Peer reviewed Journals
seron press

RESEARCH

Asymmetric Evidence and Belief Revision in ICU Deterioration Forecasting

Rafael Mendes[1] and Thiago Ribeiro[2]

[1]Departamento de Computação Aplicada, Universidade Federal do Pampa, Avenida Tiarajú 810, Bairro Ibirapuitã, Alegrete, Brazil
[2]Departamento de Sistemas de Informação, Universidade Estadual do Norte do Paraná, Rodovia PR-160 Km 0, Jardim Panorama, Cornélio Procópio, Brazil

Abstract

Intensive care forecasting is frequently described as the problem of predicting near-future clinical deterioration from recent electronic health record data. That description is correct at a high level, yet it suppresses the computational asymmetry that makes the task difficult. The structured record and the documentary record do not play equivalent roles. Vital signs and laboratory values form a sparse and irregular trace of monitored physiology, while notes, reports, and bedside assessments provide compressed interpretations whose semantic richness exceeds their temporal precision. A practical deterioration model must therefore do more than combine modalities. It must maintain a revisable internal belief about patient state while deciding how strongly numeric trajectories and narrative context should influence that belief at each moment. This paper built around representational asymmetry, belief revision, and support-aware priority formation. The proposed view treats structured signals as primary carriers of short-horizon physiologic dynamics and clinical text as a higher-level contextual operator that sharpens, softens, or redirects latent hypotheses rather than replacing the temporal stream outright. A dual-timescale memory model is introduced to separate rapidly changing monitored evidence from slowly decaying contextual evidence, followed by a belief-update rule that explicitly represents contradiction, support strength, and residual uncertainty. The discussion then turns to training objectives that prevent modality collapse, maintain a useful upper tail under severe class imbalance, and link calibrated prediction to constrained review capacity. The result is a framework in which forecasting, confidence, and triage are understood as coupled problems of sequential representation rather than as isolated steps in a standard classification pipeline.

1. Introduction

The intensive care unit is often introduced as a natural habitat for machine learning because it seems to offer everything the field usually wants [1]. Data are abundant, measurements are repeated, outcomes matter, and the operational case for early recognition of decline is obvious. Yet once one moves from that general picture to the actual mechanics of ICU prediction, the task becomes less straightforward. The record available to a model is not a single coherent stream of clinical truth. It is an uneven mixture of monitored values, laboratory updates, intervention traces, and textual compressions of human judgment. Some channels arrive frequently but say little beyond local physiology. Others arrive rarely but carry rich interpretations of trajectory, uncertainty, or concern. Some entries mark direct observation. Others mark the consequences of what clinicians have already inferred. It follows that the central problem is not simply one of multimodal fusion. It is one of deciding what sort of evidence each modality contributes to an internal estimate of the patient's state [2].

This representational asymmetry is easy to underestimate because the standard benchmark setup hides it [3]. A model receives a time window, a label indicating whether an adverse event occurred soon after, and an optimization target. Once the data have been aligned and padded, the chart appears to be a generic tensor plus perhaps a note embedding. But the original record is not generic in that way. A blood pressure value, a creatinine result, a nursing assessment, a radiology report, and the absence of a fresh lactate do not have the same semantic status. The first two are directly interpretable physiologic quantities, albeit with varying freshness. The note is a delayed natural-language interpretation of a recent trajectory. The report is a bounded document tied to an imaging event and readout process. The missing lactate is neither an observed value nor pure noise; it is a trace of what the care system chose not to do at that time. Once these distinctions are noticed, the natural question becomes not merely how to fuse all channels, but how to assign

them distinct roles inside a belief-revision process [4].

The distinction between structured physiologic evidence and textual clinical evidence is particularly important. Notes can contain impressions, suspected mechanisms, concern about trajectory, anticipated complications, and contextual information unavailable in numeric fields [5]. Yet they are poor substitutes for direct short-interval monitoring. The asymmetry reported by Sadanandan (2025) [6], in which the multimodal configuration benefits from notes while the text-only setup remains near chance, is especially suggestive here: it implies that documentation helps most when treated as an interpretive modifier of ongoing physiologic estimation rather than as an independent surrogate for the short-horizon state itself. That observation motivates the central argument of this paper. Clinical documentation should not be asked to play the same representational role as structured temporal measurements. It should instead be used to revise, contextualize, and sometimes challenge the latent state inferred from monitored variables.

This perspective changes how one thinks about the architecture. If text and structure are not symmetric evidence sources, then naïvely placing them side by side in a fusion layer risks either underusing text or letting it dominate in ways that are not temporally justified. The model needs a mechanism for dual-timescale reasoning. Structured measurements need to support rapid updates because they are closer to the moment-to-moment state of the patient. Narrative evidence needs a slower memory because its semantic content often persists beyond the minute in which it was filed, but its relevance should still decay or be revised when subsequent monitored evidence contradicts it. This implies a representation with at least two characteristic time constants and an explicit notion of support [7]. A statement in text should not carry the same influence one hour later if fresh measurements have moved the physiologic state elsewhere. A stable contextual concern, however, should not disappear merely because no new note has been filed.

A second consequence concerns uncertainty. Many predictive systems output a probability and perhaps a confidence estimate, but in ICU surveillance the uncertainty is rarely one-dimensional. Some uncertainty comes from ambiguous patient state. Some comes from sparse or stale monitoring. Some comes from conflict between what the monitored data suggest and what the documentation implies. A robust belief-state model should track these distinctions because they determine whether a patient should be escalated immediately, reviewed manually, or monitored for additional evidence. In other words, the operational meaning of uncertainty is intertwined with the representational asymmetry of the inputs.

The queue into which predictions are consumed makes this even more important. In an actual unit, scores do not exist in isolation [8]. They rank patients relative to one another, compete for review slots, and must remain coherent enough that human teams can trust the ordering. A model that is globally discriminative but has a flat or unstable high-risk tail may be worse than a slightly less discriminative model whose upper ranks are sharper and whose uncertainty is more honest. The top of the ranked list is where the asymmetry between structured and textual evidence matters most, because that is where contextual information can either rescue a subtly deteriorating patient from being overlooked or inappropriately push a low-support case into an expensive review slot.

This paper develops a framework around these observations. The first step is to formalize why ICU deterioration forecasting should be understood as belief revision rather than static classification. The second is to build a representation that distinguishes fast physiologic memory from slower contextual memory, then lets them interact through explicit support and contradiction operators. The third is to derive learning objectives that prevent representational collapse, particularly the tendency of one modality to absorb the role of the other or to exploit observation artifacts. The fourth is to connect the resulting belief state to ranked triage under limited review capacity, where support-aware calibration matters as much as discrimination. The final step is to examine how such a system behaves once deployed in a closed loop where its outputs can alter future observation and thereby the evidence available for future updates.

Its primary concern is the design of latent state, memory, update operators, loss geometry, and deployment logic under asymmetrical evidence streams [9]. The underlying clinical target remains important, but the main contribution is conceptual and systems-oriented: ICU deterioration forecasting should be treated as a problem of maintaining, revising, and acting on structured belief under unequal evidence roles, not as a generic classification task with one more modality added.

