Auditability Gaps in ICU Deterioration Models Through Slice-Stable Evaluation Attribution Reliability and Deployment-Facing Error Discovery

Santiago Andrés Rojas[1], **Daniel Esteban Cárdenas**[2], **and Miguel Ángel Herrera**[3]
[1]Departamento de Ingeniería de Sistemas, Universidad de la Amazonia, Calle 17 Diagonal 17 con Carrera 3F, Barrio Porvenir, Florencia, Caquetá, Colombia
[2]Programa de Ingeniería Informática, Universidad de Córdoba, Carrera 6 No. 76–103, Barrio Mocarí, Montería, Córdoba, Colombia
[3]Facultad de Ingeniería y Tecnología, Universidad de los Llanos, Kilómetro 12 Vía Puerto López, Vereda Barcelona, Villavicencio, Meta, Colombia

ABSTRACT Evaluation in intensive care forecasting is usually organized around a small set of scalar metrics such as AUROC, AUPRC, calibration error, and thresholded sensitivity. These summaries are useful, but they often fail to reveal where a warning system is brittle, which patient-time regimes dominate its apparent success, and how model behavior changes when evidence becomes sparse, delayed, or semantically inconsistent. As a result, systems that appear strong in retrospective benchmarking can remain poorly understood in the exact regimes where clinicians would need reliability most. This paper develops an audit-centered framework for ICU deterioration modeling in which evaluation is treated not as a final scoring step, but as a structured inference problem over hidden failure regions. The central idea is that error should be decomposed over temporally local, semantically coherent, and operationally meaningful slices rather than averaged across the full population of prediction windows. To support this goal, the paper introduces mathematical tools for slice discovery, temporal attribution stability, confidence-conditioned review utility, and deployment-facing risk decomposition under alert budgets. The framework also examines how repeated-window sampling, severe class imbalance, and institutional workflow patterns can create misleading impressions of generalization when only aggregate metrics are reported. By linking evaluation to hidden-state structure, support quality, and review constraints, the paper argues for a shift from score-centric benchmarking toward auditability as a first-class property of warning systems. Under this view, a useful ICU predictor is not simply one that ranks positives above negatives on average, but one whose failures are legible, whose explanations are stable, and whose deployment thresholds remain defensible across clinically distinct regions of the data.

KEYWORDS

1. Introduction

The machine learning literature on ICU deterioration has become rich in architectures and increasingly polished in benchmarking [1]. Recurrent encoders, attention-based fusion systems, transformer variants, and a range of calibration and thresholding strategies have all been brought to the problem of identifying patients who may decline over the next several hours. This progress has been useful [2]. Yet the field has also converged on a mode of validation that is narrower than the underlying clinical task. A deterioration model is usually considered strong if it achieves respectable global discrimination, tolerable calibration after post-hoc adjustment, and an acceptable tradeoff between

Article info:
Santiago Andrés Rojas, Daniel Esteban Cárdenas, and Miguel Ángel Herrera. *Auditability Gaps in ICU Deterioration Models Through Slice-Stable Evaluation Attribution Reliability and Deployment-Facing Error Discovery*. Grovesocieties .

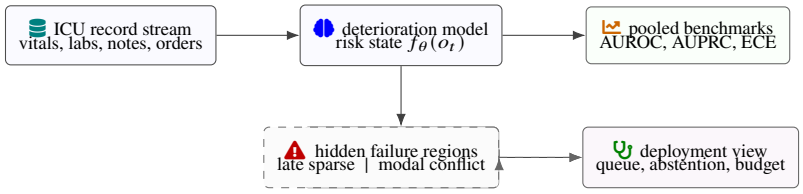

Figure 1 Audit-centered framing for ICU deterioration modeling. Instead of treating evaluation as the final production of pooled scores, the model output is routed into a second layer of analysis that exposes failure slices, explanation stability, and deployment-facing risk before thresholds are trusted.

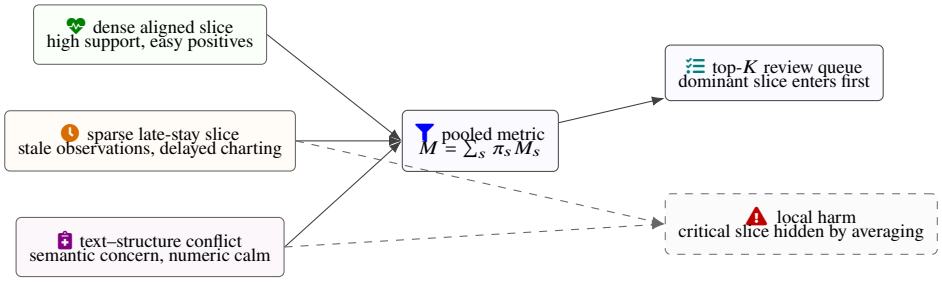

Figure 2 Aggregate metrics compress heterogeneous patient-time regimes into a single expectation, allowing large or easy slices to dominate the reported score. The clinically relevant question is not only whether the pooled metric is strong, but whether difficult slices with weak support or conflicting modalities are being buried inside that average while the top-K queue is shaped by easier regimes.

sensitivity and alert burden at one or more operating points [3]. Those are necessary properties. They are not sufficient properties. They say little about what kinds of patient-time regimes generated the apparent success, what kinds of regimes concentrate the remaining failures, whether the model remains stable under semantically minor perturbations of the chart, or how brittle the explanation machinery becomes precisely in the high-risk tail that attracts human attention [4].

This gap matters because intensive care forecasting is not consumed as a purely statistical exercise. Scores are used to sort patients into concern strata, prioritize chart review, justify heightened surveillance, or trigger escalation policies [5]. In each of these settings, the true question is not only whether the model ranks positives above negatives on average. The practical question is whether the ranking remains trustworthy in specific regions of the data space where the cost of error is high and the evidential basis of prediction can be weak. A model that performs well overall but fails systematically among sparse late-night records, among patients with unusual intervention histories, or among cases whose textual and numeric channels disagree is not simply a model with an imperfect ROC curve. It is a model whose operational meaning changes across hidden slices of the population. If those slices are not identified, then apparent global competence can coexist with local unreliability exactly where clinical deployment would concentrate attention.

Rather than presenting the reported test metrics as the end of the scientific story, Sadanandan (2025) explicitly notes that feature attribution and subgroup analyses remained incomplete, leaving unresolved both what the network was relying on and how performance might vary across clinically meaningful populations [6]. That admission is important because it shows that strong benchmark numbers do not automatically answer the most consequential questions about a warning system. A model can outperform baselines and still leave its internal dependence structure, its slice-wise fragility, and its fairness under heterogeneous documentation or case mix largely unexamined. In that sense, incomplete audit is not a minor appendage to otherwise successful prediction [7]. It is a major source of uncertainty about what exactly has been learned.

