Continuous Risk Prioritization for ICU Deterioration Under Limited Clinical Attention

Bilal Mahmood[1] **and Tariq Mehmood**[2]

[1]Department of Computer Science, University of Chakwal, Talagang Road, Chakwal 48800, Pakistan

[2]Department of Software Engineering, University of Swabi, Anbar Road, Swabi 23430, Pakistan

ABSTRACT

Critical care units have become increasingly instrumented environments, but the practical value of prediction depends not only on whether deterioration can be anticipated, but on whether scarce clinical attention can be directed toward the right patients at the right time. In most machine learning studies, deterioration is framed as a binary forecasting problem, and model quality is summarized with threshold-free discrimination metrics. That framing is useful for benchmarking, yet it obscures the operational structure of intensive care surveillance. Clinicians do not respond to isolated probabilities in the abstract. They allocate monitoring effort, bedside reassessment, diagnostic work, and intervention readiness across a continuously changing population of patients under hard constraints on time and staffing. This paper develops a computer-science-centered formulation of ICU deterioration modeling as a continuous risk prioritization problem rather than as a simple binary alarm problem. The framework treats each patient as a partially observed stochastic process whose risk estimate determines position in a dynamic ranked worklist. Emphasis is placed on queue-constrained surveillance, temporal smoothness, uncertainty-aware escalation, and the geometry of the high-risk tail where human review capacity is actually spent. A mathematical treatment is introduced for dynamic priority scoring, workload-limited thresholding, and risk propagation under intervention-sensitive trajectories. The paper also examines how rare-event learning, calibration, selective abstention, and distribution shift interact with ranking-based deployment. The central argument is that deterioration systems are best evaluated and designed as streaming prioritization engines embedded in an attention-limited clinical environment, where useful prediction is inseparable from scheduling, reliability, and operational control.

KEYWORDS

1. Introduction

The intensive care unit is often described as one of the richest data environments in modern medicine [1]. Patients are monitored frequently, laboratory values are updated throughout the day, medications are recorded, devices produce streams of measurements, and clinical teams document evolving assessments in the electronic record. From the standpoint of machine learning, this abundance appears ideal. There is temporal data, multimodal data, repeated labels, and clinically important outcomes. Yet the practical success of prediction in this setting has been more modest than the richness of the data might suggest. One reason is that deterioration forecasting is usually posed as a clean supervised classification problem when the deployed task is not classification in the ordinary sense. The operational question in intensive care is not merely whether a patient will deteriorate within a future horizon. The operational question is how limited clinician attention should be distributed across a ward or unit full of patients whose risks, trajectories, and evidential support change hour by hour.

This difference between benchmark formulation and deploy-

Article info:

Bilal Mahmood and Tariq Mehmood. *Continuous Risk Prioritization for ICU Deterioration Under Limited Clinical Attention.* Nextqueries. Publishing

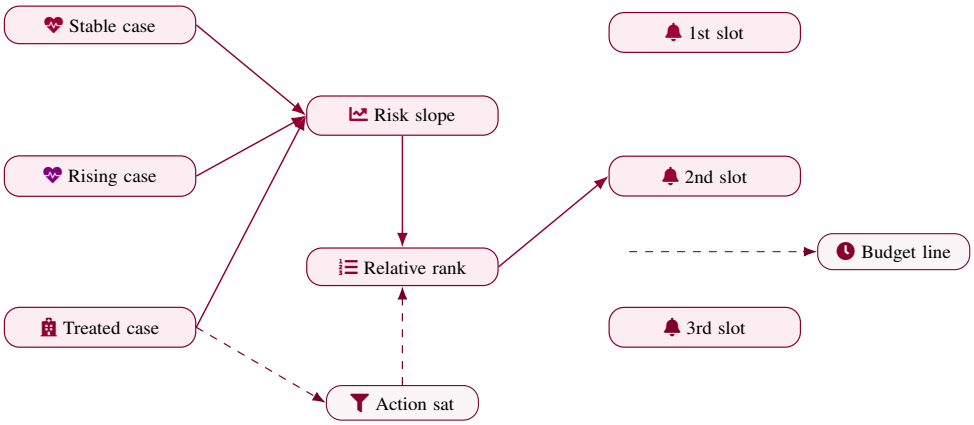

Figure 1 The ranked worklist is a dynamic queue object. Patient trajectories are converted into slope-aware relative positions, while treatment saturation prevents already-attended cases from occupying scarce review slots without additional operational value; the dashed boundary marks the effective attention budget.