2. The Mismatch Between Recorded Data and Clinical State

A common hidden assumption in predictive modeling is that the input space, although noisy, stands in a reasonably direct relation to the underlying state of interest. In the ICU that assumption weakens substantially. The relation between what is true of the patient and what is present in the chart is filtered by a clinical process that is selective, delayed, and intervention sensitive. The chart is therefore not simply a state vector with missing values. It is a representation created by repeated acts of observation and interpretation. A forecasting model that does not distinguish these acts from the patient state itself risks confound-

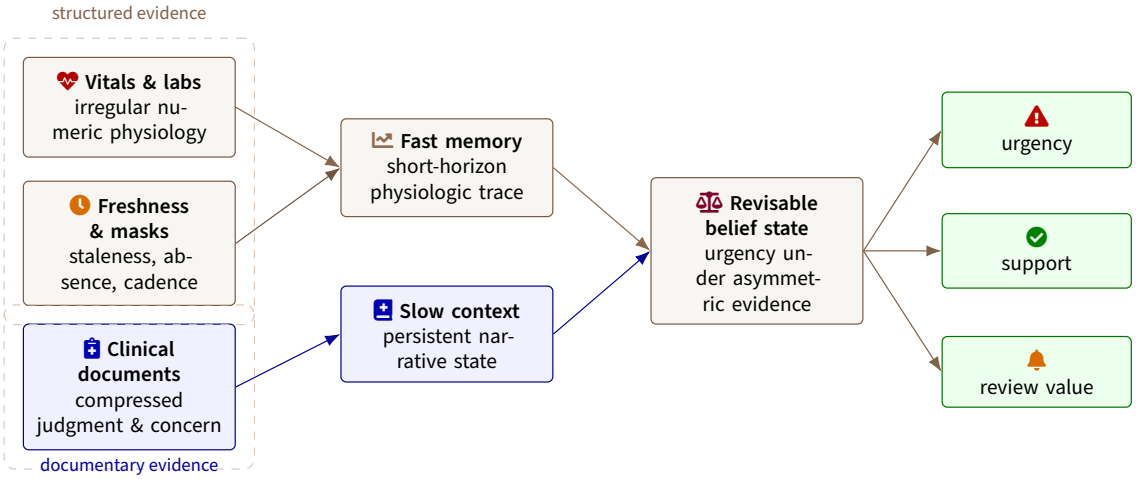


Figure 1. Asymmetric evidence roles in ICU deterioration forecasting. Structured signals drive a rapidly changing physiologic estimate, while clinical text contributes a slower contextual memory that revises rather than replaces short-horizon state estimation. The resulting belief state emits urgency, support, and review value as distinct operational quantities.

Table 1. Key components of the ICU deterioration forecasting framework

Component	Primary Role	Evidence Source	Temporal Behavior
Latent patient state $x(t)$	Hidden clinical condition	Combined signals	Continuous evolution
Observation mechanism $m(t)$	Determines measurements	Workflow + clinical context	Policy dependent
Observed record $o(t)$	Data available to model	Chart entries	Irregular updates
Intervention context $u(t)$	Treatment influence	Orders and actions	Event-driven changes

ing the thing to be inferred with the process by which it became visible.

Let $x_i(t)$ denote the latent physiologic-organizational state of patient i at continuous time t. This latent state includes not only hidden clinical severity, but also any contextual variables necessary for interpreting monitored measurements under treatment. Let $u_i(t)$ denote intervention context, and let $o_i(t)$ denote the observable record available at time t. A useful starting point is to interpose an observation mechanism $m_i(t)$, representing the current measurement and documentation policy [10]. Then

$$\begin{aligned} x_i(t+\Delta t) &= f_x\big(x_i(t), u_i(t)\big) + \epsilon_i(t) \\ m_i(t) &= f_m\big(x_i(t), u_i(t), \omega_i(t)\big) + \nu_i(t) \\ o_i(t) &= f_o\big(x_i(t), m_i(t)\big) + \eta_i(t), \end{aligned} \tag{1}$$

where $\omega_i(t)$ represents workflow or service-level factors. This decomposition is not intended as a complete causal model of the ICU. Its purpose is to state plainly that the observed record depends on a policy-like mechanism whose state is partly determined by the patient and partly by the care environment.

Once this is written down, missingness becomes more interesting. Suppose a key variable is not measured. In many domains that absence is best handled by imputation and a mask. In the ICU, the absence may be clinically benign, worrisome, or merely routine depending on context. The model should therefore not interpret absence in a uniform way. It should ask whether the current observation policy would have been expected to measure the variable if the latent state were severe. This is one reason an auxiliary observation model is helpful: it gives the system a notion of expected inspection behavior. Deviations from that expectation can then inform uncertainty or review value [11].

To see why this matters, imagine two patients with similarly sparse lab profiles. One is a stable postoperative patient on a pathway where frequent laboratory repetition is not needed. The other is a late-course unstable patient whose care team has focused on vasopressor titration and bedside assessments, leaving some labs unrefreshed. The raw mask pattern might look similar, but its interpretation should differ. In the first case, sparse observation can be consistent with low urgency and adequate support. In the second, sparse observation may coexist with high urgency and high informational deficit. A model that sees only the mask without modeling the observation regime is likely to blur these situations.

Notes produce a related but distinct mismatch. They encode observations, interpretations, and concerns that

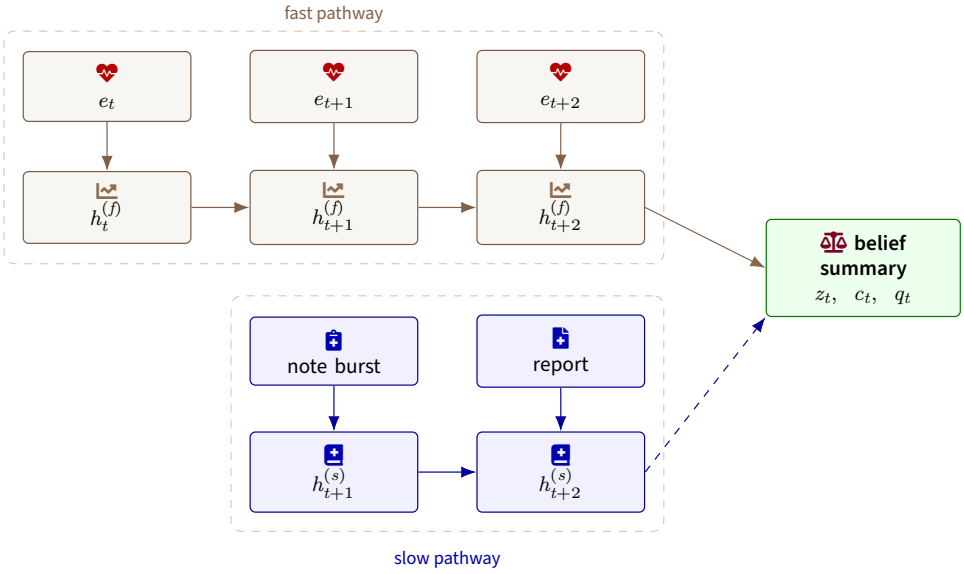


Figure 2. Dual-timescale memory architecture. Structured tokens update the fast state at each step, whereas documents update the slow contextual state only when semantically meaningful narrative evidence arrives. Belief quality depends on their agreement, their persistence, and the contradiction that remains after both streams are projected into a common latent space.