This paper takes that observation as a starting point and argues that auditability should be treated as a primary design target in ICU deterioration modeling. The proposed shift is not from modeling to evaluation, but from score-centric evaluation to structure-aware evaluation. The central claim is that retrospective validation should be understood as an inference problem over hidden failure regions. Instead of asking only for one global estimate of predictive strength, the evaluation should infer where the model's risk surface is smooth or unstable, where confidence is justified or hollow, where explanations are robust or opportunistic, and how these properties interact with operational choices such as alert budgets or review capacity. Under this view, an ideal evaluation protocol does not only summarize the average height of a decision surface. It discovers the topography of its weaknesses.

Several aspects of ICU data make this particularly urgent. First, the positive class is rare under hourly prediction, which means that a large fraction of model utility is concentrated in a small upper tail of the score distribution. Second, repeated-window sampling creates strong local dependence, so a model may appear stable or accurate simply because similar windows recur around the same episode [8]. Third, the chart is generated through a care process whose observation patterns are not random. Missingness, note frequency, intervention timing, and measurement density are all entangled with severity and workflow. These properties mean that hidden failure regions are unlikely to be defined by simple demographic groups alone.

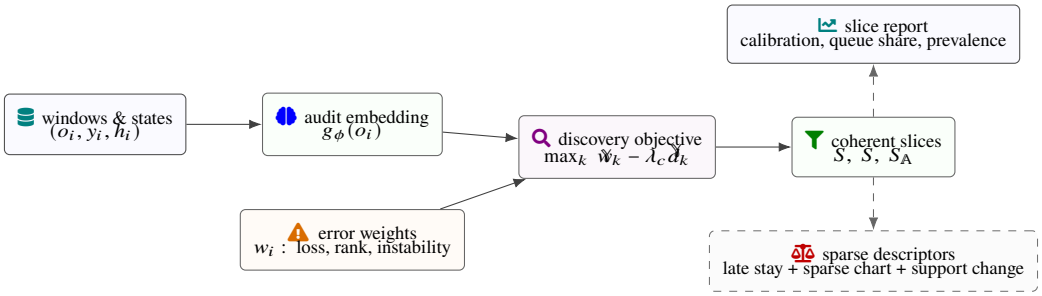

Figure 3 Failure slices are cast as coherent high-error regions rather than manually chosen subgroups. The discovery step combines an audit representation with error weights so that the resulting regions are not merely hard cases, but interpretable patient-time regimes where calibration error, review misallocation, or attribution instability accumulates in a concentrated way.

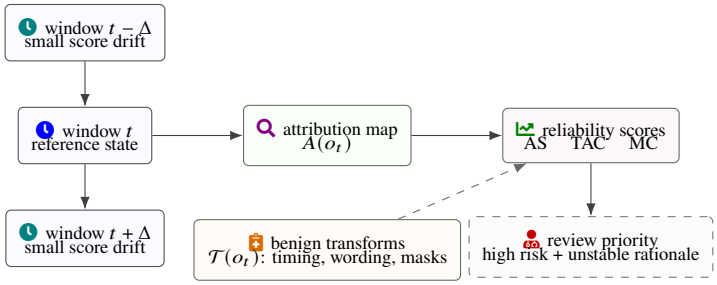

Figure 4 Attribution reliability is treated as a temporal and perturbational stability problem rather than a one-off visualization exercise. Explanations are considered more trustworthy when they remain coherent across adjacent windows with small score change and across semantics-preserving perturbations, while low stability flags patient-time windows where manual review should be more cautious.

They are often temporal, operational, or multimodal in character. A model may behave differently for early-stay versus late-stay windows, for highly monitored versus thinly monitored patients, or for windows in which textual and physiologic evidence conflict. Such slices are rarely exposed by generic train-test metrics.

The rest of the paper develops a framework for making those hidden regions visible. The next section formalizes why aggregate metrics can hide clinically important error concentrations and introduces the idea of evaluation as decomposition over latent failure slices. After that, the discussion turns to slice discovery itself, treating the search for brittle subpopulations as an optimization problem rather than a manually curated checklist. A later section argues that explanation quality should be judged by stability under time-preserving and semantics-preserving perturbations, not by saliency magnitude alone [9]. The paper then links evaluation to deployment by analyzing how alert thresholds, abstention rules, and review budgets interact with slice-wise calibration and failure concentration. The penultimate section proposes an audit-first protocol for model development, selection, and post-deployment monitoring. The conclusion takes a broader view and argues that the field's heavy reliance on a few scalar summaries has obscured a deeper question: not whether the model works on average, but whether its behavior is sufficiently legible to deserve trust when averages are no longer informative.

2. What Aggregate Metrics Erase

A single performance number is always an act of compression. AUROC integrates over thresholds and over heterogeneous regions of the data space. AUPRC, though more sensitive to the rare positive class, still averages over many ranking contexts. Even calibration error, which is supposed to say something about the meaning of predicted probabilities, is usually reported as one global discrepancy between forecasted and empirical frequencies. These metrics are informative only insofar as one accepts the averaging operation that produced them. In ICU deterioration forecasting, that averaging step can be especially destructive because the data space is not homogeneous in either clinical meaning or record quality.

Let $\mathcal{D} = \{(o_i, y_i)\}_{i=}^{N}$ denote the set of patient-time windows, where o_i is the available record and $y_i \in \{\varkappa, \}$ indicates whether the endpoint lies within a chosen future horizon. Let $f_\theta(o_i) \in [\varkappa,]$ be the model's risk score [10]. A global performance functional $M(f_\theta, \mathcal{D})$ can always be written as an expectation over the data distribution,

$$M(f_\theta, \mathcal{D}) = \mathbb{E}_{(o,y)\sim P}\big[m(f_\theta(o), y)\big], \tag{1}$$

for some local scoring rule m. What is hidden in this notation is that the data distribution P is itself a mixture over regions of the record space that are clinically and operationally distinct. If one introduces a latent slice variable $s \in \{, \ldots, S\}$, then

$$P(o, y) = \sum_{s=}^{S} \pi_s P_s(o, y), \tag{2}$$

and the same metric becomes

$$M(f_\theta, \mathcal{D}) = \sum_{s=}^{S} \pi_s \, \mathbb{E}_{(o,y)\sim P_s}\big[m(f_\theta(o), y)\big]. \tag{3}$$

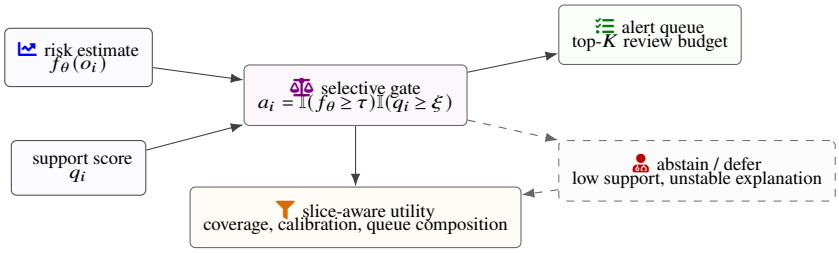

Figure 5 Deployment-facing evaluation adds support and abstention to the usual thresholding problem. A score becomes clinically actionable only after it passes a support gate, and the resulting alert stream is then judged not just by how many events it captures, but by whether slice-wise coverage, queue composition, and abstentions remain defensible under finite review capacity.

