Peer reviewed Journals

SciChronice

RESEARCH

Belief-State Design for ICU Deterioration Forecasting Under Endogenous Observation and Scarce Review Capacity

Nguyen Minh Khoa[1] and Tran Quoc Huy[2]

[1]Department of Computer Engineering, Hung Yen University of Technology and Education, Khoai Chau Campus, Dan Tien Road, Dan Tien Commune, Khoai Chau District, Hung Yen, Vietnam
[2]Faculty of Information Technology, Ha Tinh University, 447 26/3 Street, Dai Nai Ward, Ha Tinh City, Ha Tinh, Vietnam

Abstract

Critical care forecasting is often presented as a straightforward supervised learning task in which future deterioration is predicted from current electronic health record features. That framing is serviceable for benchmarking, but it hides a defining computational property of intensive care data: what becomes visible in the record is shaped by clinical attention itself. Measurements are ordered selectively, documentation density changes with concern and workflow, interventions alter both patient state and what is subsequently monitored, and the absence of information is frequently informative about the observation process rather than the patient alone. As a result, the central technical challenge is not simply prediction from incomplete data. It is inference over a partially observed dynamical system whose observation policy is endogenous to the underlying state and to clinician action. This paper develops a computer-science-centered framework that redefines ICU deterioration modeling as belief-state design under endogenous observation. The proposed view separates latent physiologic state, observation policy, intervention context, confidence, and marginal review value. A patient representation is then learned not merely to maximize event discrimination, but to support risk estimation, uncertainty formation, and prioritized review under constrained clinical capacity. The paper introduces a state-space formulation in which structured measurements, missingness patterns, and clinical notes are interpreted jointly through an observation model, a belief update mechanism, and a queue-facing utility layer. It also examines how rare-event training, temporal dependence, calibration, and support drift interact with the allocation of scarce human attention. The resulting argument is that useful ICU forecasting systems should be designed less as binary event detectors and more as engines for maintaining and acting on belief states when both the patient and the record are evolving under policy.

1. Introduction

The intensive care unit has become a canonical setting for machine learning because it offers what many domains appear to lack: repeated measurements, multimodal data, meaningful outcomes, and operational motivation for early warning [1]. It is therefore understandable that a large body of work has framed ICU deterioration prediction as a promising application of sequence modeling. Yet the apparent abundance of data obscures a more difficult issue. Intensive care records do not arise from passive sensing. They arise from active care. A clinician decides whether to order a blood gas, a nurse decides when and how to document, a radiology report appears after an image is obtained and interpreted, and a treatment change can simultaneously signal concern, alter physiology, and increase future monitoring intensity. In such a setting, the record is not merely a noisy projection of the patient. It is also a record of how the patient was inspected [2]. Any computational framework that ignores that dual role is solving an easier and less faithful problem than the one encountered in deployment.

This observation matters because it changes what should count as signal. In ordinary predictive modeling, one typically thinks of missingness as a defect in the input. One may encode it with masks, attempt to impute through temporal models, or hope that sufficiently expressive architectures will learn to cope. In critical care, however, the pattern of what is present and what is absent is often part of the mechanism of concern. Sparse measurement may reflect stability, but it can also reflect routine overnight practice, transfer timing, or a pathway in which only selected variables are watched closely. Dense measurement may reflect acute instability, but it may also arise from post-procedure protocolization or from attention already concentrated on a recognized problem. The meaning of observation therefore depends on context, and context is exactly what is often compressed away when the chart is converted into a regular matrix for prediction

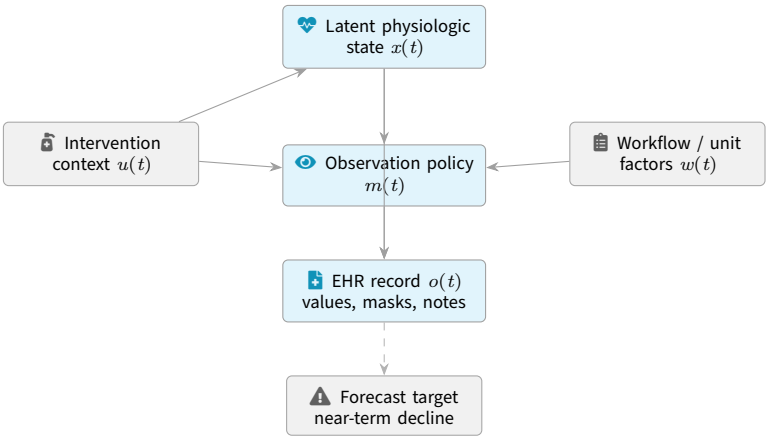

Figure 1. Endogenous observation in the ICU. The patient state does not directly produce the chart alone; interventions and workflow shape what is inspected, when it is documented, and which absences in the record are themselves informative.

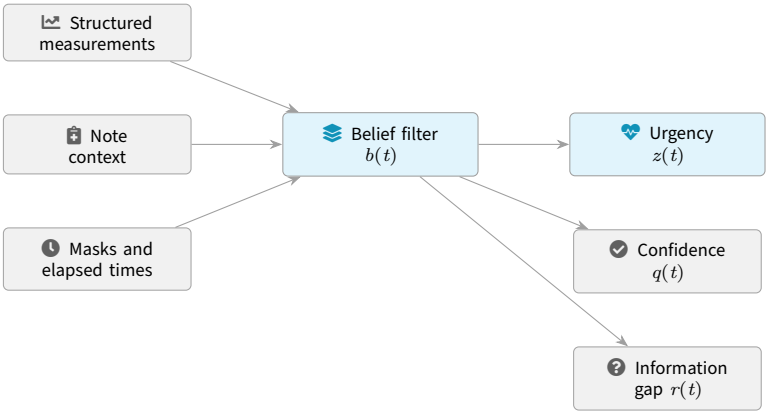

Figure 2. Belief-state decomposition. Instead of compressing all evidence into a single opaque feature vector, the representation is split into urgency, support quality, and residual informational deficit so that ranking and review can depend on more than event probability alone.

[3].

A second implication is that the clinically relevant output of a deterioration system is not exhausted by a probability of future decline. Teams do not use one scalar in a vacuum. They decide who to review now, who can wait, who is rising in urgency, and which cases remain poorly characterized enough that additional attention may itself have value. In practice, the model participates in an allocation process. A high score elevates a patient in a worklist, competes for scarce review time, and may change what is measured next. A useful forecasting system must therefore estimate not only risk but also confidence and the expected value of additional review or sensing. The right model output is closer to a structured belief state than to a naked probability.

There is a further reason to move beyond the standard framing [4]. The positive class in ICU deterioration is usually rare on an hourly grid, and the upper tail of model scores is where operational burden and benefit are concentrated. Under these conditions, average discrimination metrics can obscure the true difficulty. A system may place positive windows above negative windows often enough to achieve a respectable AUROC while still flooding the top of the list with marginal cases. Conversely, a system that is modestly weaker in global ranking may create a far more useful top queue if its scores are better calibrated, smoother over time, and more sensitive to whether current evidence is sufficient. The design objective should therefore be aligned with how the score will be consumed, not only with how it compares to a label in retrospective evaluation.