Table 2. Contrasting properties of structured signals and clinical documentation

Modality	Information Type	Strengths	Limitations
Vital signs	Physiologic measurements	Frequent updates	Limited semantic depth
Laboratory values	Quantitative biomarkers	Diagnostic specificity	Sparse timing
Clinical notes	Narrative interpretation	Rich contextual insight	Delayed timestamps
Radiology reports	Structured narrative findings	High diagnostic detail	Event-based frequency

may remain semantically relevant even when they are not temporally granular. The recorded filing time of a note is not the same as the support interval over which its content applies. A radiology report may describe a finding that remains relevant for hours or days, but a statement of bedside concern may decay more quickly if subsequent measurements contradict it [12]. This means that the mapping from note to belief update depends on both content and the evolving structured state. It is not enough to compute one embedding per note and reuse it as though text were a static covariate.

Another mismatch arises from intervention entanglement. A patient's observed hemodynamics under escalating vasopressors say something different from similar hemodynamics without support. At the same time, the intervention itself is evidence that clinicians perceived a certain urgency. Therefore the model should not treat intervention records merely as ordinary predictors or merely as targets to be excluded. They are part of the state transition and part of the support structure. A deterioration system that ignores intervention entanglement risks overreading treated stability or underreading the severity implied by the treatment burden.

This leads to a more careful definition of what a useful representation should accomplish. It should summarize the latent urgency state, preserve enough information about how that state was observed, and carry forward enough contextual evidence that support and contradiction can be assessed over time. Standard feature aggregation excels mainly at the first of these [13]. Belief-state design is motivated by the need to do all three simultaneously.

3. Dual-Timescale Memory for Asymmetric Modalities

If structured physiologic measurements and clinical documents play different epistemic roles, then their memories should not be forced into a single temporal mechanism. The physiologic stream carries local, rapidly updating evidence about short-horizon change. The document stream carries slower, semantically richer contextual evidence that is useful precisely because it compresses interpretations and concerns over a longer

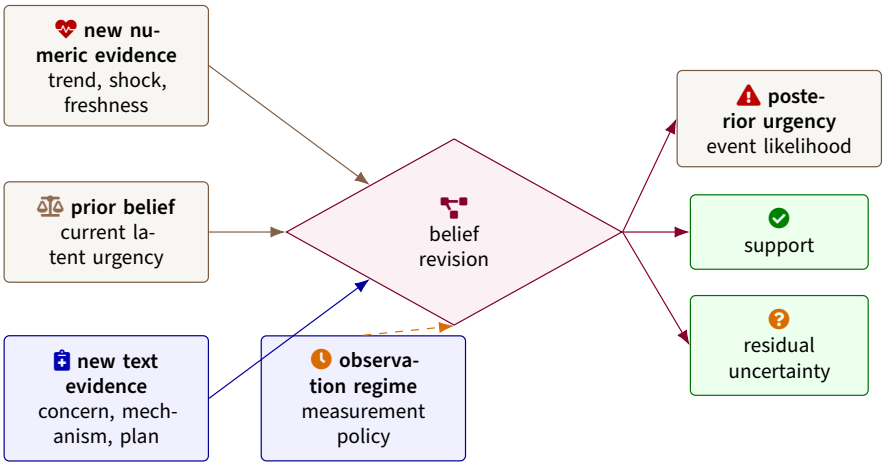


Figure 3. Support-aware belief revision. A prior state is updated by fresh physiologic evidence and slower documentary evidence, while the observation regime moderates how much trust should be placed in the update. The output separates posterior urgency from support and residual uncertainty rather than collapsing all three into one scalar score.

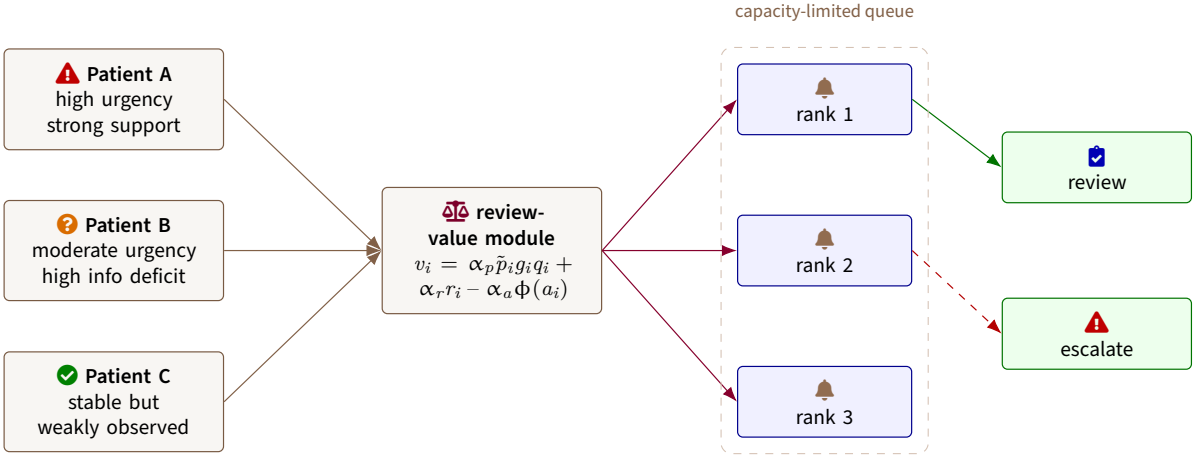


Figure 4. From belief to review value. Patient-level urgency, support, informational need, and current attention saturation are transformed into a ranked queue under limited review capacity. This representation distinguishes cases that should be reviewed for clarification from those that already warrant direct escalation.

span. A suitable model therefore needs at least two memories with different persistence and update behavior. One can think of them as a fast state and a slow context, though each may itself be high dimensional.

Let $h_i^{(f)}(t)$ denote a fast memory driven primarily by structured measurements and their freshness patterns, and let $h_i^{(s)}(t)$ denote a slower memory driven primarily by textual context, intervention narrative, and more persistent latent interpretations. The fast memory update can be written as

$$h_i^{(f)}(t+\Delta t) = (1-\kappa_i^{(f)}(t))\, h_i^{(f)}(t) + \kappa_i^{(f)}(t)\, \phi_f\big(e_i(t+\Delta t), u_i(t+\Delta t)\big), \tag{2}$$

where $e_i(t+\Delta t)$ is the structured token built from values, validity masks, and elapsed times. The gate $\kappa_i^{(f)}(t)$ should be high when genuinely informative new measurements arrive and lower when the interval contains little more than carry-forward state.

