*, vol. 5, pp. e006615–, 1 2015.

[4] B. Sadanandan, "Multimodal deep learning for early prediction of patient deterioration in the icu: Integrating time-series ehr data with clinical notes," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 10, no. 7, pp. 1–21, 2025.

[5] A. Hazazi and A. Wilson, "Leveraging electronic health records to improve management of noncommunicable diseases at primary healthcare centres in saudi arabia: a qualitative study.," *BMC family practice*, vol. 22, pp. 106–106, 5 2021.

[6] R. J. Mentz, L. K. Newby, B. Neely, J. E. Lucas, S. D. Pokorney, M. P. Rao, L. R. Jackson, M. V. Grau-Sepulveda, M. M. Smerek, P. Barth, C. L. Nelson, M. J. Pencina, and B. R. Shah, "Assessment of administrative data to identify acute myocardial infarction in electronic health records.," *Journal of the American College of Cardiology*, vol. 67, pp. 2441–2442, 5 2016.

[7] K. Marsolo, P. A. Margolis, C. B. Forrest, R. B. Colletti, and J. J. Hutton, "A digital architecture for a network-based learning health system: Integrating chronic care management, quality improvement, and research.," *EGEMS (Washington, DC)*, vol. 3, pp. 1168–1168, 8 2015.

[8] S. Nijor, G. Rallis, N. Lad, and E. Gokcen, "Patient safety issues from information overload in electronic medical records.," *Journal of patient safety*, vol. 18, pp. e999–e1003, 4 2022.

[9] L. Ohno-Machado, "Where do we stand in the maze of health information systems," *Journal of the American Medical Informatics Association*, vol. 19, pp. 497–497, 7 2012.

[10] C. U. Lehmann, "Pediatric aspects of inpatient health information technology systems," *Pediatrics*, vol. 122, pp. e1287–96, 12 2008.

[11] A. Mandal, P. Dumaru, S. Bhandari, S. Shrestha, and S. Shakya, "Decentralized electronic health record system," *Journal of the Institute of Engineering*, vol. 15, pp. 77–80, 2 2020.

[12] C. Sze, V. Simonyan, A. Sedrakyan, and B. Chughtai, "Where is research in the era of electronic health records?," *The Permanente journal*, vol. 26, pp. 154–156, 7 2022.

[13] K. J. Wells, J. R. Gordon, H. I. Su, S. Plosker, and G. P. Quinn, "Ethical considerations for informed consent in infertility research: The use of electronic health records.," *Advances in medical ethics*, vol. 2, pp. 1–5, 12 2015.

[14] E. J. Thompson, D. M. Williams, A. J. Walker, R. E. Mitchell, C. L. Niedzwiedz, T. Yang, C. F. Huggins, A. S. F. Kwong, R. J. Silverwood, G. D. Gessa, R. C. E. Bowyer, K. Northstone, B. Hou, M. J. Green, B. Dodgeon, K. J. Doores, E. L. Duncan, F. M. K. Williams, O. Collaborative, A. Steptoe, D. J. Porteous, R. R. C. McEachan, L. A. Tomlinson, B. Goldacre, P. Patalay, G. B. Ploubidis, S. V. Katikireddi, K. Tilling, C. T. Rentsch, N. J. Timpson, N. Chaturvedi, and C. J. Steves, "Risk factors for long covid: analyses of 10 longitudinal studies and electronic health records in the

uk," 6 2021.

[15] A. Hentschel, C. J. Hsiao, L. Y. Chen, L. Wright, J. Shaw, X. Du, E. Flood-Grady, C. A. Harle, C. F. Reeder, M. Francois, A. Louis-Jacques, E. Shenkman, J. L. Krieger, and D. J. Lemas, "Perspectives of pregnant and breastfeeding women on participating in longitudinal mother-baby studies involving electronic health records: Qualitative study (preprint)," 8 2020.

[16] A. Rajaram, D. Thomas, F. Sallam, A. A. Verma, and S. Rawal, "Accuracy of the preferred language field in the electronic health records of two canadian hospitals.," *Applied clinical informatics*, vol. 11, pp. 644–649, 9 2020.