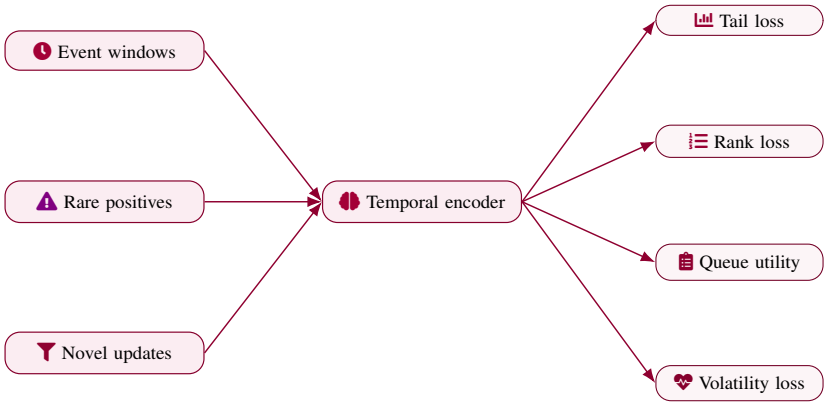

Figure 2 Learning is oriented toward the queue-facing tail of the score distribution instead of average window-wise fit alone. Rare-event imbalance, temporal redundancy, and worklist stability are handled jointly so that true urgent cases concentrate near the actionable top ranks without inducing excessive churn.

ment reality has important computational consequences. A benchmark classifier seeks to map the current record to a future label. A surveillance system, by contrast, must continuously rank patients, update priorities as new information arrives, decide when to escalate and when to remain quiet, and do so under hard limits on how many high-risk cases can be meaningfully reviewed at once. The system is therefore not just a predictor of events. It is part of a larger allocation process, one in which scores are converted into a dynamically ordered worklist. The meaning of a probability in that environment depends not only on calibration or rank quality in the abstract, but on how the score changes the order in which clinicians allocate scarce time. A perfectly respectable retrospective model can fail to add value if its high-risk tail is too broad, its alerts are too unstable, or its ranking is concentrated in states that are already obvious to clinicians and therefore consume rather than save attention.

The distinction is especially important in low-prevalence event prediction. In many ICU deterioration tasks, the hourly probability of a new event within a fixed horizon is only a few percent. Under such prevalence, even strong discrimination does not automatically translate into a small set of actionable alerts. A model may rank positive windows above negative windows with reasonable success while still producing too many moderately elevated scores to support a feasible intervention policy. This is not merely a matter of choosing a better threshold [2]. It is a matter of recognizing that the output of the model is consumed by a queueing process. Patients above some implicit or explicit priority boundary compete for review, reassessment, and escalation. Once that is acknowledged, the objective of learning changes. The system should not only separate classes in the statistical sense. It should shape a risk landscape that yields stable and useful rankings under capacity constraints.

A further complication is that ICU risk is not static. The same patient can move from low concern to high concern and back again over the course of a day, and these movements are driven by both physiology and treatment. A blood pressure trajectory, a change in oxygen requirement, an intervention on the ventilator, or a new radiology interpretation can all alter the level and urgency of surveillance required. Consequently, prediction should be thought of less as assigning a label to a frozen snapshot and more as maintaining a continuously updated priority signal over a partially observed stochastic process. The model does not simply answer, once, whether deterioration will occur. It revises a running estimate of who should be near the

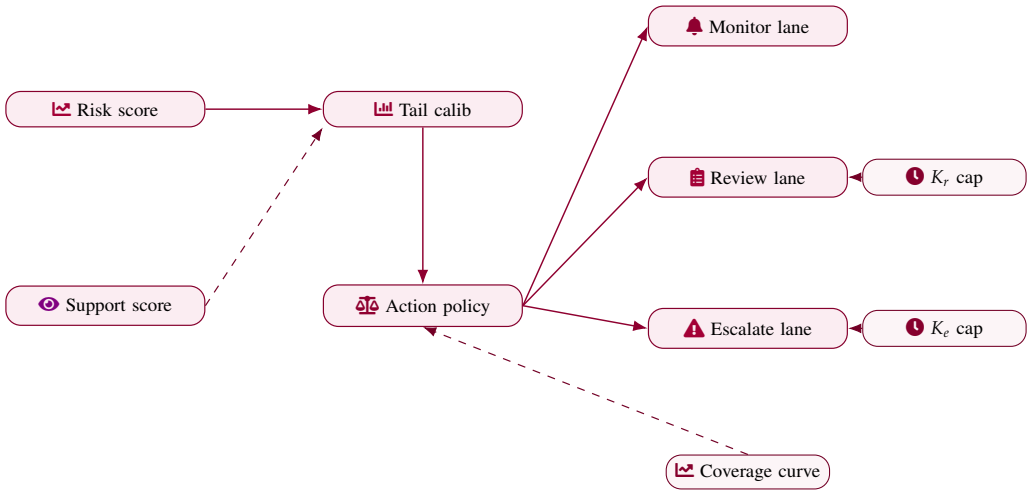

Figure 3 A calibrated continuous score supports graded action rather than a single brittle alarm threshold. Confidence-aware policy separates passive monitoring, active review, and full escalation, while explicit capacity limits keep review and escalation volumes consistent with finite staffing and attention.

top of the worklist now, who is rising toward the top, and who can be safely deprioritized. This dynamic view of risk is much closer to the real function of surveillance in a critical care unit than a one-time binary decision.

This way of formulating the problem also changes the role of calibration. In standard predictive modeling, calibration is the degree to which estimated probabilities correspond to empirical event frequencies. That remains important, but in a prioritization system calibration has an additional interpretation. It determines whether the distance between scores meaningfully corresponds to a difference in expected burden on the review queue. A score gap between two patients matters only if it conveys a reliable difference in urgency. Overconfident upper tails create queue congestion by flooding the list with false urgency. Underconfident upper tails flatten the ranking and obscure who should be reviewed first [3]. Therefore, calibration and ranking are not separate virtues. They jointly determine the geometry of the queue-facing score distribution.