In Sadanandan (2025), the text-only model trails both the structured-only and the multimodal alternatives [5], which suggests that documentation is more valuable as contextual evidence than as an independent surrogate for near-term physiologic state. That single empirical relation is easy to overlook, yet it has deep design consequences. It suggests that notes should not be asked to replace the dynamics of monitored variables [6]. Instead they should modulate how structured evidence is interpreted, how latent hypotheses

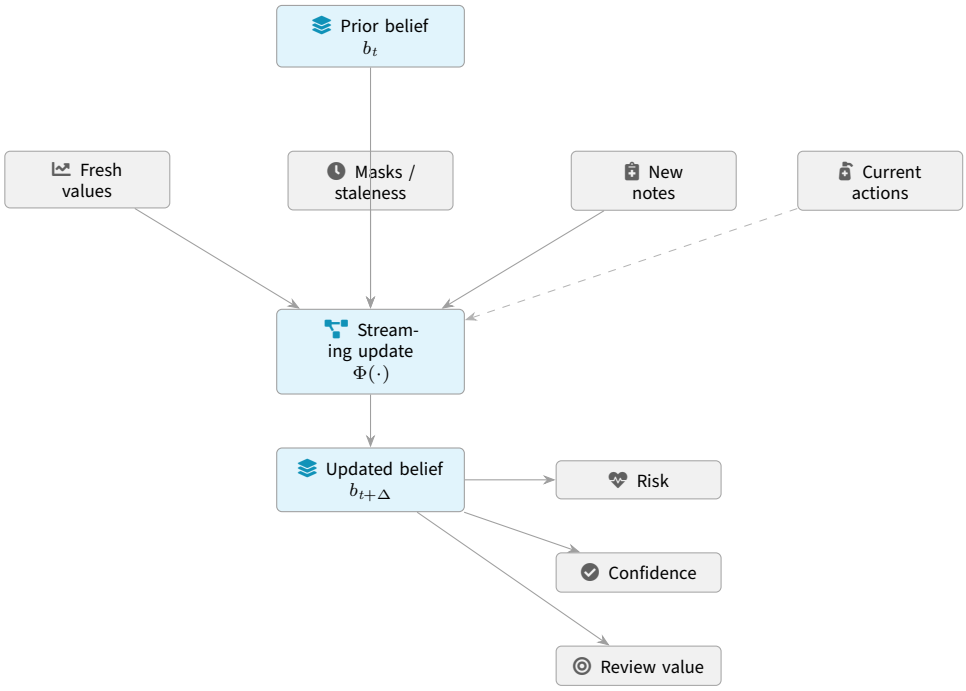

Figure 3. Streaming belief update. Each interval contributes new measurements, missingness structure, text, and intervention context to a compact filter, which refreshes the latent state and exposes downstream heads for risk, confidence, and marginal review utility.

are weighted, and how confidence changes when measured state and documented state agree or diverge. Put differently, text appears most useful when treated as an interpretive layer over belief updating rather than as another standalone predictor competing for the same role as a time series encoder.

This paper develops that argument by treating ICU deterioration forecasting as a problem of belief-state design under endogenous observation. The key claim is that the patient's latent urgency state, the policy by which evidence is gathered, and the system's estimate of what remains uncertain should all be modeled together. Once that is done, the work of the model changes. It is no longer merely learning to label windows. It is maintaining a state estimate that can drive review prioritization, selective escalation, and support-aware abstention. The following sections build that case in a different order than the usual predictive-paper progression [7]. The discussion first formalizes why observation must be considered part of the problem. It then turns to the representational consequences of treating missingness and documentation as generated phenomena. After that, it develops the link between belief states and review value, then moves to rare-event learning, counterfactual evaluation, and deployment. The intention throughout is not to overstate the novelty of any single modeling ingredient. The intention is to show that once the object of the task is named correctly, the architecture, the loss, and the decision layer all need to change with it.

2. Observation Is Part of the State

Suppose that each patient i occupies a latent clinical state $x_i(t)$ evolving in continuous time. If observations were sampled passively and uniformly, then a deterioration model could reasonably be described as learning a map from recent observations to a future outcome. Intensive care data do not satisfy that assumption [8]. Instead, observations are generated conditionally on the latent state, on intervention context, and on local clinical policy. It is therefore useful to introduce an observation-policy variable $m_i(t)$ that summarizes what the care system chose to inspect or document at time t. The model does not observe $x_i(t)$ directly. It observes a record $o_i(t)$ emitted through a process governed jointly by $x_i(t)$ and $m_i(t)$. A compact formulation is

$$\begin{aligned} x_i(t+\Delta t) &= F_x\big(x_i(t), u_i(t)\big) + \epsilon_i(t) \\ m_i(t) &= F_m\big(x_i(t), u_i(t), w_i(t)\big) \\ o_i(t) &= F_o\big(x_i(t), m_i(t)\big), \end{aligned} \tag{1}$$

where $u_i(t)$ denotes intervention context and $w_i(t)$ denotes workflow or institutional factors. This factorization is deliberately simple, but it captures the central point: the observation policy is not exogenous.

The immediate consequence is that one should not think of missing values and observation masks merely as aids to imputation [9]. They are summaries of the policy channel $m_i(t)$. For structured variables indexed by j, let $v_{i,t}^{(j)}$ be the latest available value, $a_{i,t}^{(j)} \in \{0, 1\}$

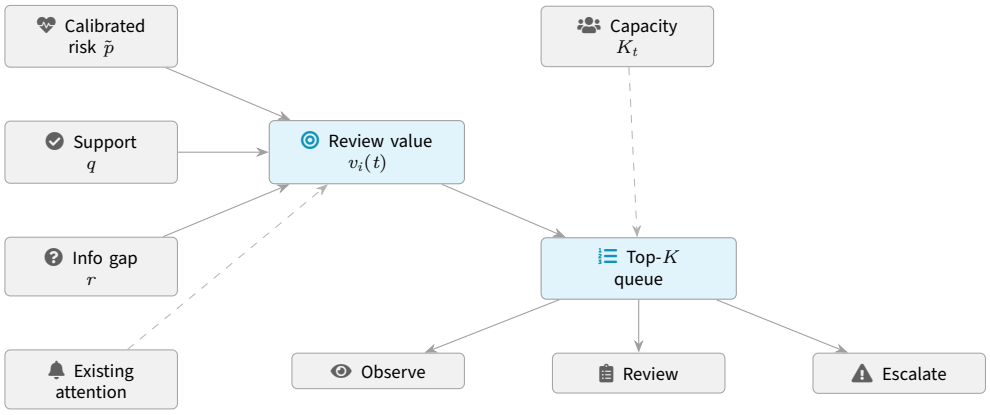

Figure 4. Queue-facing decision layer. Review priority is assembled from calibrated risk, confidence, informational deficit, and current attention saturation, then converted into a capacity-limited queue that separates simple observation from chart review and stronger escalation.

indicate whether the variable was observed during the current interval, and $\delta_{i,t}^{(j)}$ be the elapsed time since it was last observed. The usual representation

$$e_{i,t} = W_v v_{i,t} + W_a a_{i,t} + W_\delta \log(1 + \delta_{i,t}) + b \quad (2)$$

can be reinterpreted more productively. The first term says what the last known physiologic quantity was. The second and third say how the observation process has treated that quantity recently. In an endogenous observation model, those latter terms are not ancillary. They are part of what is being predicted from.