The urgency state can then be obtained from both memories:

$$z_i(t) = W_f h_i^{(f)}(t) + W_s h_i^{(s)}(t) + b_z. \tag{3}$$

However, this linear combination is too crude if left alone. What matters is not only the sum of fast and slow evidence, but whether they are coherent. Therefore a contradiction vector should be maintained:

$$c_i(t) = \big|A_f h_i^{(f)}(t) - A_s h_i^{(s)}(t)\big|. \tag{4}$$

This vector measures how different the fast and slow memories are when mapped into a common latent subspace. High contradiction does not always mean the model should lower urgency. Sometimes it should increase review value because the patient looks unstable under one view of the record and unsupported under

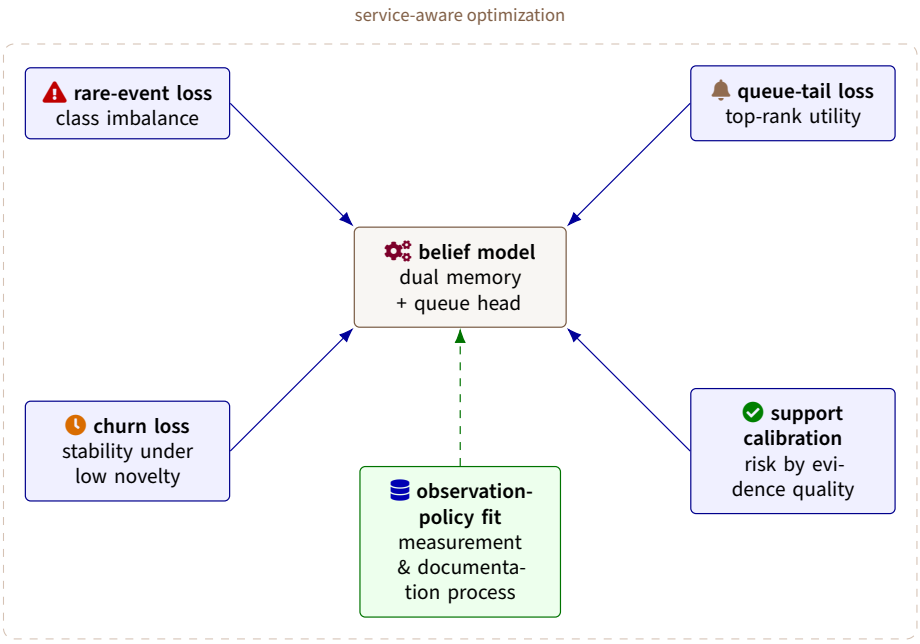


Figure 5. Training geometry under rare events. The model is optimized not only for event discrimination but also for queue concentration, support-aware calibration, low-churn prioritization, and explicit fit to the observation process that generated the chart. This keeps the top of the ranking operationally useful instead of statistically respectable but clinically flat.

Table 3. Fast and slow memory representations for multimodal evidence

Memory Type	Main Inputs	Update Trigger	Purpose
Fast memory $h^{(f)}$	Vital signs, labs	New measurements	Capture short-term dynamics
Slow memory $h^{(s)}$	Notes, contextual events	Documentation arrival	Maintain semantic context
Contradiction vector $c(t)$	Fast vs slow states	Representation divergence	Detect evidence conflict
Urgency state $z(t)$	Combined memories	Continuous inference	Estimate deterioration risk

another. What it should nearly always do, however, is affect support $q_i(t)$.

One attractive property of the dual-timescale design is that it naturally solves a problem that affects many note-based ICU predictors: temporal over-persistence of semantic embeddings. If the same note representation is allowed to influence every subsequent prediction equally, then documentation acts like a static baseline rather than as evolving context [14]. In the slow-memory framework, a note is not stored as a fixed additive vector. It alters a contextual state whose future influence depends on both its own content and subsequent fast-memory evidence. A note about a worrying trajectory can continue to matter, but its effect can be attenuated if monitored physiology later contradicts it.

The asymmetry between fast and slow memory can also be used to shape uncertainty. Diffuse fast memory with little fresh evidence should widen short-horizon uncertainty. Strong slow memory with weak fast support should produce context-heavy but cautious predictions. Agreement between the two should increase support. Let

$$q_i(t) = \sigma\Big(w_q^\top [\zeta_i^{(f)}(t), \zeta_i^{(s)}(t), c_i(t), \chi_i(t)] + b_q\Big), \quad (5)$$

where $\zeta_i^{(f)}(t)$ and $\zeta_i^{(s)}(t)$ summarize fast and slow evidence quality and $\chi_i(t)$ measures support-set familiarity. This makes support explicitly responsive to the interaction between timescales.

A further advantage of dual-timescale memory is computational. Fast updates can be performed on each chart event or hourly window, while slow updates need only occur when new documentation or meaningful semantic evidence arrives. In a streaming deployment this reduces cost and aligns processing effort with the actual semantics of the data [15]. More importantly, it keeps the model's internal dynamics interpretable: structured evidence should have the capacity to move urgency quickly, whereas text should usually move it through contextual revision rather than outright re-

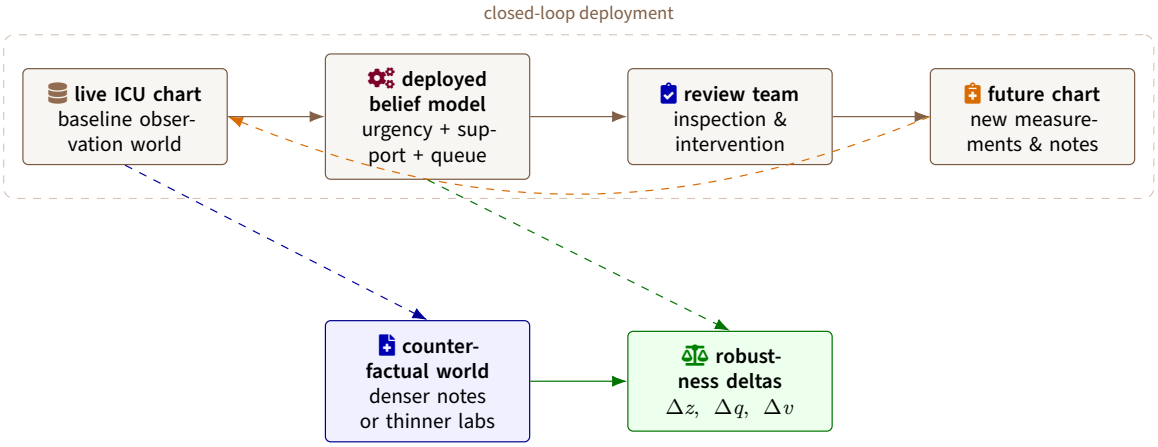


Figure 6. Closed-loop deployment and counterfactual observation worlds. Once deployed, the model changes the future chart by affecting review, intervention, and documentation patterns. Robustness should therefore be tested by perturbing observation policy itself and measuring how urgency, support, and review value move under altered documentation density or measurement cadence.

Table 4. Evidence signals influencing belief updates

Signal Type	Example	Model Interpretation	Effect on Belief
Fresh measurement	New blood pressure	High temporal relevance	Rapid urgency update
Stale observation	Old laboratory value	Reduced evidence weight	Increased uncertainty
Narrative concern	Clinician note	Contextual interpretation	Modify slow memory
Measurement absence	Missing lactate test	Observation policy trace	Adjust support estimate

placement.