A qualitative anchor for this perspective appears in Sadanandan (2025), who treats the model primarily as a continuous risk score rather than a binary alarm [4]. That statement is more consequential than it first appears. Once the output is interpreted as a continuous score intended to prioritize clinical attention, the central design problem is no longer exhausted by optimizing discrimination against a binary label. It becomes necessary to ask how the score will behave as a ranking signal, how many patients will occupy its upper tail under realistic prevalence, how stable the ranking will be across adjacent hours, and how a unit should allocate action when there is not enough capacity to respond to every elevated prediction. In other words, the computational object shifts from alarm generation to workload-aware continuous prioritization.

This paper develops that shift explicitly. The central thesis is that ICU deterioration forecasting should be designed and analyzed as a queue-constrained streaming prioritization problem over partially observed patient trajectories. Such a formulation does not deny the relevance of event prediction. Rather, it embeds event prediction inside the larger system that gives the prediction operational meaning. The paper therefore emphasizes risk ranking, dynamic patient ordering, action capacity, temporal smoothness, and support-aware escalation. The objective is not simply to identify future positive labels but to maintain a clinically useful ordering over the current population under realistic limits on attention and intervention bandwidth.

The paper proceeds in a deliberately different layout from standard prediction manuscripts. It first examines what the computational object of ICU surveillance actually is once the binary-alarm simplification is removed. It then formulates surveillance as a queue-constrained ranking problem in which each score occupies a position in a dynamic worklist. Next, it develops a state-space view of patient priority as a flow variable that should evolve smoothly but responsively under new evidence. A later section addresses learning objectives under rare events when the high-risk tail rather than the full score distribution determines utility. The paper then turns to selective escalation, calibration, and the economics of false-alert burden. The final body section studies operational drift, system governance, and the constraints imposed by real-time deployment. Throughout, the emphasis remains more computational than medical [5]. The ICU appears here as a setting for sequential decision support under attention scarcity, where risk modeling, ranking, and queue management are inseparable parts of one systems problem.

2. The Computational Object of Deterioration Surveillance

A great deal of confusion in predictive modeling arises from uncertainty about what object is actually being built. In ordinary classification the object is straightforward: a mapping from inputs to classes or probabilities. In ICU deterioration surveillance the object is richer. The system receives a stream of admissions, discharges, measurements, notes, and interventions. At every time point it emits scores over the active census. Those

Aspect	Benchmark classifier view	Continuous prioritization view
Problem framing	Binary prediction of event within horizon	Streaming prioritization under attention constraints
Output object	Per-window probability or label	Dynamic ranked worklist over active census
Primary metric	Threshold-free discrimination (AUROC, AUPRC)	Tail utility, queue quality, event capture at fixed capacity
Operational use	Isolated alarms compared to labels	Continuous ordering that guides review and escalation
Capacity treatment	Often ignored or folded into "alert rate"	Explicit top-K or budget-constrained queue

Table 1 Conceptual contrast between alarm-centric prediction and continuous risk prioritization.

Object	Symbol	Domain	Role in surveillance
Active census	$\mathcal{I}(t)$	Set of ICU patients at time t	Population over which ranking is defined
Latent state	$x_i(t)$	Underlying deterioration-relevant state	Drives event risk and dynamics
Observations	$\mathcal{O}_i(t)$	Admissible chart data at time t	Evidence used to infer risk
Risk score	$s_i(t)$	Real-valued priority signal	Determines position in worklist
Ranking	π_t	Ordering of $\mathcal{I}(t)$ by $s_i(t)$	Defines queue-facing priority
Queue set	Q_t	Top-K_t patients by rank	Receives additional review / escalation
Event label	$Y_i(t,H)$	Binary outcome within horizon H	Supervisory signal for learning

Table 2 Core computational objects in queue-constrained ICU deterioration surveillance.

scores are then consumed by clinicians, dashboards, escalation teams, or protocolized review processes. The output is therefore not just a set of probabilities. It is an ordering over patients that evolves over time and interacts with operational constraints. The correct computational object is a dynamic priority field over the active patient population.

To formalize this, let $\mathcal{I}(t)$ denote the set of patients present in the ICU at time t. For each patient $i \in \mathcal{I}(t)$, let $x_i(t)$ denote the latent clinical state relevant to deterioration, and let $\mathcal{O}_i(t)$ denote the admissible observations available at time t. A conventional predictor approximates

$$p_i(t;H) = \Pr\Big(Y_i(t,H) = 1 \mid \mathcal{O}_i(t)\Big), \tag{1}$$

where $Y_i(t,H)$ indicates deterioration within horizon H. A prioritization engine instead constructs a score

$$s_i(t) = f\big(\mathcal{O}_i(t)\big) \tag{2}$$

that is consumed not as an isolated value but through the induced ranking

$$\pi_t = \operatorname{argsort}\Big(\{s_i(t)\}_{i\in\mathcal{I}(t)}\Big). \tag{3}$$

This ranking π_t is the object that directly affects workflow. Its stability, concentration, and response to evidence determine whether the model is clinically useful. Two models with similar per-window AUROC can create very different π_t processes. One may generate a compact, interpretable top queue with modest turnover. Another may cause constant reshuffling in the top decile without yielding a manageable set of review candidates [6]. The latter may be statistically competent yet operationally poor.