Clinical notes fit the same logic, though less directly [10]. A note is not only text. It is evidence that someone decided to summarize, assess, or justify. Its appearance time, type, and density are part of $m_i(t)$, while its semantic content provides indirect information about $x_i(t)$. This means that a note contributes along two channels at once: a semantic channel and a policy channel. If a model uses only the semantic embedding and ignores the fact that the note arrived now rather than five hours earlier, it discards part of the support structure. If it uses note frequency too strongly without controlling for context, it risks learning workflow shortcuts. The architecture should therefore allow note presence and note content to influence different aspects of the belief state.

A subtle but important issue emerges when one asks what it means for a state to be "known." In passive settings, one might say that the more measurements are available, the better the state is known [11]. In an ICU this is not always true. Many measurements can coexist with severe interpretive uncertainty if the patient is unstable and therapies are changing rapidly. Conversely, relatively sparse data can be enough to know that a patient is currently stable if those data align and the observation policy is itself informative. Therefore uncertainty is not a simple monotone function of data count. It depends on whether the current evidence is relevant, fresh, and coherent. That is why the observation-policy embedding must feed confidence formation rather than only missing-value handling.

Another reason to model $m_i(t)$ explicitly is robustness. Suppose a unit changes a standard order set so that lactate is measured more routinely [12]. The distribution of observation masks changes immediately. A model that had learned to interpret repeated lactate checks as a strong acuity signal may then overestimate severity in a way not visible from value distributions alone. If observation policy is a modeled object, such a shift can be monitored and its effects localized. If it is only implicit in the weights of an opaque predictor, the resulting failure looks like mysterious drift. By naming the observation process explicitly, one turns a hidden confounder into a trackable state variable.

A final implication concerns what additional review is supposed to accomplish. If review is valuable partly because it reduces uncertainty about $x_i(t)$, then the model needs a representation of uncertainty tied to $m_i(t)$ [13]. That is, the model must know not only that the patient is risky, but also whether risk is high because the evidence is strong or because the evidential pattern itself is suspiciously sparse or selective. Once this is acknowledged, event forecasting naturally shades into active sensing. The next section makes that move more explicit by introducing a belief state that compresses both the latent patient and the way the patient has been observed.

3. Belief Compression Instead of Feature Aggregation

Most ICU models begin by aggregating features into a fixed-dimensional representation. That representation may be produced by a recurrent network, a transformer, or a structured temporal encoder, but the conceptual operation is similar: many heterogeneous observations are compressed into a vector optimized to predict the future label. A belief-state view imposes

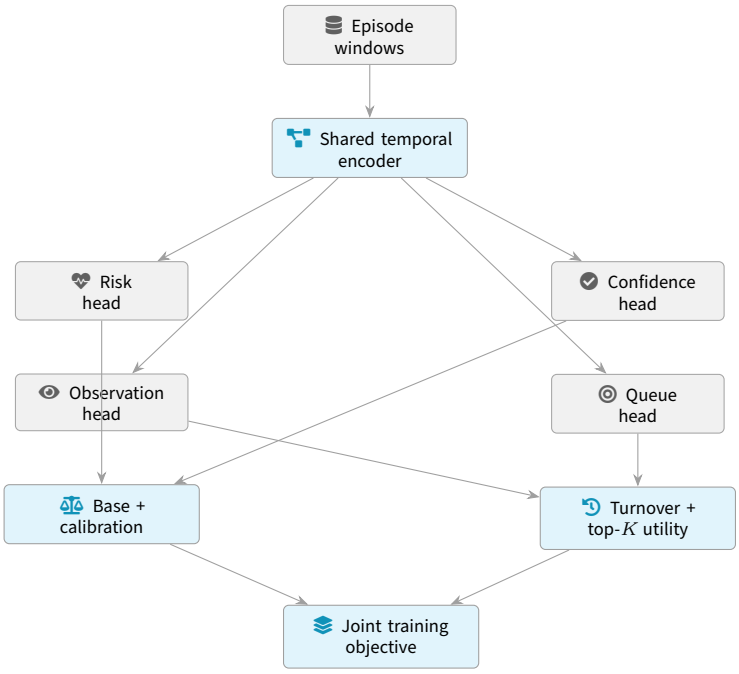

Figure 5. Training for the top of the queue. A shared encoder feeds multiple heads so the optimization can jointly shape discrimination, confidence quality, observation-policy modeling, calibration, and queue stability rather than only retrospective label prediction.

a different requirement. Compression should preserve not only predictive content but also enough structure to say how the current estimate was formed and how reliable it is. The internal representation should therefore be a belief state rather than a plain feature state [14].

Let $b_i(t)$ denote the model's belief about the relevant aspects of $x_i(t)$ given admissible evidence up to time t. The belief state need not be probabilistic in the exact Bayesian sense, but it should behave as though it summarizes a posterior over latent urgency and support. A useful decomposition is

$$b_i(t) = \big(z_i(t), q_i(t), r_i(t)\big), \tag{3}$$

where $z_i(t)$ is the latent urgency component, $q_i(t)$ the confidence or support component, and $r_i(t)$ a representation of residual uncertainty or informational deficit. The point of this split is that urgency, confidence, and informational need do not align perfectly. A single scalar risk score cannot encode all three without ambiguity.

Belief updating can then be written as a recurrent inference step rather than as one more pass through a prediction network:

$$b_i(t+\Delta t) = \Phi\big(b_i(t), o_i(t+\Delta t), m_i(t+\Delta t), u_i(t+\Delta t)\big). \tag{4}$$

Here Φ is the learned filter [15]. Structured measurements update the urgency component directly when they are fresh and relevant. Observation masks and note density update the support and informational-deficit components by indicating what has or has not been checked. Interventions alter both urgency dynamics and the interpretation of newly arriving evidence. This filtering perspective is computationally valuable because it aligns with streaming deployment. New data update the belief state incrementally instead of forcing the model to reconstruct long windows from scratch.

The belief state should also preserve modality-specific structure long enough to assess agreement. Let $z_i^{(s)}(t)$ summarize structured evidence, $z_i^{(d)}(t)$ summarize document evidence, and $z_i^{(m)}(t)$ summarize observation-policy context. A gated combination

$$z_i(t) = g_i(t) \odot z_i^{(s)}(t) + \big(1 - g_i(t)\big) \odot T_d z_i^{(d)}(t) + T_m z_i^{(m)}(t), \tag{5}$$

where $g_i(t)$ depends on freshness and support, is preferable to unconditional concatenation [16]. Such a combination allows the model to rely more heavily on text when semantic context is current and consistent, more heavily on structure when direct measurements are fresh, and more heavily on observation-policy cues when the very pattern of measurement is informative. The transforms T_d and T_m align branch dimensions without erasing branch identity entirely.