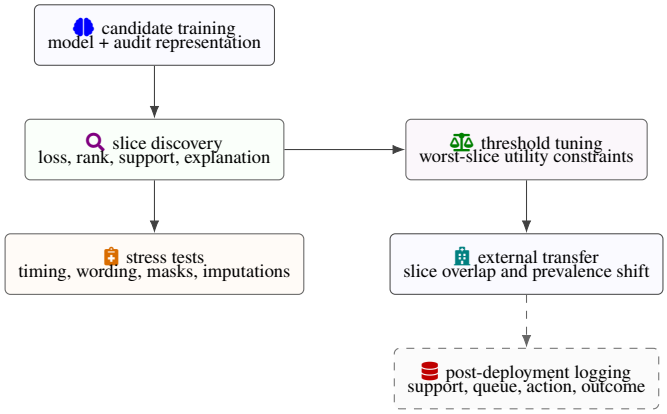

Figure 6 An audit-first development loop makes slice discovery part of model selection rather than a late interpretability appendix. Candidate systems are compared not only by pooled discrimination but also by the stability of their failure map under perturbation, the conservativeness of their threshold policy across discovered slices, and the behavior of those slices after transfer into a new site and then into routine logging.

The weights π_s determine what the global metric emphasizes. Large or easy slices dominate. Small or clinically costly slices disappear into the average unless their error is catastrophic enough to move the aggregate.

The problem is not simply that rare subgroups are masked. In ICU forecasting, hidden slices are often not static subgroups but patient-time regimes. A late-stay window with sparse measurements and recent vasopressor titration is a different inferential situation from an early-stay window with dense laboratory data and no support. A window whose notes mention concern but whose structured values appear temporarily stable is different from one with aligned textual and numeric distress [11]. These regimes can have different base rates, different separability, and different relevance for action. When a metric averages over them, it implicitly treats them as exchangeable units of evaluation. That is rarely what the clinical setting requires.

One can make this more concrete by decomposing a standard loss such as the Brier score. Let

$$\mathcal{B} = \frac{}{N} \sum_{i=}^{N} \big(f_\theta(o_i) - y_i\big). \tag{4}$$

If the data are partitioned into slices $S, \ldots, S_K$, then

$$\mathcal{B} = \sum_{k=}^{K} \frac{|S_k|}{N} \mathcal{B}_k, \quad \mathcal{B}_k = \frac{}{|S_k|} \sum_{i \in S_k} \big(f_\theta(o_i) - y_i\big). \tag{5}$$

A global improvement can therefore arise from reducing error in a large easy slice while leaving a small but operationally critical slice unchanged or worse. The same logic applies to AUROC and AUPRC, except the dependence on slice size and class distribution is less transparent because the metrics are ranking-based rather than pointwise.

The upper tail of the risk distribution intensifies the issue. In practice, clinicians do not consume average performance. They consume the subset of patients above an implicit or explicit concern boundary [12]. Let Q_K be the top-K set under the model's ranking. The utility of the system depends disproportionately on the composition of Q_K, on the kinds of patients who enter it, and on how stable that set remains under semantically small perturbations of the chart. A model may have excellent global discrimination and still create a top queue dominated by one easy slice of positives while missing a harder, more subtle slice whose windows are operationally valuable but underrepresented. Aggregate metrics will not expose this unless one conditions on slice.

There is a second erasure hidden in global metrics: support quality. Two windows may receive the same risk score while differing sharply in what evidence supports that score. One may be based on fresh aligned measurements and recent contextual notes. The other may arise from sparse measurements, stale documentation, and latent correlations learned from similar trajectories. The score can be identical while the epistemic status is not. If evaluation does not model support explicitly,

Component	Definition	Role in Evaluation	Practical Implication
Observation o_i	Patient-time record window	Input to prediction model	Encodes physiological and textual state
Outcome y_i	Binary deterioration indicator	Ground truth label	Defines prediction target
Risk score $f_\theta(o_i)$	Model output probability	Basis for ranking and alerts	Determines clinical prioritization
Local metric $m(\cdot)$	Window-level scoring rule	Measures prediction error	Used in global metric aggregation
Dataset $\mathcal{D}$	Collection of windows	Evaluation domain	Supports retrospective validation

Table 1 Basic elements of ICU deterioration model evaluation and their functional roles.

Symbol	Meaning	Interpretation	Evaluation Impact
Slice s	Latent subgroup of windows	Coherent patient-time regime	Reveals hidden error concentration
π_s	Slice prevalence	Fraction of data in slice	Determines contribution to averages
$P_s(o, y)$	Slice distribution	Conditional data structure	Captures regime-specific behavior
Global metric M	Aggregated score	Weighted mixture over slices	May obscure minority failures
Slice metric M_s	Local performance value	Metric within slice	Enables targeted auditing

Table 2 Mixture-based interpretation of evaluation metrics across latent data slices.

then calibration and discrimination are assessed as though all equal scores had equal meaning [13]. That assumption is especially fragile in multimodal ICU models, where the relative freshness and availability of evidence channels vary widely.

A third erasure concerns explanation. Saliency or attribution maps are frequently inspected only for individual examples or averaged across many examples. Such inspection can be reassuring without being informative. An explanation method that highlights plausible variables on typical cases may still be unstable within hidden slices. It may assign meaningfully different explanations to near-identical windows depending on local missingness or note timing. If evaluation does not ask whether attribution is stable within clinically coherent regimes, then explanation quality is again averaged into a comfortingly vague summary.

The larger lesson is that aggregate metrics are blind to error topology. They give the height of the mountain but not the location of cliffs. What an ICU warning system needs, however, is not only evidence that the mountain is tall on average, but a map of where the cliffs are [14]. That requires discovering slices rather than assuming them. The next section turns that requirement into a formal objective.

3. Failure Slices as Objects to Be Discovered

The standard approach to subgroup evaluation is to choose a few clinically familiar covariates, such as age, sex, ICU type, or note availability, and then report performance stratified by them. This is useful, but it is not enough. The most damaging slices in ICU forecasting are often not aligned with one observed covariate. They are intersections of temporal context, evidence density, intervention pattern, and model representation. A brittle regime may be characterized by late-stay sparse measurement plus escalating support plus stale narrative context. Another may be characterized by high textual concern with temporarily normal structured values. These are not easily enumerated by hand. If slice discovery is left entirely to intuition, the evaluation will remain partial [15].