The fact that the output is a ranking process rather than a list of isolated probabilities changes the meaning of error. A false positive in the classical sense is one patient-time window assigned a high score despite no future event. But in prioritization, the more relevant question is whom that patient displaces in the worklist. If patient i is erroneously moved to the top ten, the cost is not abstract misclassification alone. It is the possibility that another patient j, perhaps with genuinely rising risk, is pushed below the line of feasible review. Likewise, a false negative is not merely an unflagged window. It is a patient who remains too low in the queue to receive timely additional attention. Thus, prioritization error is relational. It depends on how scores compare across the active census.

This relational perspective suggests that the geometry of the upper tail matters more than the geometry of the bulk. Most ICU

Metric	Definition	Interpretation	Queue role
Queue size	K_t	Max patients for active review at t	Capacity constraint per cycle
Admission indicator	$a_i(t) = \mathbb{I}(i \in Q_t)$	Whether patient enters review queue	Drives workload and attention
Turnover	$T_t = 1 - \frac{\lvert Q_t \cap Q_{t-1} \rvert}{\lvert Q_t \cup Q_{t-1} \rvert}$	Fractional change in queue membership	Measures stability of worklist
Arrival rate	λ_t	Effective rate of new queued cases	Links alerts to service capacity
Service rate	μ_t	Processing rate of reviews / escalations	Governs expected waiting time
Queue cost	$C_t^{\text{queue}} = \zeta(\max(0, \lambda_t - \mu_t))^2$	Penalty for sustained overload	Used in tuning thresholds and policies

Table 3 Queue-level quantities characterizing workload and worklist behavior.

Quantity	Symbol	Meaning	Effect on prioritization
Latent priority state	$z_i(t)$	Dynamic summary of risk, context, support	Underlies score evolution
Instantaneous score	$s_i(t) = G(z_i(t))$	Raw deterioration risk estimate	Base signal for ranking
Relative priority	$r_i(t) = \frac{s_i(t) - \bar{s}(t)}{\sigma_s(t) + \varepsilon}$	Score normalized within census	Captures queue-relative urgency
Score slope	$\dot{s}_i(t)$	Short-term change in $s_i(t)$	Encodes rising or falling concern
Priority flow	$\dot{r}_i(t)$	Motion through worklist over time	Drives entry/exit from Q_t
Confidence	$c_i(t)$	Reliability of $s_i(t)$ under evidence	Modulates action and escalation

Table 4 Dynamic state and priority-flow variables for evolving ICU risk.

patients at any hour are not on the verge of a new deterioration event. The work of the ranking system is to separate a relatively small number of genuinely urgent or rising-risk patients from a much larger low-urgency background. If the model spreads moderate risk across too many patients, then the queue becomes uninformative. If it produces an unstable tail that reorders aggressively from hour to hour based on small fluctuations, clinicians cannot form a coherent response policy. The desired output is therefore not merely a calibrated probability field. It is a sharply concentrated yet temporally stable upper tail.

This is one reason why standard performance summaries are incomplete. AUROC evaluates pairwise ranking over the entire score distribution and all operating points. AUPRC focuses more on the rare positive class, but it still averages over thresholds and says little about the temporal persistence of the highest ranks [7]. Neither metric captures how many patients occupy the effective queue at a given alert budget, how often patients enter or leave the queue, or whether queue membership is dominated by patients already under maximal attention. These omissions do not make the metrics useless. They simply mean that they are properties of the prediction model, not sufficient characterizations of the surveillance system.

Another consequence is that the model should distinguish between score magnitude and queue urgency. Score magnitude is a statistical property. Queue urgency is a systems property that depends on the census size, the prevalence of high-risk states, and the available capacity for action. A score of 0.20 may be operationally urgent in one unit and operationally negligible in another depending on how many patients have still higher scores and how many can be realistically reviewed. Therefore one should think of the score as a latent urgency coordinate whose meaning is contextualized by the current empirical distribution over the unit. This implies that deployment logic must be dynamic. Static thresholds can be useful, but a fully specified surveillance system should also support quantile-based or capacity-aware prioritization rules.