The broader point is that multimodality in the ICU is not only about combining features. It is about combining evidence sources with different update laws. A model that respects that difference gains a more plausible internal geometry of urgency and support. It is then in a position to ask not only what the patient's risk is, but whether another look is worth the cost. That question is formalized next.

4. From Belief to Review Value

The operational relevance of a deterioration model comes from how it allocates attention. A score that cannot guide who deserves one more review, one more bedside check, or one more escalation conversation is of limited practical use no matter how respectable its benchmark statistics appear. The belief-state formulation is useful precisely because it can support a review-value calculation that goes beyond event probability.

Let $\tilde{p}_i(t)$ denote the calibrated probability of deterioration within the horizon of interest. Let $q_i(t)$ denote support, and let $r_i(t)$ denote residual informational need. Let $a_i(t)$ represent current attention saturation, for example how actively the patient is already being managed or monitored [16]. Then define review value

$$v_i(t) = \alpha_p \tilde{p}_i(t) g_i(t) q_i(t) + \alpha_r r_i(t) - \alpha_a \phi_a\big(a_i(t)\big), \quad (6)$$

where $g_i(t)$ is a lead-time-sensitive actionability function and ϕ_a discounts situations in which marginal benefit of additional automated prioritization is low because attention is already intense. This is a deliberately hybrid quantity. It combines event probability, support, informational need, and attention saturation because each of these changes whether another unit of review is worth spending.

The active review set $\mathcal{Q}(t)$ is then

$$\mathcal{Q}(t) = \arg \max_{\mathcal{Q} \subseteq \mathcal{I}(t)} \sum_{i \in \mathcal{Q}} v_i(t)$$
$$\text{s.t.} \quad |\mathcal{Q}| \le K_t, \quad (7)$$

where K_t is current capacity. This formulation makes clear that top-of-queue placement is not a simple function of deterioration risk. A patient can have high event probability but low marginal review value if their state is already obvious and heavily managed. Another can have slightly lower event probability but higher review value if support is weak and the trajectory still appears modifiable.

This queue view also clarifies why support matters operationally. If two patients have the same risk but one is strongly supported and the other weakly supported, the strong-support case may deserve escalation while the weak-support case may deserve review or additional observation. Support should therefore

Table 5. Factors determining support and uncertainty estimates

Factor	Description	Evidence Source	Influence on Support
Data freshness	Recency of measurements	Structured signals	Higher freshness increases support
Modal agreement	Consistency across modalities	Fast and slow memory	Agreement strengthens belief
Observation density	Frequency of monitoring	Chart patterns	Sparse data lowers support
Context familiarity	Similar prior states	Support set comparison	Familiar states increase confidence

Table 6. Elements contributing to review value in patient prioritization

Variable	Meaning	Operational Effect	Typical Scenario
Risk probability $\tilde{p}$	Estimated deterioration chance	Raises priority ranking	Rapid physiologic decline
Support q	Evidence reliability	Determines escalation confidence	Consistent measurements
Information need r	Value of additional evidence	Encourages manual review	Sparse monitoring
Attention saturation a	Existing clinical focus	Reduces marginal value	Already intensively managed patient

not merely shrink a probability. It should help decide what kind of attention to spend [17]. This is difficult to encode if the model emits only a single scalar. It becomes straightforward when urgency, support, and informational need remain separate inside the belief state.

A useful extension is to model several action tiers. Let $\mathcal{A} = \{\text{none}, \text{watch}, \text{review}, \text{escalate}\}$, and define

$$V_i(a, t) = \mathbb{E}\big[B_i(a, t) - C_i(a, t) \mid b_i(t)\big]. \tag{8}$$

The action policy selects

$$a_i^\star(t) = \arg\max_{a \in \mathcal{A}} V_i(a, t) \tag{9}$$

under action-specific capacity limits. This is more realistic than a single threshold because it preserves the difference between high urgency with low support and high urgency with high support. The former may lead to review. The latter may lead to immediate escalation. A model that collapses both to one alarm threshold forces the clinical system to reconstruct distinctions that the representation should have carried.

One can also interpret $r_i(t)$ as a limited form of value-of-information. If the patient's queue position is highly sensitive to evidence that is currently stale or absent, then additional review can be worthwhile even at only moderate risk [18]. This is especially relevant in the ICU because the record is endogenous. A model that notices not only urgency but also strategic uncertainty can guide attention toward cases where the chart itself is not keeping pace with the patient. This is a more nuanced use of machine learning than simple early warning, and it may be one of the few ways to make predictions useful when event probability alone is not enough to justify intervention.

The review-value formulation therefore provides the bridge from belief representation to operational deployment. It also creates a more appropriate target for learning. If the system is judged by the usefulness of the top of the queue, then the loss must care about that queue, not only about all windows equally. The next section addresses that issue directly.

5. Learning Under Rare Events Without Flattening the Queue

A common failure in ICU prediction is that the model learns a score field that is statistically respectable and operationally flat. The positive class is rare, adjacent windows are strongly dependent, and large parts of the score distribution never matter for action. If the loss is dominated by the bulk of the data, the model can optimize average discrimination while leaving the high-value tail too crowded, too unstable, or too poorly calibrated for practical use. Service-aware learning should therefore concentrate on the part of the ranking that feeds the queue [19].

Table 7. Action tiers derived from belief-state assessment

Action Level	Trigger Pattern	Operational Response	Goal
None	Low urgency and strong support	Continue routine monitoring	Avoid unnecessary alerts
Watch	Moderate urgency signals	Increased automated tracking	Early trajectory awareness
Review	High uncertainty or moderate risk	Human chart assessment	Clarify patient state
Escalate	High risk with strong support	Immediate clinical response	Prevent deterioration

Table 8. Training objectives supporting queue-oriented learning

Loss Component	Function	Optimization Target	Benefit
Base classification loss	Predict deterioration events	Global discrimination	Detect clinical decline
Queue-focused loss	Emphasize top-ranked patients	Utility concentration	Improves triage ordering
Churn regularization	Penalize unstable rankings	Temporal consistency	Stable worklists
Support calibration	Align support with evidence quality	Reliable confidence	Trustworthy prioritization

A class-imbalance-aware base loss remains useful:

$$\mathcal{L}_{\text{base}} = -\sum_{i,t} \alpha y_i(t)(1-\hat{p}_i(t))^{\gamma} \log \hat{p}_i(t) - \sum_{i,t}(1-\alpha)(1-y_i(t))\hat{p}_i(t)^{\gamma} \log(1-\hat{p}_i(t)). \tag{10}$$

But this objective is only the background. To shape the queue, define soft queue weights over the active census:

$$w_i(t) = \frac{\exp(\beta v_i(t))}{\sum_{j\in\mathcal{I}(t)} \exp(\beta v_j(t))}, \tag{11}$$

where $v_i(t)$ is the predicted review value. If $v_i^{\star}(t)$ is a proxy target for true review utility, then the queue-focused term

$$\mathcal{L}_{\text{queue}} = -\sum_{t} \sum_{i\in\mathcal{I}(t)} w_i(t) v_i^{\star}(t) \tag{12}$$

encourages the model to pack review utility into the top of the ranked list rather than spreading moderate scores widely.