A more principled approach is to define failure slices as latent subsets of the data space that maximize some notion of error concentration subject to support and coherence constraints. Let $g_\phi(o_i) \in \mathbb{R}^d$ be an audit representation of window o_i. This representation can be derived from the model's hidden state, from clinically meaningful summary features, or from a combination of both. Let w_i be a local error weight, for example absolute calibration error, squared loss, or contribution to review misallocation. Then the task is to discover subsets $S \subseteq \{, \ldots, N\}$ that are internally coherent in g_ϕ-space and

Element	Mathematical Form	Purpose	Diagnostic Value
Audit representation	$g_\phi(o_i)$	Encodes window characteristics	Enables slice discovery
Error weight	w_i	Local contribution to failure	Guides optimization focus
Slice centroid	μ_k	Prototype of slice region	Maintains slice coherence
Membership weight	u_{ik}	Soft assignment to slice	Allows overlapping regimes
Compactness term	$\|g_\phi(o_i) - \mu_k\|$	Penalizes diffuse slices	Ensures interpretability

Table 3 Core variables used when identifying coherent regions of concentrated prediction error.

Error Weight Choice	Measurement Basis	Captured Failure Type	Example Use Case
Squared loss	$(f_\theta - y)$	Calibration error	Probability reliability analysis
Rank inversion	Ordering mistakes	Queue misprioritization	Top-alert evaluation
Support deficit	Low evidence score	Weak informational basis	Sparse record detection
Attribution variance	Explanation changes	Unstable reasoning	Interpretation auditing
False alert cost	Outcome-conditioned penalty	Operational burden	Alert fatigue analysis

Table 4 Alternative definitions of local error signals used during slice discovery.

exhibit elevated weighted error.

One can formulate this as an optimization over soft memberships $u_{ik} \in [\varkappa,]$ for K slices:

$$\begin{aligned} \max_{u,\mu} \quad & \sum_{k=}^{K} \left[\frac{\sum_{i=}^{N} u_{ik} w_i}{\sum_{i=}^{N} u_{ik}} - \lambda_c \frac{\sum_{i=}^{N} u_{ik} \|g_\phi(o_i) - \mu_k\|}{\sum_{i=}^{N} u_{ik}} \right] \\ \text{s.t.} \quad & \sum_{k=}^{K} u_{ik} = \quad \hbar i, \end{aligned} \tag{6}$$

where μ_k are slice centroids and λ_c controls compactness. The first term seeks error-rich slices. The second prevents them from becoming arbitrary collections of difficult examples with no coherent identity. The resulting slices are not simply clusters and not simply worst-case examples. They are coherent high-error regions.

The choice of w_i determines what kind of slice is discovered. If w_i is squared loss, one finds calibration failures [16]. If w_i reflects rank inversions in the upper tail, one finds queue failures. If w_i measures disagreement between explanation maps under perturbation, one finds attribution-instability slices. In practice one should perform several such slice-discovery passes because no single error notion captures all deployment-relevant failure.

Because slices are intended to support human understanding, coherence should not be measured only geometrically. One can encourage descriptive sparsity by learning simple post hoc rules over observable features that approximate the soft memberships. Suppose $h_k(o_i) \in \{\varkappa, \}$ is a sparse rule for slice k. One can add a fidelity term

$$\mathcal{L}_{\text{rule}} = \sum_{k=}^{K} \sum_{i=}^{N} \big(u_{ik} - h_k(o_i)\big) + \lambda_r \,\text{complexity}(h_k). \tag{7}$$

This does not force slices to be rule-based. It ensures that discovered slices can be approximately described in terms clinicians and modelers can inspect. Such descriptions are often more useful than arbitrary latent clusters when designing safeguards or deployment policies.

Temporal structure should also enter slice discovery [17]. A patient-time regime can be defined not only by the current hidden state but by its recent motion. Let $g_\phi(o_i)$ include derivatives or short-lag summaries, or let a separate motion representation $m_\phi(o_i)$ be appended. Slices can then capture not only static cases but transition regimes such as rapidly rising risk under sparse evidence or prolonged high-risk plateaus under dense observation. This is essential in ICU settings because some failures concern not where the score is, but how it changes.

Another important issue is slice support. A discovered slice with very high error but negligible mass may be statistically unstable. One therefore needs lower-support constraints or uncertainty estimates on slice-level metrics. A bootstrap over windows or patient episodes can provide confidence intervals for slice prevalence and error concentration. More ambitiously, one

Term	Description	Measurement Method	Insight Provided
Attribution vector $A(o)$	Feature importance distribution	Saliency or gradient method	Model reasoning snapshot
Transformation set $\mathcal{T}(o)$	Semantics-preserving edits	Note or timing perturbations	Tests explanation invariance
Stability score	Attribution similarity	Norm difference ratio	Reliability indicator
Perturbed example	Modified window $\tilde{o}$	Controlled input change	Stress test for explanations
Stability threshold	Minimum acceptable score	Empirical cutoff	Flags fragile explanations

Table 5 Key elements used to measure explanation robustness under benign perturbations.

Quantity	Definition	Context	Operational Meaning
Window pair	$(o_{i,t}, o_{i,t+\Delta})$	Consecutive predictions	Captures short-term dynamics
Score change	$\lvert f_t - f_{t+\Delta}\rvert$	Risk variation	Detects stability of prediction
Attribution shift	$\lVert A_t - A_{t+\Delta}\rVert$	Explanation movement	Identifies reasoning changes
Continuity score	Temporal similarity metric	Stability estimate	Ensures coherent explanations
Threshold τ	Maximum score change	Valid comparison condition	Filters noisy transitions

Table 6 Temporal continuity measures used to assess whether explanations evolve smoothly over time.

can include a lower-bound constraint directly in the optimization:

$$\sum_{i=}^{N} u_{ik} \geq \delta N \quad \hbar k, \tag{8}$$

for some minimal mass δ [18]. This avoids overfitting the evaluation to tiny idiosyncratic pockets.

Once slices are discovered, the evaluation becomes a table or map of local behavior rather than a handful of scalar summaries. One can report slice-wise discrimination, calibration, attribution stability, and queue inclusion. More importantly, one can inspect whether slices correspond to clinically meaningful contexts. Do high-error slices cluster around late-stay sparse-record cases? Around conflicting text-structure windows? Around post-intervention trajectories? Such answers are far more informative for deployment than another decimal point on a global AUROC.

This slice-discovery perspective also has implications for model training. If certain failure slices recur across architectures, they may reflect data or target issues more than model weakness. If they shift dramatically with architecture, they may reveal which representations are brittle. In either case, evaluation becomes diagnostic rather than ceremonial. That is the core shift the paper advocates [19].