Confidence $q_i(t)$ should not be derived from score magnitude. It should be learned from belief-state prop-

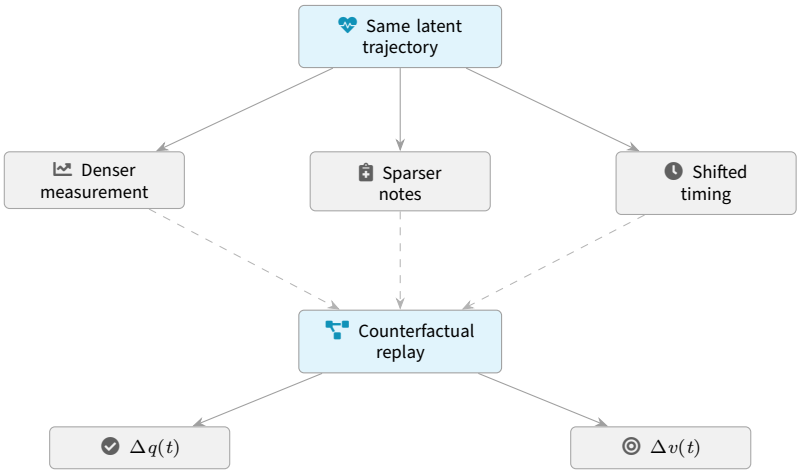

Figure 6. Counterfactual observation worlds. The same underlying patient trajectory is replayed under altered measurement density, note availability, or timing so that robustness can be judged by how confidence and review value move when the support process changes.

erties. A simple head is

$$q_i(t) = \sigma\Big(w_q^\top [\Delta_i^{(sd)}(t), \zeta_i(t), \chi_i(t)] + b_q\Big), \qquad (6)$$

where $\Delta_i^{(sd)}(t)$ is structured-document agreement, $\zeta_i(t)$ summarizes evidence freshness and density, and $\chi_i(t)$ is a support-distance feature measuring how far the current belief lies from familiar regions of training-time state space. This makes confidence a learned synthesis of agreement, freshness, and familiarity rather than an opaque byproduct of the classifier head.

The informational-deficit term $r_i(t)$ is the piece most absent from conventional ICU models. It is the model's estimate of how much useful uncertainty remains because the patient has not been observed in certain important ways [17]. One need not make this term fully causal or counterfactual to make it useful. It is enough to learn a proxy for how much the current action recommendation would change if selected evidence were refreshed. In a practical network, $r_i(t)$ can be trained against approximations to future uncertainty reduction or against simulated observation-refresh outcomes. The presence of this term is what allows the model to recommend attention not only because risk is high, but because uncertainty is consequential and still reducible.

A final benefit of belief compression is interpretive discipline. Since the model tracks urgency, confidence, and residual informational need separately, it becomes easier to explain why a patient is high priority. The answer need not be a brittle token-level attribution. It can be a structured statement: urgency is high because recent hemodynamics worsened and treatment intensified; confidence is moderate because documentation is stale and respiratory measurements are sparse; review value is high because additional information could materially change triage [18]. This is much closer to the operational use case of a surveillance engine than a bare score.

The belief-state view sets up the next step naturally. If the model carries forward not only an estimate of risk but also an estimate of what remains worth learning, then the review queue can be optimized with respect to more than outcome probability. The next section formalizes that idea by treating review allocation as a value-of-information problem under limited capacity.

4. Review Allocation as Value of Information

A review slot in the ICU is not valuable merely because it lands on a high-risk patient. It is valuable because it may change what is known, what is done, or how quickly the team responds. This is why a service-aware forecasting system should order patients by expected value of review rather than by event probability alone. The expected value of review has at least three components: predicted deterioration risk, expected actionability if that risk is real, and expected information gain if uncertainty is still consequential [19]. A patient can score highly on one component and weakly on another. The queue should reflect the combination.

Let $\tilde{p}_i(t)$ denote the calibrated event risk for patient i, $q_i(t)$ denote confidence, and $r_i(t)$ denote informational deficit. Define $g_i(t)$ as a lead-time-sensitive actionability weight, decreasing when the event is either too distant or too immediate to benefit much from one more review. Then a simple expected review value is

$$v_i(t) = \alpha_p \tilde{p}_i(t) g_i(t) q_i(t) + \alpha_r r_i(t) - \alpha_a \phi_a\big(a_i(t)\big), \qquad (7)$$

where $\phi_a(a_i(t))$ discounts cases already receiving high-intensity attention. This formula is deliberately not purely probabilistic. It recognizes that a patient with moderate event risk but large reducible uncertainty

Table 1. Core variables in the ICU state–observation formulation.

Component	Symbol	Description	Role in Model
Latent state	$x_i(t)$	Hidden physiologic condition	Governs patient dynamics
Intervention context	$u_i(t)$	Treatments and actions	Alters state evolution
Observation policy	$m_i(t)$	What clinicians choose to measure	Determines visibility of data
Observed record	$o_i(t)$	EHR measurements and notes	Input to inference system
Workflow factors	$w_i(t)$	Operational conditions	Influences observation behavior

Table 2. Representation of structured observations and missingness signals.

Feature Type	Notation	Meaning
Latest value	$v_{i,t}^{(j)}$	Most recent structured measurement
Observation indicator	$a_{i,t}^{(j)}$	Whether variable was measured now
Elapsed time	$\delta_{i,t}^{(j)}$	Time since last measurement
Embedding vector	$e_{i,t}$	Encoded observation representation
Projection weights	W_v, W_a, W_δ	Learnable embedding parameters

may deserve a slot ahead of a patient with slightly higher risk but minimal marginal benefit from more review.

The review queue is then the solution to [20]

$$\mathcal{Q}(t) = \arg\max_{\mathcal{Q}\subseteq\mathcal{I}(t)} \sum_{i\in\mathcal{Q}} v_i(t)$$
$$\text{s.t.} \quad |\mathcal{Q}| \le K_t. \tag{8}$$

Because K_t is limited, queue membership is inherently comparative. Every admitted patient displaces another. This is why the value term $\phi_a(a_i(t))$ matters. Without it, the queue can become dominated by already saturated cases whose risk is extreme but whose marginal benefit from automated reprioritization is low.

In a more refined system, review value should distinguish between kinds of review. Some patients warrant chart review, some warrant bedside reassessment, and some warrant data refresh such as a repeated lab or closer monitoring.

Let the action set be $\mathcal{A} = \{\text{none}, \text{observe}, \text{review}, \text{escalate}\}$. Then the model can estimate

$$V_i(a,t) = \mathbb{E}\big[B_i(a,t) - C_i(a,t) \mid b_i(t)\big], \tag{9}$$

for each action a, where $b_i(t)$ is the belief state [21]. The chosen action is the highest-value one subject to capacity constraints in each action tier. This turns the surveillance problem into a small control problem, not merely a ranking problem. The queue becomes a policy over scarce review resources.

An important consequence is that confidence should modulate action nonlinearly. A high-risk, low-confidence patient is not equivalent to a low-risk patient. For such a case, the best action may be to request or prioritize more information rather than to escalate immediately. This means the system should reserve the strongest escalation for regions of the belief state where both urgency and support are high, while allowing moderate-support high-risk cases to enter lower-cost review tiers. This provides a principled route to selective escalation rather than crude alert suppression [22].