Repeated windows create another distortion. One deterioration episode can yield many similar positive windows, which tempts the model to learn broad positive plateaus rather than useful transitions in priority. A novelty measure $n_i(t)$, derived from belief-state change or newly arrived evidence, can correct for this. One can penalize queue churn when novelty is low:

$$\mathcal{L}_{\text{churn}} = \sum_{t} \sum_{i\in\mathcal{I}(t)} (1-n_i(t)) \left| w_i(t) - w_i(t-1) \right|. \tag{13}$$

This term does not force stability universally [20]. It says that worklist movement should usually be paid for with new information. That is the right inductive bias when the system's main job is prioritization under overlapping windows.

Confidence and support should also enter the learning objective. If the model is to separate high urgency with strong support from high urgency with weak support, then $q_i(t)$ must be trained toward meaningful evidence-quality targets. Let $\rho_i(t)$ be a proxy support label derived from freshness, modality agreement, and support-set familiarity. Then

$$\mathcal{L}_{\text{conf}} = \sum_{i,t} \Big(q_i(t) - \rho_i(t) \Big)^2 \tag{14}$$

gives the support head a concrete task. A calibration term stratified by support can further regularize score meaning:

$$\mathcal{L}_{\text{cal}} = \sum_{m} \omega_m \left(\frac{1}{|B_m|} \sum_{(i,t)\in B_m} \hat{p}_i(t) - \frac{1}{|B_m|} \sum_{(i,t)\in B_m} y_i(t) \right)^2, \tag{15}$$

where bins B_m are defined jointly over risk and support. This helps prevent the top queue from being populated by scores that are rank-useful but numerically unstable under the very evidence regimes where review is expensive.

Since observation policy is explicitly modeled, the latent state should also be encouraged to explain measurement and documentation behavior. Let $\hat{m}_i(t)$ be

Table 9. Counterfactual observation scenarios for robustness testing

Scenario	Policy Change	Expected Model Response	Evaluation Focus
Measurement densification	More frequent labs	Stable urgency, higher support	Shortcut detection
Observation thinning	Reduced monitoring	Increased uncertainty	Informational deficit handling
Note inflation	Higher documentation rate	Limited urgency change	Narrative robustness
Delayed documentation	Late note entries	Gradual context decay	Temporal sensitivity

Table 10. Operational signals monitored after deployment

Metric	Description	Monitored Quantity	Purpose
Queue urgency mean	Average risk in review set	$z_i(t)$ distribution	Detect priority drift
Support level	Confidence of queued cases	$q_i(t)$ trends	Identify evidence degradation
Information deficit	Uncertainty among top cases	$r_i(t)$ values	Evaluate review efficiency
Queue turnover	Patient changes across cycles	Rank stability	Maintain actionable lists

the model's estimate of the observation-policy features and $m_i(t)$ the realized ones. Then a policy-fit term [21]

$$\mathcal{L}_{\text{obs}} = -\sum_{i,t} \Big(m_i(t) \log \hat{m}_i(t) + (1 - m_i(t)) \log(1 - \hat{m}_i(t)) \Big) \tag{16}$$

or an appropriate vector analogue constrains the model to represent not only outcomes but also the care process that produced the record. This is important because many queue failures come from hidden dependence on observation patterns.

The full objective is therefore service facing:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_q \mathcal{L}_{\text{queue}} + \lambda_t \mathcal{L}_{\text{churn}} \\ + \lambda_f \mathcal{L}_{\text{conf}} + \lambda_c \mathcal{L}_{\text{cal}} + \lambda_o \mathcal{L}_{\text{obs}} + \lambda_w |0\Theta|_2^2. \end{aligned} \tag{17}$$

Model selection should then favor the system that produces the most useful queue under realistic capacity, turnover, and support calibration, not simply the one with the best global discrimination. This is not a concession to deployment pragmatism. It is the mathematically consistent thing to do once the deployed object has been correctly specified.

6. Counterfactual Observation Worlds

If observation policy is part of the signal, then robustness must be tested by changing observation policy. Standard robustness analyses often perturb values or mask entries randomly. That is useful for generic stress testing, but it misses the more interesting question: what happens when the world changes the way the ICU actually changes? What happens if note density increases because a new documentation template is introduced, if a laboratory is ordered more routinely under a new protocol, or if the same patient trajectory is inspected more or less intensely than in development data? A belief-state model should be judged by whether its queue, support, and review-value behavior remain sensible under such counterfactual observation worlds.

Let π denote a modified observation policy, and let $\mathcal{O}_i^{(\pi)}(t)$ be the resulting record under that policy. One can then compare urgency, support, and review value under baseline and modified policies:

$$\begin{aligned} \Delta z_i^{(\pi)}(t) &= |0 z_i^{(\pi)}(t) - z_i(t)|_2, \\ \Delta q_i^{(\pi)}(t) &= q_i^{(\pi)}(t) - q_i(t), \\ \Delta v_i^{(\pi)}(t) &= v_i^{(\pi)}(t) - v_i(t). \end{aligned} \tag{18}$$

The crucial issue is not invariance [22]. In many cases the model should respond. The issue is directional correctness. If stable patients receive more routine monitoring, urgency should not inflate merely because the mask pattern changed. If a high-risk trajectory becomes more weakly observed, support should usually drop and informational need should rise, even if urgency remains elevated. If documentation is delayed, contextual influence should weaken gracefully rather than disappearing chaotically or being treated as current when it is not.

Counterfactual worlds are especially useful for identifying observation shortcuts. If synthetic note inflation or measurement densification pushes many low-risk patients upward in the queue, then the model

has learned to treat attention traces as direct surrogates for danger rather than as context for belief. Conversely, if structured thinning causes urgency to collapse even when contextual evidence remains strong, then the model has underused the slower memory. These tests therefore diagnose representational imbalance as much as they diagnose generalization.

They also provide a route to evaluating informational-deficit estimates. If $r_i(t)$ is large, then a targeted synthetic refresh of the evidence should materially change the belief state [23]. If it does not, the model's estimate of review value may be poorly grounded. In this way, counterfactual observation worlds provide supervision not only for robustness but for the model's sense of what is worth learning next.

7. Deployment Changes the Data

Once a deterioration model is deployed into an ICU, it stops being a passive reader of the chart and starts becoming part of the mechanism that writes the future chart. A patient placed high in the queue is more likely to be reviewed, measured, documented, and perhaps treated differently. This means the model enters a closed loop with the observation process. The future data distribution is no longer simply an external property of the unit. It is partly shaped by the model's own actions and by how humans respond to them.

This matters first for evaluation. A case ranked highly that later avoids overt deterioration is not obviously a false positive. It may be a genuinely useful early signal whose downstream effects altered the course of care. Conversely, a model that appears more accurate after deployment may partly owe that improvement to the fact that top-ranked patients are now observed more densely and therefore easier to classify [24]. Both phenomena mean that one cannot interpret post-deployment performance without logging what the system did to the observation world.