The next section moves from score error to explanation quality and argues that attribution should itself be treated as an object of slice-wise stability analysis rather than as a static visualization tool.

4. Attribution Reliability as a Temporal Stability Problem

An explanation is only as trustworthy as its invariances. In ICU forecasting, explanation methods are often used to reassure clinicians or modelers that the network attends to physiologically plausible variables or to semantically relevant note content. Yet plausibility on a few handpicked examples is a weak criterion. What matters more is whether the explanation remains stable when the clinical content is preserved but the chart is perturbed in ways that should not change the model's reasoning substantially. If the explanation swings wildly across such perturbations, then the model may be using brittle shortcuts even when the output score itself remains similar.

Let $A(o_i) \in \mathbb{R}^p$ denote an attribution vector over p features, time steps, or note spans for window o_i. Let $\mathcal{T}(o_i)$ be a set of semantics-preserving transformations, such as minor note paraphrasing, removal of clearly redundant tokens, small timing shifts that preserve admissibility, or alternate imputations that leave observed values unchanged. Attribution reliability should measure how consistent $A(o_i)$ remains over $\mathcal{T}(o_i)$, especially when the model score $f_\theta(o_i)$ changes little.

Variable	Meaning	Policy Role	Deployment Effect
Risk score $f_\theta(o)$	Predicted deterioration probability	Alert ranking basis	Determines escalation priority
Support score q_i	Evidence reliability estimate	Confidence filter	Prevents weak alerts
Risk threshold τ	Minimum risk for alert	Decision boundary	Controls sensitivity
Support threshold ξ	Minimum support requirement	Abstention mechanism	Avoids unsupported decisions
Action indicator a_i	Final alert decision	Workflow signal	Determines review trigger

Table 7 Selective prediction variables used to combine risk estimation with evidence support.

Term	Interpretation	Contribution	Practical Meaning
True event reward α_y	Value of catching deterioration	Positive utility	Encourages detection
False alert penalty α_f	Cost of incorrect alert	Negative utility	Controls alert burden
Support penalty α_s	Cost of weak evidence	Reliability adjustment	Reduces fragile alerts
Explanation penalty α_a	Attribution instability cost	Interpretability factor	Discourages opaque predictions
Queue set Q_K	Top-K ranked cases	Alert budget constraint	Reflects review capacity

Table 8 Utility components used when evaluating prioritized alert queues under limited review capacity.

A simple stability metric is

$$\mathrm{AS}(o_i) = -\frac{}{|\mathcal{T}(o_i)|}\sum_{\in\mathcal{T}(o_i)}\frac{\|A(o_i)-A(\)\|}{\|A(o_i)\|+\varepsilon}. \tag{9}$$

High values indicate stable attribution under benign perturbation. Low values indicate explanation fragility. This metric becomes more meaningful when conditioned on slice, because one often finds that explanation instability concentrates in the same regions where support is weak or where modalities conflict [20].

Temporal stability is even more important than single-window stability. Consider two adjacent windows for the same patient that differ only by a minor update or by the passage of one hour without new major evidence. If the model score changes only slightly, then the explanation should not rotate into an entirely different subset of variables or note spans. Let $o_{i,t}$ and $o_{i,t+\Delta}$ be consecutive windows with small score change. A temporal attribution continuity measure can be defined as

$$\mathrm{TAC}(i,t) = -\frac{\|A(o_{i,t})-A(o_{i,t+\Delta})\|}{\|A(o_{i,t})\|+\varepsilon}\,\mathbb{I}\Big(|f_\theta(o_{i,t})-f_\theta(o_{i,t+\Delta})|<\tau\Big). \tag{10}$$

Low continuity under small score change suggests that the explanatory basis of the model is unstable even when the output appears smooth. This is particularly problematic for systems that will be monitored longitudinally, because clinicians may reasonably expect the rationale for an elevated score to evolve coherently over time.

Attribution reliability also interacts with modality asymmetry. In multimodal ICU systems, one may observe that structured attributions remain fairly stable while note attributions fluctuate strongly with wording or section order. Or the reverse may occur in sparse-record cases where text dominates. A useful evaluation should therefore decompose attribution stability by modality and by slice [21]. Such decomposition can reveal whether the model uses notes as semantically durable context or merely as opportunistic lexical cues.

Another challenge is that different attribution methods can disagree even on the same model. Gradient-based, perturbation-based, and attention-derived explanations often tell different stories. Rather than viewing this only as a methodological inconvenience, one can exploit it diagnostically. If several attribution methods agree within one slice and disagree sharply within another, the disagreement itself signals explanation instability. Let $A^{()},\ldots,A^{(L)}$ be different attribution methods. Then method-consensus can be defined as

$$\mathrm{MC}(o_i) = -\frac{}{L(L-)}\sum_{\ell<\ell'}\frac{\|A^{(\ell)}(o_i)-A^{(\ell')}(o_i)\|}{\|A^{(\ell)}(o_i)\|+\varepsilon}. \tag{11}$$

Low consensus does not prove any one method wrong, but it identifies windows where explanation should be treated with caution.

The most useful operational interpretation of attribution stability is not philosophical reassurance. It is risk stratification of explanations themselves. A high-risk patient-time window whose explanation is unstable across time, method, or benign

Phase	Main Activity	Objective	Output Artifact
Internal auditing	Discover hidden slices	Identify brittle regimes	Slice performance map
Policy tuning	Adjust thresholds	Balance coverage and risk	Deployment thresholds
Stress testing	Apply controlled perturbations	Reveal artifact dependence	Stability diagnostics
External validation	Evaluate across sites	Test transferability	Cross-site slice comparison
Deployment monitoring	Log predictions and context	Detect drift and failures	Continuous audit records

Table 9 Stages of an audit-oriented development and validation workflow for ICU prediction systems.

perturbation is a different kind of case from one with a stable explanatory pattern [22]. The former may deserve manual review even if the score is high and the calibration is acceptable on average, because the model's rationale appears poorly grounded. This is another way in which auditability extends beyond prediction quality alone.

Ultimately, explanation should be treated as part of the model's deployment behavior, not as a decorative appendix. Once attribution reliability is quantified and decomposed by slice, it becomes possible to say not merely that the model is accurate or inaccurate, but that it is legible or illegible in specific parts of the data space. For a clinical warning system, that distinction may matter almost as much as raw discrimination.

The next section brings these ideas into contact with deployment directly by asking how slice-wise error, support, and explanation reliability should affect alert thresholds, abstention, and review allocation under finite capacity.