One can express this with tiered decision boundaries. For example,

$$\begin{aligned} a_i(t) &= \text{escalate} \quad \text{if} \quad \tilde{p}_i(t) \ge \tau_e \quad \text{and} \quad q_i(t) \ge \xi_e, \\ a_i(t) &= \text{review} \quad \text{if} \quad v_i(t) \ge \tau_r \quad \text{and} \quad a_i(t) \ne \text{escalate}, \end{aligned} \tag{10}$$

with τ_e, ξ_e, τ_r adapted to capacity. This makes explicit that risk and confidence enter escalation differently from how risk and informational deficit enter review.

The value-of-information perspective also yields a way to think about active sensing without overcommitting to algorithmic intervention. The system need

Table 3. Decomposition of the belief state used for monitoring and decisions.

Belief Component	Symbol	Interpretation	Operational Use
Urgency state	$z_i(t)$	Estimated deterioration level	Drives risk scoring
Confidence level	$q_i(t)$	Strength of evidence	Modulates escalation
Information deficit	$r_i(t)$	Missing or stale evidence	Signals review value
Belief vector	$b_i(t)$	Combined representation	Used in decision layer

Table 4. Modal contributions to the latent urgency representation.

Evidence Source	Symbol	Information Type	Effect on Belief
Structured signals	$z_i^{(s)}$	Physiologic measurements	Direct urgency update
Document signals	$z_i^{(d)}$	Clinical notes semantics	Contextual interpretation
Observation policy	$z_i^{(m)}$	Measurement patterns	Support and uncertainty cues
Fusion gate	$g_i(t)$	Adaptive weighting	Combines modalities

not autonomously order tests to benefit from modeling informational deficit. It can simply surface patients whose ranking is highly sensitive to missing or stale evidence [23]. In many cases, that is enough to guide human review. A patient near the queue boundary with large $r_i(t)$ is exactly the kind of case where targeted reassessment can improve allocation quality.

Finally, this formulation makes clear why calibration and confidence should be consumed together. If $\tilde{p}_i(t)$ is well calibrated but $q_i(t)$ is absent, then review allocation conflates high-probability well-supported cases with high-probability weakly supported ones. If $q_i(t)$ is available but $\tilde{p}_i(t)$ is badly calibrated, then review value can be mispriced in the upper tail. The queue needs both to be meaningful. The next section therefore turns to how the model should be trained so that the top of the queue reflects review value rather than only event discrimination.

5. Loss Design for the Top of the Queue

A service-aware model must be trained for the part of the score distribution that actually matters. In the ICU, that part is small. Most patient-time windows will never be near the review boundary, and many positives will be repeated adjacent windows from the same episode [24]. If the loss is dominated by these abundant but operationally unimportant cases, the resulting worklist can be mathematically good and clinically clumsy. Therefore the loss should be engineered around three facts: positives are rare, adjacent windows are dependent, and only a small tail of the ranking consumes meaningful attention.

The base loss can still be focal or imbalance-aware cross-entropy:

$$\mathcal{L}_{\text{base}} = -\sum_{i,t} \alpha y_i(t)(1-\hat{p}_i(t))^{\gamma} \log \hat{p}_i(t) - \sum_{i,t} (1-\alpha)(1-y_i(t))\hat{p}_i(t)^{\gamma} \log(1-\hat{p}_i(t)). \tag{11}$$

But this should be thought of as merely establishing a sensible background geometry for risk. By itself it does not optimize the thing the unit consumes.

To push review value into the top queue, define a soft queue weight

$$w_i(t) = \frac{\exp(\beta\, v_i(t))}{\sum_{j\in\mathcal{I}(t)} \exp(\beta\, v_j(t))}, \tag{12}$$

where β controls concentration. Then a queue-enrichment term is [25]

$$\mathcal{L}_{\text{queue}} = -\sum_{t} \sum_{i\in\mathcal{I}(t)} w_i(t)\, v_i^{\star}(t), \tag{13}$$

where $v_i^{\star}(t)$ is a proxy target for true review value. This target can be assembled from event occurrence, lead time, current treatment saturation, and perhaps delayed evidence of whether more information would have changed the downstream trajectory. The point is not to claim that $v_i^{\star}(t)$ is observed perfectly. The point is to train the model toward useful queue composition.

Because adjacent windows are redundant, the model should care more about state changes than about repeated persistence. A novelty term $n_i(t)$ can be defined

Table 5. Inputs used to compute expected review value.

Queue Variable	Symbol	Meaning	Influence
Calibrated risk	$\tilde{p}_i(t)$	Event probability estimate	Primary severity signal
Confidence	$q_i(t)$	Evidence reliability	Controls escalation strength
Information deficit	$r_i(t)$	Potential information gain	Promotes review cases
Actionability weight	$g_i(t)$	Lead-time sensitivity	Prioritizes actionable windows

Table 6. Action tiers used in review allocation.

Action	Description	Typical Trigger
None	No intervention	Low risk and high confidence
Observe	Continue monitoring	Moderate uncertainty
Review	Chart or clinician review	High review value
Escalate	Immediate clinical action	High risk and strong support

from the change in belief state or newly arrived evidence. Then examples can be reweighted by $\epsilon + n_i(t)$. Alternatively, one can penalize queue movement when novelty is low:

$$\mathcal{L}_{\text{turn}} = \sum_t \sum_{i \in \mathcal{I}(t)} (1 - n_i(t)) \left| w_i(t) - w_i(t-1) \right|. \quad (14)$$

This regularizer encourages the top of the queue to move when new information justifies movement, not because of small numerical instabilities in the score.

A related issue is that many labels are positive at different distances from the event, but not all such positives have equal review value [26]. This can be handled by defining a lead-time-sensitive target

$$u_i(t) = y_i(t) \exp\left(- \frac{d_i(t)}{\tau_d} \right), \quad (15)$$

where $d_i(t)$ is time to event and τ_d controls scale. This target can weight the queue-enrichment term or supervise a separate actionability head. The result is a queue that prefers not just positive cases, but cases whose position near the top is likely to matter.

Confidence should be trained as an operational variable, not merely as a side product of the encoder. If $\rho_i(t)$ is a support-quality proxy derived from evidence freshness, observation density, agreement across modalities, and support-set familiarity, then

$$\mathcal{L}_{\text{conf}} = \sum_{i,t} \Big(q_i(t) - \rho_i(t) \Big)^2 \quad (16)$$

encourages the confidence head to track evidence quality. A second term can encourage calibration stratified by confidence: [27]

$$\mathcal{L}_{\text{cal}} = \sum_m \omega_m \left(\frac{1}{|B_m|} \sum_{(i,t) \in B_m} \hat{p}_i(t) - \frac{1}{|B_m|} \sum_{(i,t) \in B_m} y_i(t) \right)^2, \quad (17)$$

with bins B_m formed jointly over score and confidence. This prevents the model from becoming sharply discriminative in the top queue while hiding severe probability distortion under some support regimes.

Observation-policy learning should also appear in the objective. Let $\hat{m}_i(t)$ be the predicted observation pattern and $m_i(t)$ the realized one. Then

$$\mathcal{L}_{\text{obs}} = - \sum_{i,t} \Big(m_i(t) \log \hat{m}_i(t) + (1 - m_i(t)) \log(1 - \hat{m}_i(t)) \Big) \quad (18)$$

or an appropriate multivariate extension encourages the latent state to explain why the record looks the way it does. This serves as a regularizer against implausible state representations that fit outcomes but ignore measurement policy. It also sharpens the model's understanding of informational deficit, because one can compare expected and actual observation patterns.