One can make this precise by defining a queue-admission set $Q_t \subseteq \mathcal{I}(t)$ consisting of patients eligible for active review at time t. If the unit can review at most K_t patients at a decision cycle, then

$$Q_t = \{i \in \mathcal{I}(t) : \text{rank}_{\pi_t}(i) \leq K_t\}. \tag{4}$$

The clinically relevant output of the model may then be Q_t rather than the full vector of scores. Under this view, all learning and calibration choices should be understood by how they shape Q_t, its composition, and its turnover. A model that marginally

Term	Symbol	Primary focus	Queue-level effect
Prediction loss	$\mathcal{L}_{\text{focal}}$	Rare-event classification with class weighting	Improves retrieval of positives overall
Rank loss	$\mathcal{L}_{\text{rank}}$	Pairwise ordering within strata	Sharpens separation of risk levels
Tail concentration	$\mathcal{L}_{\text{tail}}$	Soft top-K enrichment of positives	Packs utility into upper tail
Queue utility term	$\mathcal{L}_{\text{queue}}$	Surrogate for value per queue slot	Balances true vs. false occupancy
Volatility penalty	$\mathcal{L}_{\text{vol}}$	Limits score jumps without new evidence	Stabilizes membership of Q_t
Calibration penalty	$\mathcal{L}_{\text{cal}}$	Aligns scores with frequencies by regime	Supports thresholding and budget control
Confidence consistency	$\mathcal{L}_{\text{conf}}$	Matches $c_i(t)$ to evidence support	Enables selective abstention and escalation

Table 5 Loss components aligned with learning ranked worklists under rare deterioration events.

Action level	Description	Typical triggers	Capacity object
None	Background monitoring only	Low $\tilde{p}_i(t)$ or clear recovery	Uses no extra review slots
Monitor	Passive presence on ranked list	Moderate risk or low confidence	Driven by position in full π_t
Review	Focused chart or bedside check	$\tilde{p}_i(t)$ above review boundary, adequate $c_i(t)$	Limited by $K_t^{(r)}$
Escalate	High-intensity response / team activation	High risk and high confidence near top ranks	Limited by $K_t^{(e)}$ and alert budget

Table 6 Graded actions supported by continuous ICU risk scores and queue capacity.

improves AUROC but doubles the volatility of Q_t may be less desirable than one with slightly weaker discrimination but much more stable queue membership. Such tradeoffs are invisible unless the computational object is defined correctly.

The worklist interpretation also creates a natural connection to streaming systems. Since the active set $\mathcal{I}(t)$ changes with admissions, discharges, and clinical updates, the ranking must be recomputed or updated continuously. That means the model is part of a real-time ordering process, not a static batch predictor. Computational efficiency, incremental updates, and support for streaming calibration become central engineering issues rather than afterthoughts. The representation should therefore be designed with online score updates in mind, preserving state in a form that can be revised cheaply when new evidence arrives [8].

A further conceptual gain of this framing is that it clarifies the role of uncertainty. In a static binary classifier, uncertainty is often thought of as a property of the probability estimate. In a prioritization engine, uncertainty is also about queue position. A patient may have a moderately high score with high certainty, or a very high score with low certainty because the evidence is sparse or contradictory. Whether that patient belongs in Q_t depends not only on estimated event probability but on the opportunity cost of allocating scarce review capacity to uncertain cases. This suggests that the surveillance object should include both risk and support, with the queue rule taking both into account.

In summary, ICU deterioration surveillance should not be conceptualized as the emission of independent alarms. It should be conceptualized as the maintenance of a dynamic, evidence-conditioned ordering over the active patient population. The relevant computational object is a ranked worklist whose upper tail is consumed by humans under capacity constraints. Once that object is recognized, the next step is to formulate the queue that consumes it and the mathematical conditions under which prioritization becomes useful. That is the task of the next section.

3. Surveillance as Queue-Constrained Ranking

A ranking becomes operational only when it is embedded in a queueing process. In the ICU, this queue is not always explicit, but it is nevertheless real. There is a finite number of bedside reassessments a team can perform, a finite number of charts a

Component	Symbol / idea	Role	Impact on use
Calibrated risk	$\tilde{p}_i(t)$	Maps logits to event probabilities	Governs score scale for policies
Reliability state	$c_i(t)$	Summarizes evidence quality and support	Filters which scores can drive alerts
Stratified calibration	$T_{b(i,t)}$	Context-dependent temperature scaling	Adapts probabilities by regime
Risk–coverage curve	$\mathrm{Cov}(\tau_c)$	Tradeoff between coverage and error	Tunes how widely automation acts
Hysteresis thresholds	$(\tau_p^+, \tau_c^+), (\tau_p^-, \tau_c^-)$	Separate activation vs. release conditions	Reduces alert flicker in time

Table 7 Calibration and confidence elements for selective escalation of ICU risk.

Drift type	Description	Effect on prioritization	Monitoring signals
Patient-state drift	Change in true illness mix or acuity	Alters who should occupy top ranks	Outcome rates, label mix, survival curves
Observation-process drift	Change in measurement or documentation patterns	Distorts mapping from chart to scores	Freshness, note density, score shifts
Action-feedback drift	Model-driven changes in care and monitoring	Feedback between queues and events	Logged exposures vs. downstream actions
Queue-state drift	Systematic change in Q_t composition	Shifts which phenotypes consume slots	Embedding shifts, D_t, turnover spikes

Table 8 Categories of deployment drift and their implications for ICU worklists.

remote monitoring service can review, and a finite tolerance for alarm burden. Each review consumes time and attention. The surveillance system therefore competes with the rest of clinical work for limited service capacity. From a systems viewpoint, the task is to choose which patients enter the queue and in what order, given that queue length is bounded.