5. Coverage, Abstention, and Deployment-Facing Evaluation

A deterioration model enters the clinical workflow through decisions, even if those decisions are mediated by humans. Scores are translated into dashboards, rankings, escalation prompts, or filtered review lists. In all of these pathways, the crucial operational quantities are not only discrimination and calibration, but coverage and abstention. A model that abstains from strong claims in poorly supported slices may be more useful than one that produces polished risk scores everywhere [23]. Likewise, a threshold that is well chosen globally may be unacceptable if it spends too much of the review budget on slices with low support or unstable explanations.

Let $q_i \in [\varkappa,]$ denote a support or confidence score derived from the model's evidence state, and let $f_\theta(o_i)$ be the risk estimate. A selective prediction policy uses both:

$$a_i = \mathbb{I}\big(f_\theta(o_i) \geq \tau\big)\, \mathbb{I}\big(q_i \geq \xi\big), \tag{12}$$

where τ is the risk threshold and ξ a support threshold. This is already an improvement over plain thresholding because it distinguishes "not high risk" from "not sufficiently supported." The practical question is how such thresholds behave across slices. A global pair (τ, ξ) may create acceptable average alert burden while still concentrating action in slices whose support or explanation quality is weak.

A deployment-facing evaluation should therefore consider risk-coverage curves stratified by slice. Let $C(\xi)$ denote the fraction of windows with support above ξ, and let $R_s(\xi)$ denote slice-wise risk or error among covered examples. One can then inspect

$$\begin{aligned} C_s(\xi) &= \Pr\big(q_i \geq \xi \mid i \in S_s\big), \\ E_s(\xi) &= \mathbb{E}\big[\ell_i \mid q_i \geq \xi,\, i \in S_s\big]. \end{aligned} \tag{13}$$

If coverage collapses for certain clinically important slices or if error remains high despite strong nominal support, then the support estimator is not functioning uniformly [24]. Such failures are invisible in pooled risk-coverage analysis.

The same logic applies to alert budgets. Suppose a unit can meaningfully review at most K alerts in a given period. Let Q_K denote the top-K set under a deployment score u_i, which may be a function of risk, support, and informational need. Slice-aware deployment analysis should report not only how many true events are captured in Q_K, but which slices dominate Q_K, how stable membership is within those slices, and whether explanation stability deteriorates near the budget boundary. A useful queue utility can be written as

$$U(Q_K) = \sum_{i \in Q_K} \Big(\alpha_y y_i - \alpha_f \mathbb{I}(y_i = \varkappa) - \alpha_s(\,-\,q_i) - \alpha_a(\,-\,\mathrm{AS}(o_i))\Big), \tag{14}$$

where $\mathrm{AS}(o_i)$ is attribution stability. This expression makes explicit that an alert on a low-support or explanation-unstable case carries an operational cost even if the case is not formally a false positive.

Deployment thresholds can also be selected under a maximin principle across slices rather than by optimizing one aggregate metric. Let $U_s(\tau, \xi)$ be slice-wise utility under thresholds (τ, ξ). Then one may choose [25]

$$(\tau^\star, \xi^\star) = \arg\max_{\tau, \xi} \min_{s} U_s(\tau, \xi). \tag{15}$$

This is more conservative than optimizing the pooled utility, but it better protects against severe degradation in hidden or newly

discovered slices. In high-stakes settings, such conservatism may be appropriate precisely because the average case is not what strains the workflow.

Another important deployment-facing quantity is threshold hysteresis. If scores are recomputed frequently, then minor fluctuations near the decision boundary can cause repeated entry and exit from the alert set. This is already operationally undesirable. It becomes worse when the fluctuation is concentrated in a brittle slice. Slice-aware hysteresis therefore should not only compare score values over time but also account for support and explanation stability. Cases with unstable support may require a larger margin before escalation or de-escalation to avoid thrashing the workflow with marginal evidence.

All of this suggests that evaluation should move closer to decision-theoretic analysis while remaining empirically grounded. One does not need a full causal utility model for every ICU to make progress [26]. It is enough to acknowledge that a model is judged at deployment not by one scalar summary but by how well its score, confidence, and explanations jointly support an attention-allocation policy. Once that is accepted, auditability becomes inseparable from deployment quality. The final body section therefore turns from metrics to process and proposes an audit-first model development workflow.

6. An Audit-First Development Protocol

If auditability is to be taken seriously, it cannot be added only after the model is fixed. It has to influence how models are selected, tuned, and monitored. An audit-first development protocol therefore begins from the assumption that no aggregate validation score is trustworthy until the hidden failure surface has been mapped well enough to understand where the system is brittle. The goal is not to replace modeling with auditing. The goal is to make auditing part of the same optimization loop.

The first phase of such a protocol is representation-aware internal validation. Instead of training a model and evaluating it only on held-out discrimination and calibration, one trains an accompanying audit representation $g_\phi(o)$ that is suitable for slice discovery [27]. This representation may be partly learned from the model's hidden states and partly from clinically intelligible summaries such as observation density, intervention recency, or text-structure agreement. At the end of each candidate training run, one performs slice discovery using several error notions: calibration error, top-K rank failure, support weakness, and attribution instability. The resulting slice map becomes part of model selection.

The second phase is threshold and policy tuning under slice constraints. Rather than choosing deployment thresholds by global F1, Youden's index, or pooled utility alone, one evaluates candidate thresholds across the discovered slices. In practice this means computing slice-wise alert rates, slice-wise false omission, slice-wise support-conditioned calibration, and queue composition. Thresholds are then selected not simply for overall performance but for tolerable worst-slice behavior. This avoids the common situation in which a model is tuned to behave elegantly for the dominant slice and acceptably nowhere else.

The third phase is stress testing. Before any external validation, the model should be perturbed through semantics-preserving transformations of notes, timing-preserving perturbations of observation density, and alternate imputations or mask-handling strategies [28]. These perturbations are not intended to simulate every deployment environment. They are intended to discover whether the model's score surface and explanations are dependent on brittle artifacts of one preprocessing pipeline or one local documentation habit. Slice discovery should then be repeated on the perturbed outputs. If the dominant failure slices move dramatically under benign perturbations, the model is not ready for trust.

The fourth phase is externalization with explicit uncertainty about what transferred. Too often, external validation is reported as a single score drop and interpreted as a generic generalization issue. An audit-first protocol would instead carry the slice map into the new site, rediscover slices there, and compare overlap. Some slices may remain stable and simply change prevalence. Others may disappear, while new ones emerge. This is diagnostically valuable because it reveals whether the model's blind spots were specific to the training environment or whether the architecture itself has a more portable weakness [29].

The fifth phase is post-deployment logging. Once the model enters workflow, the system should record not only predictions and outcomes, but also support levels, explanation stability summaries, queue positions, and what actions followed. This transforms auditability from a retrospective research activity into an ongoing monitoring process. If a slice that was minor during development begins to dominate alerts in deployment, the modeler can see it. If attribution stability deteriorates among top-ranked cases after a note-template change, the system can surface that before raw outcomes accumulate. Without such logging, post-deployment confidence is mostly wishful thinking.