The full loss then becomes

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{base}} &+ \lambda_q \mathcal{L}_{\text{queue}} + \lambda_t \mathcal{L}_{\text{turn}} \\ &+ \lambda_f \mathcal{L}_{\text{conf}} + \lambda_c \mathcal{L}_{\text{cal}} + \lambda_o \mathcal{L}_{\text{obs}} + \lambda_w |0\Theta|0_2^2. \end{aligned} \quad (19)$$

Table 7. Training objectives shaping risk prediction and queue behavior.

Loss Component	Symbol	Objective	Benefit
Base classification	$\mathcal{L}_{base}$	Event prediction	Handles imbalance
Queue enrichment	$\mathcal{L}_{queue}$	Improve top ranking	Better review lists
Turnover regularizer	$\mathcal{L}_{turn}$	Stabilize queue movement	Reduces noise shifts
Confidence loss	$\mathcal{L}_{conf}$	Align support estimates	Reliable uncertainty

Table 8. Counterfactual observation experiments for robustness analysis.

Counterfactual Test	Modification	Evaluation Goal
Measurement thinning	Reduce lab frequency	Test confidence response
Observation inflation	Add extra measurements	Detect shortcut learning
Note delay	Shift documentation time	Assess semantic robustness
Support suppression	Remove selected signals	Measure uncertainty behavior

The purpose of this construction is not to make the loss elaborate for its own sake [28]. It is to ensure that the training objective reflects what the deployed system must do: form a stable, confidence-aware, review-efficient queue under observation bias.

A final point concerns validation. The best model is not necessarily the one with the best whole-distribution discrimination. It is the one with the highest utility in the review queue under realistic capacity, acceptable turnover, and tolerable calibration error in the high-confidence top tail. These criteria may prefer a model that looks slightly worse on a conventional leaderboard. That is not a weakness of the model. It is a sign that the evaluation object has finally been aligned with the operational one.

6. Counterfactual Observation Worlds

A model that treats observation policy as signal should be tested by changing that policy in thought experiments [29]. Retrospective train-test splits do not reveal how strongly the system depends on the current meaning of missingness, note density, or measurement timing. If a unit begins measuring more aggressively, if note-writing becomes sparser, or if the patient's state is equally severe but less explicitly documented, will the queue still behave sensibly? These are not peripheral questions. They are the right robustness questions for a system built around endogenous observation.

Counterfactual evaluation therefore should include modified observation worlds. One type of test delays or thins structured measurements while keeping latent trajectories approximately the same. Another delays or suppresses note arrivals. Another alters observation density in ways consistent with policy change rather than random corruption. Let $\mathcal{O}_i^{(\pi)}(t)$ denote the observed record under a counterfactual observation policy π. Then one can compare [30]

$$\begin{aligned} \Delta_v^{(\pi)}(t) &= v_i^{(\pi)}(t) - v_i(t), \\ \Delta_q^{(\pi)}(t) &= q_i^{(\pi)}(t) - q_i(t). \end{aligned} \tag{20}$$

The key question is not whether the score changes at all. It should. The key question is whether the score changes in the right way. If structured support thins moderately, confidence should usually fall and review value may rise for information-seeking reasons rather than collapsing or remaining falsely stable. If note density changes but structured evidence remains strong, urgency should not shift wildly.

A particularly informative counterfactual concerns observation inflation. If a model has learned that dense measurement implies acuity, then artificially increasing measurement density for stable patients should not push them into the queue if actual values remain benign. Running this test reveals whether the model is using observation policy as context or as a shortcut [31]. Conversely, artificially suppressing measurements in high-risk trajectories should increase uncertainty and perhaps review value, but should not necessarily erase urgency if other evidence remains strong.

One can also use counterfactual re-observation to estimate value-of-information more directly. Suppose a candidate measurement j is hypothetically refreshed. The difference in queue utility before and after the hypothetical update approximates the review value of ob-

Table 9. Operational signals logged during deployment for monitoring.

Deployment Signal	Logged Variable	Purpose	Monitoring Outcome
Queue placement	Rank position	Track prioritization	Detect drift
Review activity	Review indicator	Measure human response	Closed-loop effects
Observation change	Measurement density	Identify policy shifts	Support recalibration
Intervention timing	Treatment updates	Detect escalation patterns	Outcome interpretation

taining it:

$$V_{i,t}^{(j)} = \mathbb{E}\Big[v_i^{(j)}(t^+) - v_i(t) \mid \mathcal{O}_i(t)\Big]. \tag{21}$$

This quantity need not be perfect to be useful. It tells the system whether the case is near a decision boundary that could be clarified cheaply. In deployment, even a rough estimate can help distinguish high-risk cases that need escalation from moderate-risk cases that need better information first.

Counterfactual tests should also examine queue-level behavior [32]. Under altered observation policies, does the top-K queue remain enriched? Does turnover become erratic? Does calibration in high-confidence regions remain acceptable? These questions are more informative than whole-population accuracy because the queue is where service capacity is actually spent.

The broader methodological point is that robustness in ICU surveillance should be tested against plausible changes in how evidence is gathered, not only against noise in the values already present. This follows directly from the belief that observation policy is part of the state. A model built on that belief should be stress-tested accordingly.

7. Deployment Is a Closed Loop

Once a service-aware deterioration model is deployed, it stops being a passive estimator. It begins changing the world that it predicts. High-ranked patients may receive more chart review, more bedside visits, more notes, more labs, or earlier intervention [33]. As a result, future observations no longer arise from the predeployment policy alone. They arise from a mixture of ordinary care and model-mediated attention. This means deployment creates a closed loop. If the system is retrained without accounting for this, it can end up learning from its own downstream effects as if they were natural data.

Closed-loop deployment creates three challenges. The first is interpretive. A patient who received intensive review because of the model and then did not deteriorate may look like a false positive even though the signal may have been clinically useful. The second is statistical [34]. Observation density and intervention timing become conditioned on the model's own outputs. The third is operational. If the model changes the support structure of the very patients it ranks highly, then its confidence model and calibration map may drift in nontrivial ways.

The practical response is careful logging. A deployment-grade system should record not only scores and outcomes, but also whether a patient entered the review queue, whether additional measurements followed, whether note activity changed, and whether escalatory interventions occurred soon after queue elevation. These are not optional analytics. They are the minimum record needed to distinguish model-mediated trajectory changes from ordinary drift.

One should also monitor queue quality over time [35]. If the same queue size is maintained but the average confidence within the queue falls, then the unit may be spending scarce attention on weaker evidence than before. If queue turnover rises while patient acuity appears stable, the system may be responding to observation-policy drift rather than to true urgency. If subgroup composition of the queue changes after a documentation workflow update, then the model may have become biased toward certain observation regimes. All of these are service-level failures even if AUROC on delayed labels remains superficially stable.