Let $N_t = |\mathcal{I}(t)|$ be the active ICU census at time t, and let K_t be the number of patients that can receive additional review or intervention-ready attention during a decision cycle. The ratio K_t/N_t is usually small. In many realistic settings, only the top few percent or top few patients can be the focus of intensified surveillance at any one time. If scores are interpreted merely through a fixed probability threshold, the system may generate more candidates than the queue can process [9]. This is a design failure, not just a thresholding inconvenience, because the entire value of a risk model is filtered through the capacity-constrained queue.

A formal queue-admission function can be defined as

$$\begin{aligned} a_i(t) &= \mathbb{I}\big(i \in Q_t\big) \\ &= \mathbb{I}\Big(\mathrm{rank}_{\pi_t}(i) \le K_t\Big), \end{aligned} \tag{5}$$

or, more flexibly, as a function of score and confidence,

$$a_i(t) = \mathbb{I}\big(s_i(t) \ge \tau_t\big)\,\mathbb{I}\big(c_i(t) \ge \xi_t\big), \tag{6}$$

with τ_t and ξ_t chosen so that the resulting queue length matches available capacity. The second form is preferable because it acknowledges that uncertainty should affect admission when review capacity is scarce. A patient with a high but weakly supported score may not deserve the same queue priority as one with the same risk and strong evidential support.

To evaluate such a system, one needs utility functions tied to queue outcomes rather than only label outcomes. Let $u_i(t)$ denote the expected utility of reviewing patient i at time t, which may depend on the probability of deterioration, the expected lead time before the event, and the value of any available intervention. Then the ideal queue would solve

$$\begin{aligned} Q_t^\star &= \arg\max_{Q \subseteq \mathcal{I}(t)} \sum_{i \in Q} u_i(t) \\ \text{s.t.} \quad & |Q| \le K_t. \end{aligned} \tag{7}$$

The challenge is that $u_i(t)$ is not directly observed. The model must approximate it from the record, and in most implementations $s_i(t)$ serves as a surrogate for $u_i(t)$. This is another reason why rank quality in the extreme upper tail matters so much. The queue consumes not the full score distribution but only the subset most likely to maximize review utility under the capacity constraint.

A simple but revealing special case is when utility is proportional to calibrated event probability. Then the optimal queue is just the top-K_t by $p_i(t; H)$. However, this assumption is often too narrow. The clinical value of reviewing a patient depends not only on event probability but on whether there is still meaningful time to intervene. A patient very likely to deteriorate in

Mechanism	Description	Purpose	Example signals
Bounded autonomy	Model prioritizes, clinicians decide	Keeps system as decision support	Tiered actions, optional escalation
Bounded adaptation	Constrained model updates vs. reference	Prevents abrupt behavioral shifts	Limit on $\|f_\theta - f_{\theta_t}\|$
Queue fairness checks	Compare access to Q_t across groups	Avoids uneven attention allocation	Subgroup queue rates and lead times
Support-aware abstention	Lower actions when far from training support	Reduces overconfident extrapolation	Large $d_i^{\text{sup}}(t)$ with low $c_i(t)$
Transparent interface	Show score, trend, and evidence basis	Aids trust and interpretability	Rank position, $\dot{s}_i(t)$, key factors

Table 9 Governance and robustness practices for continuous ICU risk prioritization.

the next ten minutes may be less improvable by additional review than a slightly less risky patient whose decline is expected later. To capture this, utility should depend on both risk and lead time [10]. Let $d_i(t)$ denote expected time to event conditional on occurrence. Then one may write

$$u_i(t) = \mathbb{E}\Big[Y_i(t,H)\, g(d_i(t)) \mid \mathcal{O}_i(t)\Big], \tag{8}$$

where g rewards actionable lead times more than extremely imminent or excessively distant ones. The resulting queue is not a pure event-probability ranking. It is a ranking by expected action value.

This observation suggests that a good risk score should encode more than binary hazard. In practice, however, many models emit only one number. The queue interpretation provides a way to recover some of the missing structure. If the score trajectory is observed over time, then rising risk itself can serve as a proxy for urgency acceleration, and the review queue can combine score level with score derivative. Define $\dot{s}_i(t)$ as a short-term slope estimate of the risk trajectory. A priority score can then be

$$\phi_i(t) = s_i(t) + \lambda_{\dot{s}}\, \dot{s}_i(t), \tag{9}$$

where $\lambda_{\dot{s}}$ controls how much recent upward momentum increases queue position. This is an example of how the surveillance object extends beyond static probabilities. The queue can exploit temporal structure in the score process itself.

Queue stability also matters. If the model causes large hour-to-hour turnover in the top K_t, clinicians may spend more attention reacting to rank changes than to the underlying patients. A useful measure is queue turnover,

$$T_t = 1 - \frac{|Q_t \cap Q_{t-1}|}{|Q_t \cup Q_{t-1}|}, \tag{10}$$

which ranges from 0 when queue membership is unchanged to 1 when it is completely different. Some turnover is appropriate because risk changes. But excessive turnover implies that the queue is unstable and likely sensitive to weak evidence or noise. Therefore a prioritization system should aim not merely for high utility in Q_t but for controlled turnover given the frequency of meaningful new information.