It matters second for drift. Since observation policy is modeled explicitly, any shift in note density, mask patterns, or review saturation should be monitored as a first-order quantity, not buried inside downstream score change. Let

$$g_i(t) = [z_i(t), q_i(t), r_i(t)] \tag{19}$$

denote the patient-level belief summary. A queue-level drift metric can be formed by comparing the mean of $g_i(t)$ over the review set to a reference distribution:

$$D_{\text{queue}}(t) = \left| 0 \frac{1}{|\mathcal{Q}(t)|} \sum_{i \in \mathcal{Q}(t)} g_i(t) - \mu_g^{\text{ref}} \right| 0 \ . \tag{20}$$

If urgency remains similar but support systematically declines in the queue, the system may still be maintaining a full worklist while spending more of it on uncertain cases. That is exactly the kind of operational degradation ordinary discrimination metrics can miss.

It matters third for adaptation. A service-aware model should generally be adapted from the outside inward. The safest first adjustment is usually recalibration, especially support-stratified calibration. The second is re-tuning of queue thresholds or capacities [25]. Retraining the latent belief architecture should be more conservative, because abrupt changes there can alter the semantics of urgency and support in ways users cannot easily track. A bounded update rule

$$\theta_{t+1}^{\star} = \arg\min_{\theta} \mathcal{L}_{\text{recent}}(\theta)$$
$$\text{s.t.} \quad \mathbb{E}\left[|0 f_{\theta}(\mathcal{O}) - f_{\theta_t}(\mathcal{O})| 0_2 \right] \le \epsilon \tag{21}$$

protects continuity of queue meaning while still allowing the system to respond to drift.

There is also a governance implication. Since the model's worklist is built from urgency, support, and informational need, those quantities should remain visible enough for humans to understand why a case is high priority. A deployment that only displays a scalar rank wastes the representational discipline of the model. A more honest interface can indicate whether a patient is high in the queue because urgency is strongly supported, because urgency is high but support is weak, or because the expected value of review is being driven mainly by informational deficit. These distinctions help preserve bounded autonomy. The model can suggest where attention would be well spent without pretending that all top-ranked cases are equivalent.

What emerges from this closed-loop view is that a deterioration model is not just a predictor. It is a participant in the evolution of the clinical record [26]. That is precisely why belief-state design is more suitable than bare classification. A belief-state model can expose what changed in the system when its own deployment changes the environment. A static predictor generally cannot.

8. Conclusion

There is a recurring habit in ICU prediction research of treating the most difficult part of the problem as if it had already been solved by preprocessing. Once the chart has been gridded, normalized, aligned, masked, and embedded, the remaining question is supposed to be how to maximize predictive performance on a future event label. The deeper claim of this paper is that the essential difficulty lies before that stage and survives all of those transformations. The ICU record is not merely incomplete. It is intentionally incomplete in patterns shaped by judgment, workflow, intervention, and attention. A forecasting system that fails to acknowledge that fact will often still learn something useful, but it will learn through a medium whose asymmetries it has not named. The result is usually a model that looks more certain than it should, responds too weakly or too strongly to text, uses missingness as a shortcut when

it should treat it as a support signal, and gives clinicians one number where the clinical problem requires a structured state of belief.

That is why the paper has argued for belief-state revision rather than feature aggregation [27]. A feature vector can be powerful while still being semantically impoverished. It can drive a classifier quite well and yet offer no principled place to represent whether urgency is well supported, whether contradictory evidence is accumulating, whether another review would mostly confirm what is already obvious, or whether the chart itself is strategically underinformative. A belief state is not magic. It does not make these distinctions trivial, and it does not guarantee that the model will learn them correctly. What it does provide is a formal place to store them, revise them, and make them matter for action. That is already a major improvement over forcing every uncertainty and every action tradeoff into the shape of one scalar probability.

The asymmetry between structured monitoring and documentation is especially instructive here. Physiologic data, even when sparse and irregular, remain the main carrier of short-horizon state change. Documentation, by contrast, is richer in abstraction than in granularity. When the two are treated as interchangeable branches of the same fusion problem, either text becomes too weak to matter or it becomes too influential for the wrong reasons. Treating text as contextual revision over a fast physiologic state is a more natural fit for what documentation is [28]. It explains why notes can materially improve a multimodal system while still performing poorly on their own in near-term deterioration tasks. It also provides a practical route to controlling how narrative context enters the queue: not as a timeless embedding, but as a slower memory whose influence should decay, persist, or reverse depending on subsequent monitored evidence.

The queue itself is another place where the conventional framing has been too small. Clinical teams do not interact with prediction one patient at a time in a vacuum. They interact with ranked sets under scarcity. Every review slot is spent at the expense of another. Every unnecessary escalation has an opportunity cost. Every unstable top-of-list reshuffle consumes attention. Once the model is understood as a producer of queue positions rather than merely of labels, several things follow immediately. Tail concentration matters more than average discrimination. Support-aware calibration matters more than a single global reliability statistic [29]. Turnover in the top of the list matters more than score stability in the bulk of the distribution. And the right immediate target is review value, not event probability alone.

The framework proposed here is deliberately more demanding than a standard deterioration classifier. It asks the model to represent urgency, support, and informational need. It asks the learning objective to care about queue utility, not only about label separation. It asks evaluation to move beyond train-test metrics and toward counterfactual observation worlds in which documentation density or measurement cadence changes. It asks deployment to be treated as a closed loop in which the model partly changes the evidence it later receives. Those demands are not embellishments added to make the system look sophisticated. They arise from the basic fact that in an ICU the chart is a product of care, not just a readout of biology.

There is a temptation, especially in technical work, to conclude by reducing everything again to a compact summary: more context matters, better modeling matters, deployment matters. That would miss the main point [30]. The main point is that the abstraction of the task has been too narrow. If one insists on seeing ICU deterioration only as event prediction under missing data, then many of the most consequential phenomena remain hidden inside preprocessing and threshold selection. If one instead sees it as belief revision under endogenous observation and scarce review capacity, then a richer and more coherent research agenda appears. That agenda includes observation-policy modeling, dual-timescale memory, support-aware calibration, value-of-information estimation, queue-facing losses, counterfactual observation robustness, and closed-loop governance. It also includes a humbler understanding of what a good model is supposed to provide. Not certainty where certainty is impossible, and not alarms where one needs structured prioritization, but a revisable internal estimate of urgency and support that can help a clinical team spend its attention where it is most likely to matter.

A model built in that spirit will still make mistakes. It will still inherit limitations from the data, from the deployment environment, and from the impossibility of fully observing critically ill patients in real time. But those mistakes will occur inside a representation that at least acknowledges the right sources of uncertainty. That is the deeper advantage of the belief-state view. It does not solve the ICU. It stops pretending that the ICU is a simpler problem than it really is [31].

References

[1] B. K. Rai, "Patient-controlled mechanism using pseudonymization technique for ensuring the security and privacy of electronic health records," International Journal of Reliable and Quality E-Healthcare, vol. 11, pp. 1–15, 2 2022.

[2] A. K. Saiod, D. van Greunen, and A. Veldsman, Handbook of Large-Scale Distributed Computing in Smart Healthcare - Electronic Health Records: Benefits and Challenges for Data Quality. Springer International Publishing, 8 2017.