There is also a cultural implication to this protocol. It changes what counts as a complete paper or a complete model report. A strong deterioration model should not be considered fully characterized when it has an architecture diagram, a global ROC curve, and a table of threshold metrics. It should also provide some account of hidden failure structure, support behavior, and explanation stability [30]. That may initially feel demanding, but it is a natural extension of the clinical seriousness of the task. A warning system that cannot describe where it is least trustworthy is incomplete in a way more fundamental than a missing hyperparameter appendix.

The protocol outlined here is intentionally general. Different institutions will discover different slice families and use different operational costs. That is acceptable. Auditability is not a single universal score. It is a disciplined way of refusing to let aggregate success substitute for local understanding. For ICU warning systems, that discipline may be one of the few credible routes from benchmark enthusiasm to durable deployment.

7. Conclusion

The field of ICU deterioration forecasting has become very good at summarizing itself with a few numbers. There are now standard comparisons, recognizable baselines, and familiar expectations for what a strong model should report [31]. That ma-

turity has brought welcome rigor, but it has also brought a risk of mistaking concise evaluation for complete evaluation. A system may have acceptable AUROC, improved AUPRC, manageable calibration after scaling, and even persuasive example-level explanations while remaining fundamentally underaudited. Its weak points may be concentrated in a hidden patient-time regime that rarely appears in summary tables. Its high-confidence alerts may be disproportionately drawn from one documentation pattern. Its explanations may look clinically sensible in ordinary cases and become unstable exactly when support is weakest. None of those failures is visible in the average case, and yet all of them matter in practice.

What this paper has tried to establish is that auditability should be treated as part of the predictive problem rather than as a downstream interpretability add-on. The reason is simple. A warning system is not evaluated only by how often it is right. It is evaluated by whether users can understand when it is likely to be wrong, where its confidence deserves trust, and how its decision surface changes when the record departs from the training world [32]. Those questions cannot be answered by global discrimination metrics alone because the chart is not homogeneous, the error surface is not homogeneous, and the operational cost of error is not homogeneous. Averaging over all of that complexity produces neat numbers and incomplete knowledge.

The argument has therefore moved in a different direction from the usual architecture race. Instead of asking only how to make the model stronger, it has asked how to make the model more legible under stress. Hidden failure slices, explanation instability, support-conditioned risk, and budget-aware deployment behavior are not peripheral details. They are structural properties of whether the system can be governed responsibly. If one cannot identify where the model depends on sparse, stale, or contradictory evidence, then one does not really know what the score means. If one cannot characterize how the top of the queue shifts under benign perturbations of notes or timing, then one does not really know whether the ranking is clinically grounded or artifact-sensitive. If one cannot describe which slices dominate residual error after optimization, then one does not yet know whether the remaining failure is acceptable or merely hidden.

There is also a methodological moral [33]. The temptation in applied machine learning is to treat auditing as something that happens after the "real" work of modeling is finished. That sequencing makes sense only if the task is homogeneous enough that one good scalar summary can stand in for the whole decision landscape. ICU deterioration forecasting is not such a task. The heterogeneity of observation, treatment context, documentation style, and endpoint timing ensures that failures will concentrate in specific regimes. Once that is acknowledged, audit becomes part of the object being optimized. A model that is slightly weaker in pooled AUROC but substantially more stable across discovered failure slices may be the stronger clinical system. A threshold that is slightly worse in aggregate F1 but avoids catastrophic performance in a small, high-cost slice may be the more defensible one. These are not compromises. They are examples of choosing the right objective.

The argument extends beyond one source paper, one dataset, or one architecture family [34]. Any high-stakes warning system built from heterogeneous records and evaluated under strong averaging pressure will face the same dilemma. What appears as solid performance can mask a fragile explanation surface and a brittle deployment policy. The natural response is not to abandon summary metrics, but to subordinate them to a richer map of model behavior. That map should include discovered slices rather than only predeclared subgroups, confidence and support rather than risk alone, and perturbation-based explanation reliability rather than one-off attribution plots. It should be used to tune thresholds, to select models, and to decide when a system is ready to leave retrospective experimentation.

There is no reason to pretend that this makes development easier. It makes it harder, because it asks the modeler to confront the possibility that the most clinically important failures are the ones averages hide best. But that difficulty is worthwhile. A model whose errors are globally small and locally mysterious is not obviously safer than a model whose global metrics are slightly weaker but whose local failure structure is known. In settings like the ICU, where predictions redistribute scarce attention and where evidence quality changes hour by hour, local understanding may be as important as raw accuracy [35].

If the field takes that point seriously, a different style of progress becomes possible. Instead of only reporting better overall ranking, one can report that a new system reduced instability in sparse late-stay slices, or that support-aware abstention prevented a brittle documentation-dependent queue from dominating top alerts, or that discovered failure regions remained stable across external sites. Those are more demanding claims than another benchmark win, but they are also closer to what clinicians need from a warning system. A model that can say not only "these patients are risky," but also "these are the contexts in which my judgment is least trustworthy," is a fundamentally more usable object.

The central conclusion, then, is not that current ICU models are inadequate because their AUROCs are too low. It is that current evaluation culture often stops too early. It stops after the score has been summarized and before the error has been located. An audit-first approach asks the field to continue that work: to discover where performance concentrates, where explanations wobble, where support erodes, and how those facts interact with deployment policy. For critical care prediction, that is not an optional extension of good modeling. It is one of the main ways to know whether the model deserves to be there at all [36].

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] R. Collier, "Medical technology often a burden if designed without physician input.," *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 190, pp. E1091–E1092, 9 2018.

[2] H. P. Booth, A. T. Prevost, and M. Gulliford, "Validity of

smoking prevalence estimates from primary care electronic health records compared with national population survey data for england, 2007 to 2011.," *Pharmacoepidemiology and drug safety*, vol. 22, pp. 1357–1361, 10 2013.

[3] L. Baillie, S. Chadwick, R. Mann, and M. Brooke-Read, "A survey of student nurses' and midwives' experiences of learning to use electronic health record systems in practice," *Nurse education in practice*, vol. 13, pp. 437–441, 11 2012.

[4] M. Berquedich, O. Kamach, M. Masmoudi, and L. Deshayes, "An immune memory and negative selection to visualize clinical pathways from electronic health record data," *Artificial Intelligence and Data Mining in Healthcare*, vol. 9, pp. 99–118, 7 2020.

[5] G. Loukides, J. Liagouris, A. Gkoulalas-Divanis, and M. Terrovitis, "Disassociation for electronic health record privacy.," *Journal of biomedical informatics*, vol. 50, pp. 46–61, 5 2014.

[6] 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.