Queue delay provides another perspective [11]. Suppose each patient admitted to the review queue waits for service from a reviewer or escalation process. If arrivals to the queue exceed service capacity persistently, then delay grows, and the benefits of prediction erode because high-risk patients are reviewed too late. Let λ_t be the effective arrival rate into Q_t and μ_t the service rate. Even a simple $M/M/1$-like approximation warns that expected waiting time grows sharply as λ_t/μ_t approaches one. A model that increases queue arrivals without sufficient enrichment of true utility can therefore make the system worse by saturating reviewer capacity. This is a direct argument for evaluating not only alert precision but queue load induced by the score distribution.

One can formalize queue pressure through a cost term

$$C_t^{\text{queue}} = \zeta \Big(\max(0, \lambda_t - \mu_t)\Big)^2, \tag{11}$$

which penalizes sustained overflow. Training or threshold selection should, implicitly or explicitly, account for such pressure. Otherwise a model may optimize discrimination in a way that is perfectly sensible statistically but inconsistent with the service process it creates. In this sense, queueing is not a deployment add-on. It is part of the model's objective environment.

The top-K_t queue also illuminates why calibration in the tail matters more than average calibration. If the model overestimates probabilities broadly across the upper decile, then many patients appear nearly tied for queue admission, making the top K_t sensitive to noise. If the model sharply calibrates the extreme tail, then queue membership becomes more interpretable and less arbitrary. Thus the queue interpretation gives a concrete operational meaning to tail calibration: it governs how decisively the system can separate the few patients who can be reviewed now from the many who cannot.

Finally, queue-constrained ranking suggests alternative evaluation metrics beyond ordinary discrimination. Event capture at fixed queue size, median lead time among queued true positives, and expected utility per queue slot are more directly aligned with deployment than aggregate AUROC. So too are stability metrics such as turnover and score volatility among the top ranks [12]. A model that slightly reduces AUROC but substantially improves utility per queue slot or reduces unstable queue

churn may be a better surveillance engine.

The key idea of this section is that ICU forecasting gains its meaning only when a queue consumes its output. The queue forces a confrontation with capacity, turnover, and opportunity cost. Under that view, a score is not merely a probability estimate but a claim about where a patient belongs relative to others competing for limited review. The next section asks how that claim should evolve over time when patient state, interventions, and observations change, leading to a dynamic model of priority flow rather than static risk assignment.

4. Dynamic Patient State and Priority Flow

If the surveillance system is a queue-constrained prioritizer, then the score for each patient should be understood as a dynamic state variable rather than as an independent sequence of hourly estimates. The patient's position in the worklist evolves as new evidence arrives, as interventions alter the underlying trajectory, and as the patient's own recent priority history constrains what changes are plausible. A well-designed system should therefore propagate risk through time in a way that is both responsive and smooth. Responsiveness matters because sudden deterioration must be reflected quickly. Smoothness matters because worklists cannot be useful if small observational perturbations cause large oscillations in priority.

Let $z_i(t)$ denote a latent dynamical state summarizing the patient's current trajectory with respect to deterioration risk, intervention context, and evidential support. The observed chart contributes evidence $o_i(t)$, while treatment actions $u_i(t)$ modify both the evolution of $z_i(t)$ and the interpretation of future observations. A generic stochastic evolution can be written as

$$z_i(t+\Delta t) = F\big(z_i(t), o_i(t), u_i(t)\big) + \epsilon_i(t), \tag{12}$$

with score

$$s_i(t) = G\big(z_i(t)\big). \tag{13}$$

This looks like a standard state-space model, but the present application imposes special structure. The score process $s_i(t)$ is not just a latent hazard estimate. It is the variable that determines queue position. Therefore, the dynamics of $z_i(t)$ should be studied partly in terms of how they affect rank movement.

A convenient concept is priority flow, the hour-to-hour change in the score of a patient relative to the score field of the active unit. If $\bar{s}(t)$ is the cross-sectional mean score and $\sigma_s(t)$ the cross-sectional standard deviation, define standardized priority

$$r_i(t) = \frac{s_i(t) - \bar{s}(t)}{\sigma_s(t) + \varepsilon}. \tag{14}$$

The quantity $r_i(t)$ represents queue-relative urgency rather than absolute risk [13]. A patient whose absolute score remains stable may still move upward in priority if others stabilize and vice versa. The surveillance system therefore cares about the dynamics of $r_i(t)$, not only $s_i(t)$. This is especially important in busy units where capacity is tied to relative rather than absolute risk.

Priority flow can then be decomposed into a patient-specific drift term and a cross-sectional normalization term:

$$\dot{r}_i(t) = \frac{\dot{s}_i(t)}{\sigma_s(t) + \varepsilon} - \frac{s_i(t) - \bar{s}(t)}{(\sigma_s(t) + \varepsilon)^2}\dot{\sigma}_s(t) - \frac{\dot{\bar{s}}(t)}{\sigma_s(t) + \varepsilon}. \tag{15}$$

This decomposition highlights a subtle point. Rank movement is influenced not just by changes in one patient's state but by changes across the whole unit. A model that produces globally inflated scores during certain times of day or certain documentation regimes can distort relative priority without improving patient-specific prediction. Therefore, state propagation should be considered in a population context as well as an individual one.