Adaptation under deployment should therefore proceed in layers. The safest first move is usually recalibration of the score and confidence map. The next is adjustment of review thresholds or capacities. Core retraining of the latent-state model should be more conservative, because large changes there can alter queue semantics in opaque ways [36]. A constrained update rule is useful:

$$\theta_{t+1}^{\star} = \arg\min_{\theta} \mathcal{L}_{\text{recent}}(\theta)$$
$$\text{s.t.} \quad \mathbb{E}\big[|0 f_{\theta}(\mathcal{O}) - f_{\theta_t}(\mathcal{O})|0_2\big] \le \epsilon. \tag{22}$$

This guards against abrupt shifts in worklist behavior that would otherwise undermine trust and complicate human adaptation.

A final deployment principle is bounded autonomy. Because the system allocates attention rather than making definitive diagnoses, it should remain transparent about why a patient is high in the queue and how much support that placement has. The queue should therefore surface urgency, confidence, and perhaps whether

the recommendation is driven mainly by high event risk or by high informational value. Such distinctions help users understand whether to escalate, to review, or simply to obtain more information. In a closed-loop environment, that transparency is not only good interface design. It is part of maintaining safe human control over a system that affects what is measured next [37].

8. Conclusion

The usual framing of ICU deterioration prediction starts too late. It begins after the record has already been flattened into inputs and after the observation process has been treated as a side issue. A more faithful starting point is earlier and less convenient: what the model sees is not the patient alone, but the patient as filtered through selective attention, selective measurement, selective documentation, and selective intervention. Once that is admitted, the nature of the modeling problem changes. The task is not simply to infer who will deteriorate. It is to infer what is currently believed about the patient, how strongly that belief is supported, what remains importantly unknown, and how scarce review effort should respond.

That change in viewpoint has several consequences. It means masks and elapsed times are not only technical accessories but traces of clinical policy [38]. It means notes are not just semantic vectors but signs that someone chose to compress and communicate a trajectory. It means uncertainty should be decomposed into support quality and informational deficit rather than hidden inside a single probability. It means the model's most useful output may not be an alarm at all, but a queue position backed by an estimate of why more attention is or is not worth spending. And it means the model must be judged by what it does to the review queue, not only by what it does to a held-out label.

A service-aware system built on those principles is more demanding than a standard classifier. It requires a latent-state model, an observation-policy model, a confidence head, a review-value head, and a queue-facing decision layer. It requires losses that sculpt the top of the ranking rather than the score field in general. It requires counterfactual tests that alter measurement patterns instead of only perturbing values [39]. It requires deployment logging that acknowledges that the model changes what gets measured after it is introduced. None of these demands are small. But they are not artificial complexity. They are the complexity that was already present in the ICU and merely hidden by the simpler formulation.

There is also a deeper lesson here about clinical machine learning. In many domains, better prediction is imagined as the straightforward result of more data and more expressive models. In the ICU, more data do not eliminate the harder issue that the data are action-conditioned. A forecasting system that ignores that fact may still recover useful correlations, but it will forever struggle to distinguish high risk from high uncertainty, dense observation from dense concern, and score magnitude from marginal utility of review [40]. A system that models the observation world as part of the clinical world stands a better chance of handling these distinctions honestly.

The argument of this paper is therefore not that one particular architecture is mandatory. It is that the target of design should be a belief state, not a window label. A belief state is richer because it carries urgency, support, and informational need simultaneously. That richness matters because it gives the surveillance engine somewhere to put the things clinicians actually care about but ordinary classifiers cannot express well: whether the evidence is fresh, whether the state estimate is fragile, whether a patient is already saturated with attention, whether another review could still change something, and whether the queue is being spent wisely.

A long-term research agenda follows naturally from this reframing. One direction is better estimation of value of information under realistic clinical constraints, so that review queues can separate high-risk patients from high-uncertainty patients without conflating them. Another is explicit modeling of the support process for notes and structured observations, so that documentation and missingness can be interpreted less as crude proxies and more as policy traces [41]. Another is closed-loop evaluation, where a model is judged partly by how it changes observation patterns and intervention timing after deployment. Yet another is subgroup analysis framed in terms of access to scarce attention rather than only in terms of score calibration. These are not side questions. They are the questions that become visible only after the problem is named correctly.

If there is a single practical takeaway, it is this: an ICU deterioration model should not aspire merely to be a better event detector. It should aspire to be a better steward of uncertainty under scarce attention. That is a harder ambition, because it forces the system to confront missingness, support, intervention, queue pressure, and feedback all at once. But it is also a more realistic ambition, and therefore a more useful one [42].

References

[1] I. Bayramli, V. Castro, Y. Barak-Corren, E. M. Madsen, M. K. Nock, J. W. Smoller, and B. Y. Reis, "Predictive structured-unstructured interactions in ehr models: A case study of suicide prediction," 8 2021.

[2] K. B. Wagholikar, S. V. Joshi, V. Vernekar, Y. Ostrovsky, S. D. Desai, P. B. Magdum, S. B. Wakle, S. Jain, A. Zagade, R. Patel, and S. N. Murphy, "Cloud services for patient cohort identification using the informatics for integrating biology and the bedside platform.," BioMed research international, vol. 2020, pp. 2851713–2851713, 7 2020.

[3] L. Warren and H. Chi, "Acm southeast regional conference - securing ehrs via cpma attribute-based encryption on cloud systems," in Proceedings of the 2014 ACM Southeast Regional Conference, pp. 20–7, ACM, 3 2014.

[4] B. M. Hollister and V. L. Bonham, "Should electronic health record-derived social and behavioral data be used in precision medicine research?," AMA journal of ethics, vol. 20, pp. 873–880, 9 2018.

[5] 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.

[6] C. D. Williams and M. J. Kelley, "Automated extraction of quality oncology practice initiative (qopi) quality measures from the veterans health administration (vha) electronic health record system.," Journal of Clinical Oncology, vol. 29, pp. e16560–e16560, 5 2011.

[7] 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.

[8] R. L. Richesson, "Data standards in diabetes patient registries.," Journal of diabetes science and technology, vol. 5, pp. 476–485, 5 2011.

[9] S. Mataraso, V. Socrates, F. Lekschas, and N. Gehlenborg, "Halyos: A patient-facing visual ehr interface for longitudinal risk awareness," 4 2019.

[10] C. Violán, Q. Foguet-Boreu, E. Hermosilla-Pérez, J. M. Valderas, B. Bolíbar, M. Fàbregas-Escurriola, P. Brugulat-Guiteras, and M. Ángel Muñoz-Pérez, "Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity.," BMC public health, vol. 13, pp. 251–251, 3 2013.

[11] M. Jaboyedoff, M. Rakic, S. Bachmann, C. Berger, M. Diezi, O. Fuchs, U. Frey, A. Gervaix, G. As, M. A. Grotzer, U. Heininger, C. R. Kahlert, D. Kaiser, M. V. Kopp, R. Lauener, T. J. Neuhaus, P. Paioni, K. M. Posfay-Barbe, G. P. Ramelli, U. Simeoni, G. D. Simonetti, C. Sokollik, B. D. Spycher, and C. E. Kuehni, "Swisspeddata: Standardising hospital records for the benefit of paediatric research," 6 2021.

[12] V. Vydiswaran, X. Zhao, and D. Yu, "Data science and natural language processing to extract information in clinical domain," in Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), pp. 352–353, ACM, 1 2022.