[17] D. Agniel, I. S. Kohane, and G. M. Weber, "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study.," *BMJ (Clinical research ed.)*, vol. 361, pp. k1479–, 4 2018.

[18] F. Alsohime, M.-H. Temsah, A. Al-Eyadhy, F. A. Bashiri, M. Househ, A. Jamal, G. M. Hasan, A. Alhaboob, M. Alabdulhafid, and Y. S. Amer, "Satisfaction and perceived usefulness with newly-implemented electronic health records system among pediatricians at a university hospital.," *Computer methods and programs in biomedicine*, vol. 169, pp. 51–57, 12 2018.

[19] I. Bayramli, V. Castro, Y. Barak-Corren, E. M. Madsen, M. K. Nock, J. W. Smoller, and B. Y. Reis, "Predictive structured-unstructured interactions in ehr models: A case study of suicide prediction.," *NPJ digital medicine*, vol. 5, pp. 15–15, 1 2022.

[20] J. R. Enriquez, J. A. de Lemos, S. V. Parikh, D. N. Simon, L. Thomas, T. Y. Wang, P. S. Chan, J. A. Spertus, and S. R. Das, "Modest associations between electronic health record use and acute myocardial infarction quality of care and outcomes results from the national cardiovascular data registry," *Circulation. Cardiovascular quality and outcomes*, vol. 8, pp. 576–585, 10 2015.

[21] S. E. Spratt, K. Pereira, B. B. Granger, B. C. Batch, M. Phelan, M. J. Pencina, M. L. Miranda, E. Boulware, J. E. Lucas, C. L. Nelson, B. Neely, B. A. Goldstein, P. Barth, R. L. Richesson, I. L. Riley, L. Corsino, E. R. M. Hinz, S. A. Rusincovitch, J. B. Green, A. B. Barton, C. E. Kelley, K. Hyland, M. Tang, A. Elliott, E. Ruel, A. Clark, M. Mabrey, K. L. Morrissey, J. P. Rao, B. Hong, M. Pierre-Louis, K. Kelly, and N. E. Jelesoff, "Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 24, pp. e121–e128, 9 2016.

[22] P. Dhawan, A. Rastogi, and D. Saisanthiya, "Revolutionizing healthcare with blockchain based record keeping," in *2023 World Conference on Communication & Computing (WCONF)*, pp. 1–6, IEEE, 7 2023.

[23] T. C. Kariotis, M. Prictor, S. Chang, and K. Gray, "Impact of electronic health records on information practices in mental health contexts: Scoping review (preprint)," 5 2021.

[24] R. K. Johnson, K. M. Marker, D. Mayer, J. Shortt, D. Kao, K. C. Barnes, J. T. Lowery, and C. R. Gignoux, "Covid-19 surveillance in the biobank at the colorado center for personalized medicine: Observational study (preprint)," 2 2022.

[25] M. Noaeen, S. Amini, S. Bhasker, Z. Ghezelsefli, A. Ahmed, O. Jafarinezhad, and Z. S. H. Abad, "Unlocking the power of ehrs: Harnessing unstructured data for machine learning-based outcome predictions," 2 2023.

[26] S. Cui, J. Wang, X. Gui, T. Wang, and F. Ma, "Automed: Automated medical risk predictive modeling on electronic health records.," *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, vol. 2022, pp. 948–953, 12 2022.

[27] J. Piasecki, E. Walkiewicz-Żarek, J. Figas-Skrzypulec, A. Kordecka, and V. Dranseika, "Ethical issues in biomedical research using electronic health records: a systematic review.," *Medicine, health care, and philosophy*, vol. 24, pp. 1–26, 6 2021.

[28] S. Denaxas and K. I. Morley, "Big biomedical data and cardiovascular disease research: opportunities and challenges," *European heart journal. Quality of care & clinical outcomes*, vol. 1, pp. 9–16, 6 2015.

[29] T. Quaini, A. Roehrs, C. A. da Costa, and R. da Rosa Righi, "A model for blockchain-based distributed electronic health records," *IADIS INTERNATIONAL JOURNAL ON WWW/INTERNET*, vol. 16, pp. 66–79, 12 2018.