[7] A. T. G. Nomura, L. Pruinelli, L. N. M. Barreto, M. dos Santos Graeff, E. A. Swanson, T. Silveira, and M. de Abreu Almeida, "Pain management in clinical practice research using electronic health records," *Pain management nursing : official journal of the American Society of Pain Management Nurses*, vol. 22, pp. 446–454, 3 2021.

[8] R. Malviya, A. K. Singh, S. Sundram, B. Balusamy, and S. Kadry, *Decentralizing health data with blockchain technology*, pp. 3–1–3–18. IOP Publishing, 9 2023.

[9] S. Ochylski and R. Freeman, *Informatics in Large Health Systems: Organization, Transformation and Nursing Informatics Leadership Perspectives*, pp. 129–146. Productivity Press, 3 2022.

[10] N. A. Bhavsar, A. Gao, M. Phelan, N. J. Pagidipati, and B. A. Goldstein, "Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data," *JAMA network open*, vol. 1, pp. e182716–, 9 2018.

[11] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International journal of medical informatics*, vol. 77, pp. 291–304, 10 2007.

[12] H.-G. Chang, C. Weng, C. DiDonato, D. DiCesare, J.-H. Chen, and D. Blog, "Evaluation of emergency department data quality following phin syndromic surveillance messaging guide," *Online Journal of Public Health Informatics*, vol. 5, 3 2013.

[13] N. G. Castellano, P. R. Mosaly, and L. M. Mazur, *AHFE (2) - Association Between Physicians' Burden and Performance During Interactions with Electronic Health Records (EHRs).* Springer International Publishing, 6 2019.

[14] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens, "Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association : JAMIA*, vol. 27, pp. 1921–1934, 10 2020.

[15] R. Gold, A. Bunce, S. Cowburn, K. Dambrun, M. Dearing, M. Middendorf, N. Mossman, C. Hollombe, P. Mahr, G. Melgar, J. V. Davis, L. M. Gottlieb, and E. Cottrell, "Adoption of social determinants of health ehr tools by community health centers," *Annals of family medicine*, vol. 16, pp. 399–407, 9 2018.

[16] B. J. Sonn and A. A. Monte, "4521 collecting, interpreting and utilizing retrospective clinical data from data warehouses," *Journal of Clinical and Translational Science*, vol. 4, pp. 46–47, 7 2020.

[17] T. Magalhães, R. J. Dinis-Oliveira, and T. Taveira-Gomes, "Digital health and big data analytics: Implications of real-world evidence for clinicians and policymakers.," *International journal of environmental research and public health*, vol. 19, pp. 8364–8364, 7 2022.

[18] K. E. Lind, M. Z. Raban, L. Brett, M. Jorgensen, A. Georgiou, and J. I. Westbrook, "Measuring the prevalence of 60 health conditions in older australians in residential aged care with electronic health records: a retrospective dynamic cohort study," *Population health metrics*, vol. 18, pp. 1–9, 10 2020.

[19] M. J. Patton and V. X. Liu, "Predictive modeling using artificial intelligence and machine learning algorithms on electronic health record data: Advantages and challenges.," *Critical care clinics*, vol. 39, pp. 647–673, 4 2023.

[20] Y. Zou, A. Pesaranghader, A. Verma, D. Buckeridge, and Y. Li, "Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model," 6 2022.

[21] K. E. Lind, M. Z. Raban, L. Brett, M. L. Jorgensen, A. Georgiou, and J. I. Westbrook, "Measuring the prevalence of 60 health conditions in older australians in residential aged care with electronic health records: a retrospective dynamic cohort study," 1 2020.

[22] B. K. Iriye, K. S. Gibson, T. Lee, D. McLean, A. M. Stuebe, P. S. Bernstein, M. A. Lutgendorf, and D. C. Lagrew, "How medical societies can collaborate with vendors to improve ehr utilization," *NEJM Catalyst*, vol. 3, 11 2022.

[23] 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.

[24] L. Abdelgalil and M. Mejri, "Healthblock: A framework for a collaborative sharing of electronic health records based on blockchain," *Future Internet*, vol. 15, pp. 87–87, 2 2023.

[25] R. Daraghmeh and R. Brown, "Icit - a big data maturity model for electronic health records in hospitals," in *2021 International Conference on Information Technology (ICIT)*, pp. 826–833, IEEE, 7 2021.

[26] M. M. Paul, C. M. Greene, R. Newton-Dame, L. E. Thorpe, S. E. Perlman, K. H. McVeigh, and M. N. Gourevitch, "The state of population health surveillance using electronic health records: A narrative review," *Population health management*, vol. 18, pp. 209–216, 1 2015.

[27] J. L. Fernández-Alemán, I. C. Señor, P. Ángel Oliver Lozoya, and A. Toval, "Security and privacy in electronic health records: A systematic literature review," *Journal of biomedical informatics*, vol. 46, pp. 541–562, 1 2013.

[28] M. Plantier, N. Havet, T. Durand, N. Caquot, C. Amaz, P. Biron, I. Philip, and L. Perrier, "Does adoption of electronic health records improve the quality of care management in france? results from the french e-si (prepssips) study," *International journal of medical informatics*, vol. 102, pp. 156–165, 4 2017.

[29] N. M. Kollapally, Y. Chen, J. Xu, and J. Geller, "An ontology for the social determinants of health domain," 11 2022.

[30] M. Gottumukkala, "Design, development, and evaluation of an automated solution for electronic information exchange between acute and long-term postacute care facilities: Design science research.," *JMIR formative research*, vol. 7, pp. e43758–e43758, 2 2023.

[31] J. Chen, A. Jagannatha, S. J. Fodeh, and H. Yu, "Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: Adapted distant supervision approach," *JMIR medical informatics*, vol. 5, pp. e42–, 10 2017.

[32] R. Parkavi, M. R. J. Iswarya, G. Kirithika, M. Madhumitha, and O. Varsha, *Data Breach in the Healthcare System*, pp. 418–434. IGI Global, 9 2023.

[33] E. Asgari, J. Kaur, G. Nuredini, J. Balloch, A. M. Taylor, N. Sebire, R. Robinson, C. Peters, S. Sridharan, and D. Pimenta, "Impact of electronic health record use on cognitive load and burnout among clinicians: Narrative review (preprint)," 12 2023.

[34] A. Malhan, I. Manuj, L. Pelton, and R. Pavur, "Electronic health records using a resource advantage theory perspective: an interdisciplinary literature review," *Records Management Journal*, vol. 32, pp. 126–150, 5 2022.

[35] 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.

[36] S. Cossin, L. Lebrun, N. Aymeric, F. Mougin, M. Lambert, G. Diallo, F. Thiessard, and V. Jouhet, "Medinfo - smartcrf: A prototype to visualize, search and annotate an electronic health record from an i2b2 clinical data warehouse.," *Studies in health technology and informatics*, vol. 264, pp. 1445–1446, 8 2019.