[13] V. Granit, A.-L. Grignon, J. Wuu, J. S. Katz, D. Walk, S. Hussain, J. Hernandez, C. E. Jackson, J. Caress, T. Yosick, N. Smider, and M. Benatar, "Harnessing the power of the electronic health record for als research and quality improvement: Create capture-als and the als toolkit.," Muscle & nerve, vol. 65, pp. 154–161, 11 2021.

[14] S. Batra and S. Sachdeva, Managing Large-Scale Standardized Electronic Health Records, pp. 201–219. Springer India, 10 2016.

[15] L. J. Chavez, J. E. Richards, P. Fishman, K. Yeung, A. Renz, L. M. Quintana, S. Massimino, and R. B. Penfold, "Cost of implementing an evidence-based intervention to support safer use of antipsychotics in youth.," Administration and policy in mental health, vol. 50, pp. 725–733, 6 2023.

[16] I. M. Miake-Lye, A. M. Cogan, S. Mak, J. Brunner, S. Rinne, C. E. Brayton, A. Krones, T. E. Ross, J. T. Burton, and M. Weiner, "Transitioning from one electronic health record to another: A systematic review.," Journal of general internal medicine, vol. 38, pp. 956–964, 10 2023.

[17] R. Foraker, S. Patel, Y. Khan, M. A. Bauman, and J. Bower, "Abstract 20: Cardiovascular health trends in electronic health record data (2010-2015): the guideline advantage," Circulation, vol. 135, 3 2017.

[18] W. Zhu and N. Razavian, CHIL - Variationally regularized graph-based representation learning for electronic health records. ACM, 4 2021.

[19] A. Sheikhtaheri, S. M. T. Jabali, E. Bitaraf, A. TehraniYazdi, and A. Kabir, "A near real-time electronic health record-based covid-19 surveillance system: An experience from a developing country.," Health information management : journal of the Health Information Management Association of Australia, vol. 53, pp. 145–154, 7 2022.

[20] M. C. Lim, M. V. Boland, C. A. McCannel, A. Saini, M. F. Chiang, K. D. Epley, and F. Lum, "Adoption of electronic health records and perceptions of financial and clinical outcomes among ophthalmologists in the united states.," JAMA ophthalmology, vol. 136, pp. 164–170, 2 2018.

[21] H. Angier, R. Gold, C. Crawford, J. P. O'Malley, C. J. Tillotson, M. Marino, and J. E. DeVoe, "Linkage methods for connecting children with parents in electronic health record and state public health insurance data.," Maternal and child health journal, vol. 18, pp. 2025–2033, 2 2014.

[22] B. Lo, L. Sequeira, G. Strudwick, D. Jankowicz, K. Almilaji, A. Karunaithas, D. Hang, and T. Tajirian, "Accuracy of physician electronic health record usage analytics using clinical test cases.," 10 2022.

[23] Y. Zou, A. Pesaranghader, A. Verma, D. Buckeridge, and Y. Li, "Modeling electronic health record data using a knowledge-graph-embedded topic model," 6 2022.

[24] A. Sinha, L. A. Stevens, F. Su, N. M. Pageler, and D. S. Tawfik, "Measuring electronic health record use in the pediatric icu using audit-logs and screen recordings.," Applied clinical informatics, vol. 12, pp. 737–744, 8 2021.

[25] M. T. Quasim, M. M. Mobarak, K. U. Nisa, M. Meraj, and M. Z. Khan, "Blockchain-based secure health records in the healthcare industry," in 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), vol. 2020, pp. 545–549, IEEE, 4 2023.

[26] D. C. Classen, A. J. Holmgren, Z. Co, L. P. Newmark, D. L. Seger, M. Danforth, and D. W. Bates, "National trends in the safety performance of electronic health record systems from 2009 to 2018.," JAMA network open, vol. 3, pp. e205547–, 5 2020.

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

[28] K. G. Blumenthal, W. W. Acker, Y. Li, N. S. Holtzman, and L. Zhou, "Allergy entry and deletion in the electronic health record," Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology, vol. 118, pp. 380–381, 1 2017.

[29] K. Fujita, K. Onishi, T. Takemura, and T. Kuroda, "The improvement of the electronic health record user experience by screen design principles," Journal of medical systems, vol. 44, pp. 21–, 12 2019.

[30] R. Li, Y. Chen, and J. H. Moore, "Integration of genetic and clinical information to improve imputation of data missing from electronic health records," Journal of the American Medical Informatics Association : JAMIA, vol. 26, pp. 1056–1063, 4 2019.

[31] K. Marsolo, "Informatics and operations–let's get integrated," Journal of the American Medical Informatics Association : JAMIA, vol. 20, pp. 122–124, 9 2012.

[32] M. F. Kacamarga, A. Budiarto, and B. Pardamean, "A platform for electronic health record sharing in environments with scarce resource using cloud computing," International Journal of Online and Biomedical Engineering (iJOE), vol. 16, pp. 63–76, 8 2020.

[33] S. E. Golden, E. R. Hooker, S. Shull, M. Howard, K. Crothers, R. F. Thompson, and C. G. Slatore, "Validity of veterans health administration structured data to determine accurate smoking status.," Health informatics journal, vol. 26, pp. 1507–1515, 11 2019.

[34] N. Zong, A. Wen, S. Moon, S. Fu, L. Wang, Y. Zhao, Y. Yu, M. Huang, Y. Wang, G. Zheng, M. M. Mielke, J. R. Cerhan, and H. Liu, "Computational drug repurposing based on electronic health records: a scoping review.," NPJ digital medicine, vol. 5, pp. 77–77, 6 2022.

[35] A. V. Masoe, A. S. Blinkhorn, K. Colyvas, J. Taylor, and F. A. Blinkhorn, "Reliability study of clinical electronic records with paper records in the nsw public oral health service.," Public health research & practice, vol. 25, pp. e2521519–, 3 2015.

[36] P. D. Rebouças, "Electronic health records in dental education: A scoping review and quantitative analysis of publications," Journal of Clinical Medical Research, pp. 1–16, 10 2022.

[37] A. Beresniak, A. Schmidt, D. Dupont, M. Sundgren, D. Kalra, and G. D. Moor, "Improving performance of clinical research: Development and interest of electronic health records," BioMed research international, vol. 2015, pp. 864549–864549, 10 2015.

[38] P. S. Sen and N. Mukherjee, "A case study on implementation of health data standards for smart healthcare," International Journal of Engineering Research in Computer Science and Engineering, vol. 9, pp. 17–25, 6 2022.

[39] T. Sawant, P. Idayakumar, A. Sabkale, and K. Pampattiwar, "Decentralized ehr storage using blockchain," ECS Transactions, vol. 107, pp. 6397–6405, 4 2022.

[40] K. J. S. Brady, A. Legler, and W. G. Adams, "Exploring opportunities for household-level chronic care management using linked electronic health records of adults and children: A retrospective cohort study.," Maternal and child health journal, vol. 24, pp. 829–836, 5 2020.

[41] S. M. Albert, P. McCracken, T. Bui, J. Hanmer, G. S. Fischer, J. Hariharan, and A. E. James, "Do patients want clinicians to ask about social needs and include this information in their medical record?," BMC health services research, vol. 22, pp. 1275–1275, 10 2022.

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