[30] C. Yan, Z. Zhang, S. Nyemba, and Z. Li, "Generating synthetic electronic health record data using generative adversarial networks: Tutorial (preprint)," 9 2023.

[31] T. K. Colicchio, W. H. Liang, P. I. Dissanayake, C. V. D. Rosario, and J. J. Cimino, "Physicians' perceptions about a semantically integrated display for chart review: A multi-specialty survey.," *International journal of medical informatics*, vol. 163, pp. 104788–104788, 5 2022.

[32] J. McNeely, S. D. Gallagher, M. Mazumdar, N. Appleton, J. Fernando, E. Owens, E. Bone, N. Krawczyk, J. Dolle, R. K. Marcello, J. Billings, and S. Wang, "Sensitivity of medicaid claims data for identifying opioid use disorder in patients admitted to 6 new york city public hospitals.," *Journal of addiction medicine*, vol. 17, pp. 339–341, 10 2022.

[33] T. Xu, Y. Chen, D. Zeng, and Y. Wang, "Self-matched learning to construct treatment decision rules from electronic health records.," *Statistics in medicine*, vol. 41, pp. 3434–3447, 5 2022.

[34] J. S. Wittenborn, A. Y. Lee, E. A. Lundeen, P. Lamuda, J. Saaddine, G. L. Su, R. Lu, A. Damani, J. S. Za-

wadzki, C. P. Froines, J. Z. Shen, T.-P. H. Kung, R. T. Yanagihara, M. Maring, M. M. Takahashi, M. Blazes, and D. B. Rein, "Validity of administrative claims and electronic health registry data from a single practice for eye health surveillance.," *JAMA ophthalmology*, vol. 141, pp. 534–534, 6 2023.

[35] M. F. Ansari, B. Dash, S. Swayamsiddha, and G. Panda, "Use of blockchain technology to protect privacy in electronic health records- a review," in *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 144–149, IEEE, 1 2023.

[36] M. A. Miller and C. Stoeckert, "Semantic triples from "a collaborative, realism-based, electronic healthcare graph: Public data, common data models, and practical instantiation"," 4 2019.

[37] G. Alfiansyah, A. S. Fajeri, M. W. Santi, and S. J. Swari, "Evaluasi kepuasan pengguna electronic health record (ehr) menggunakan metode eucs (end user computing satisfaction) di unit rekam medis pusat rsupn dr. cipto mangunkusumo," *Jurnal Penelitian Kesehatan "SUARA FORIKES" (Journal of Health Research "Forikes Voice")*, vol. 11, pp. 258–263, 4 2020.

[38] C. L. M. Joseph, A. Tang, D. W. Chesla, M. M. Epstein, P. A. Pawloski, A. B. Stevens, S. C. Waring, B. K. Ahmedani, C. C. Johnson, and C. D. Peltz-Rauchman, "Demographic differences in willingness to share electronic health records in the all of us research program.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 29, pp. 1271–1278, 4 2022.

[39] G. W. Daniel, E. Ewen, V. J. Willey, C. L. Reese, F. Shirazi, and D. C. Malone, "Efficiency and economic benefits of a payer-based electronic health record in an emergency department.," *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, vol. 17, pp. 824–833, 7 2010.

[40] P. K. Sari and S. Yazid, *IWBIS - Design of Blockchain-based Electronic Health Records for Indonesian Context: Narrative Review*. IEEE, 10 2020.

[41] S. Y. Jung, H. Hwang, K. Lee, D. Lee, S. Yoo, K. Lim, H.-Y. Lee, and E. Kim, "User perspectives on barriers and facilitators to the implementation of electronic health records in behavioral hospitals: Qualitative study (preprint)," 3 2020.

[42] A. J. Holmgren, M. Kuznetsova, D. C. Classen, and D. W. Bates, "Assessing hospital electronic health record vendor performance across publicly reported quality measures.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 28, pp. 2101–2107, 8 2021.

[43] S. E. Perlman, K. H. McVeigh, L. E. Thorpe, L. E. Jacobson, C. M. Greene, and R. C. Gwynn, "Innovations in population health surveillance: Using electronic health records for chronic disease surveillance.," *American journal of public health*, vol. 107, pp. 853–857, 4 2017.