[3] N. H. Ismail, Z. Jahali, and Y. Zolkefli, "Nurses' understanding of ethical dimension of using electronic health records (ehrs)," INTERNATIONAL JOURNAL OF CARE SCHOLARS, vol. 6, pp. 59–68, 7 2023.

[4] H. Angier, S. Likumahuwa, S. C. Finnegan, T. Vakarcs, C. C. Nelson, A. Bazemore, M. Carrozza, and J. E. De-

Voe, "Using geographic information systems (gis) to identify communities in need of health insurance outreach: An ochin practice-based research network (pbrn) report," Journal of the American Board of Family Medicine : JABFM, vol. 27, pp. 804–810, 11 2014.

[5] S. Hoffman, Electronic Health Records and Medical Big Data. Cambridge University Press, 12 2016.

[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. S. Dighe, "Electronic health record optimization for artificial intelligence.," Clinics in laboratory medicine, vol. 43, pp. 17–28, 12 2022.

[8] X. Zhang, S. Li, Z. Cheng, R. Callcut, and L. Petzold, "Domain adaptation for trauma mortality prediction in ehrs with feature disparity," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1145–1152, IEEE, 12 2021.

[9] L. Choi, C. Beck, E. McNeer, H. L. Weeks, M. L. Williams, N. T. James, X. Niu, B. Abou-Khalil, K. A. Birdwell, D. M. Roden, C. M. Stein, C. A. Bejan, J. C. Denny, and S. L. V. Driest, "Development of a system for postmarketing population pharmacokinetic and pharmacodynamic studies using real-world data from electronic health records.," Clinical pharmacology and therapeutics, vol. 107, pp. 934–943, 2 2020.

[10] N. Maganti, H. Tan, L. M. Niziol, S. Amin, A. Hou, K. Singh, D. Ballouz, and M. A. Woodward, "Natural language processing to quantify microbial keratitis measurements," Ophthalmology, vol. 126, pp. 1722–1724, 6 2019.

[11] A. J. Holmgren and N. C. Apathy, "Assessing the impact of patient access to clinical notes on clinician ehr documentation.," Journal of the American Medical Informatics Association : JAMIA, vol. 29, pp. 1733–1736, 7 2022.

[12] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nature reviews. Genetics, vol. 13, pp. 395–405, 5 2012.

[13] L. A. Rodriguez, T. W. Thomas, H. Finertie, D. Wiley, W. T. Dyer, P. E. Sanchez, M. Yassin, S. Banerjee, A. Adams, and J. A. Schmittdiel, "Identifying predictors of homelessness among adults in a large integrated health system in northern california.," The Permanente journal, vol. 27, pp. 56–71, 3 2023.

[14] S. Thuraisingam, P. Chondros, M. M. Dowsey, T. Spelman, S. Garies, P. F. M. Choong, J. Gunn, and J.-A. Manski-Nankervis, "Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment.," BMC medical informatics and decision making, vol. 21, pp. 297–, 10 2021.

[15] I. Abunadi and R. L. Kumar, "Bsf-ehr: Blockchain security framework for electronic health records of patients," Sensors (Basel, Switzerland), vol. 21, pp. 2865–, 4 2021.

[16] N. S. Pagad, P. N, K. K. Almuzaini, M. Maheshwari, D. Gangodkar, P. Shukla, and M. Alhassan, "Clinical text data categorization and feature extraction using medical-fissure algorithm and neg-seq algorithm.," Computational intelligence and neuroscience, vol. 2022, pp. 5759521–16, 3 2022.

[17] Y. J. Juhn, E. Ryu, C. i. Wi, K. S. King, S. R. Brufau, C. Weng, S. Sohn, R. Sharp, and J. Halamka, "An individual-level socioeconomic measure for assessing algorithmic bias in health care settings: A case for houses index," 8 2021.

[18] G. C. Wood, X. Chu, C. Manney, W. E. Strodel, A. T. Petrick, J. Gabrielsen, J. Seiler, D. J. Carey, G. Argyropoulos, P. N. Benotti, C. D. Still, and G. S. Gerhard, "An electronic health record-enabled obesity database," BMC medical informatics and decision making, vol. 12, pp. 45–45, 5 2012.

[19] R. Kavitha, E. Kannan, and S. Kotteswaran, "Implementation of cloud based electronic health record (ehr) for indian healthcare needs," Indian Journal of Science and Technology, vol. 9, pp. 1–5, 2 2016.

[20] A. K. Barrett, "Technological appropriations as workarounds: Integrating electronic health records and adaptive structuration theory research," Information Technology & People, vol. 31, pp. 368–387, 4 2018.

[21] E. Smith, J. Lapinski, J. Lichty-Hess, and K. Pier, "Using health information technology and data to improve chronic disease outcomes in federally qualified health centers in maryland.," Preventing chronic disease, vol. 13, pp. E178–, 12 2016.

[22] S. Saigaonkar and V. Narawade, "Predicting chronic diseases using clinical notes and fine-tuned transformers," in 2022 IEEE Bombay Section Signature Conference (IBSSC), IEEE, 12 2022.

[23] M. Udoh, H. Okeke, L. Edet, and N. Osuchukwu, "Healthcare providers' knowledge and perceptions of electronic health records and perceived effect on health service delivery in tertiary health facilities in uyo, nigeria," World Journal of Public Health, 9 2023.

[24] S. V. Kiran, I. Kaur, K. Thangaraj, V. Saveetha, R. K. Grace, and N. Arulkumar, "Machine learning with data science-enabled lung cancer diagnosis and classification using computed tomography images," International Journal of Image and Graphics, vol. 23, 11 2021.

[25] R. Gibson, Personal Health Records and Health Record Banks, pp. 69–88. Productivity Press, 2 2018.

[26] 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.

[27] V. C. Kaggal, R. K. Elayavilli, S. Mehrabi, J. J. Pankratz, S. Sohn, Y. Wang, D. Li, M. M. Rastegar, S. P. Murphy, J. L. Ross, R. Chaudhry, J. D. Buntrock, and H. Liu, "Toward a learning health-care system – knowledge delivery at the point of care empowered by big data and nlp," Biomedical informatics insights, vol. 8, pp. 13–22, 6 2016.

[28] X. Nie, A. Zhang, J. Chen, Y. Qu, and S. Yu, "Time-enabled and verifiable secure search for blockchain-empowered electronic health record sharing in iot," Security and Communication Networks, vol. 2022, pp. 1–15, 12 2022.

[29] R. Adlam and B. Haskins, "A permissioned blockchain approach to electronic health record audit logs," in Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications, pp. 1–7, ACM, 9 2020.

[30] H. Hemingway, F. W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. P. Maggioni, G. J. M. W. van Thiel, M. Cronin, G. Brobert, P. Vardas, S. D. Anker, D. E. Grobbee, and S. Denaxas, "Big data from electronic health records for early and late translational cardiovascular research: challenges and potential," European heart journal, vol. 39, pp. 1481–1495, 8 2017.

[31] S. M. Kabene, R. W. Leduc, and C. J. Gibson, Data Security in Electronic Health Records, pp. 182–194. IGI Global, 8 2010.