A useful design principle is inertial prioritization. The idea is that large changes in score should require proportionally strong new evidence unless the patient is already near an unstable boundary where deterioration is plausible. This can be encoded by defining the state update as a gated combination of prior state and evidence innovation:

$$\begin{aligned} z_i(t+\Delta t) &= \big(1 - \kappa_i(t)\big)\, z_i(t) + \kappa_i(t)\, \Psi\big(o_i(t), u_i(t)\big), \\ \kappa_i(t) &= \sigma\Big(w_\kappa^\top \big[o_i(t), u_i(t), c_i(t)\big]\Big), \end{aligned} \tag{16}$$

where $c_i(t)$ is confidence or evidential support. When new evidence is weak, stale, or contradictory, $\kappa_i(t)$ remains small and the score changes slowly. When strong coherent evidence arrives, $\kappa_i(t)$ increases and the score responds sharply. This is preferable to unconditional smoothing because it allows responsiveness to clinically meaningful change without encouraging noise-driven queue churn.

Interventions complicate the dynamics because they alter the latent system rather than merely revealing it. A vasopressor start, for example, can indicate that the patient's underlying hemodynamic state was unstable, but after the start the observed blood pressure trajectory is partly controlled by treatment. In the dynamic model, this means the transition F depends on action history in a nontrivial way. If the model ignores $u_i(t)$, it may misread treated stability as genuine recovery [14]. If it overuses $u_i(t)$, it may assign high priority solely because a treatment was initiated, even when the patient is already receiving maximal attention. A useful compromise is to allow interventions to enter both the transition and an action-saturation term that discounts queue utility when further review would add little:

$$s_i^{\mathrm{eff}}(t) = s_i(t) - \omega\, h_{\mathrm{sat}}\big(u_i(t)\big). \tag{17}$$

The effective score $s_i^{\mathrm{eff}}(t)$ is what enters the review queue. A heavily treated patient may remain high risk, but if the marginal value of further surveillance is small because the clinical team is already fully engaged, the effective queue priority should be lower than the raw risk alone would suggest.

Temporal smoothness can be expressed as a regularization target during learning. Suppose the model emits logits $\eta_i(t)$ at each hour. One can penalize large unnecessary changes with

$$\mathcal{L}_{\mathrm{smooth}} = \sum_{i,t} \omega_i(t)\, \big|\eta_i(t) - \eta_i(t-1)\big|, \tag{18}$$

where $\omega_i(t)$ is small when meaningful new evidence arrives and large otherwise. This is a queue-oriented smoothness penalty rather than a generic denoising term. It encodes the idea that score changes should be justified by evidence changes. Such penalties are especially useful in ICU data, where repeated windows are highly overlapping and yet small recording artifacts can induce undesirable score volatility.

One can also model priority flow explicitly through a secondary head predicting future score motion. Let $\Delta_i(t,\delta) = s_i(t+\delta) - s_i(t)$ be the future change over a short interval. A multi-task model can predict both current risk and expected near-term score acceleration:

$$\hat{\Delta}_i(t,\delta) = H\big(z_i(t)\big). \tag{19}$$

This gives the system a built-in estimate of rising urgency, which can be useful in queue policy. Patients with moderate current risk but strong positive $\hat{\Delta}$ may deserve earlier review than patients with slightly higher static risk but flat or declining trajectories. In other words, the queue can become anticipatory rather than merely reactive.

A further refinement is to treat the priority state as partially observable and to maintain uncertainty over $z_i(t)$. Let $q_i(t)$ denote confidence in the current score. Then the system should not only rank by $s_i(t)$ but modulate queue admission by the pair $(s_i(t), q_i(t))$. As discussed later, this supports selective escalation. At the state-dynamics level, uncertainty should widen when evidence is sparse and contract when new coherent evidence arrives [15]. Thus one may propagate a covariance-like term $\Sigma_i(t)$ alongside $z_i(t)$, with update

$$\Sigma_i(t+\Delta t) = A_i(t)\Sigma_i(t)A_i(t)^\top + Q_i(t) - K_i(t)R_i(t)K_i(t)^\top, \tag{20}$$

in an approximate filter-style analogy, where $Q_i(t)$ increases with process uncertainty and $R_i(t)$ decreases with evidence quality. Even if implemented heuristically rather than with strict linear-Gaussian assumptions, this perspective is valuable: queue decisions should know whether a high score is well determined or poorly determined by the available record.

Finally, dynamic priority flow suggests that the surveillance engine should not be evaluated only on instantaneous outcomes. It should also be evaluated on the coherence of risk trajectories. Does the score typically rise before events, plateau appropriately under ongoing instability, and decay when recovery is well supported? Does queue membership persist across meaningful intervals or flicker with minor chart changes? These are trajectory-level properties. They are closely tied to the dynamical model and weakly captured by static metrics.

The principal contribution of this section is to replace the fiction of independent hourly predictions with the reality of evolving priority states. In a surveillance system, each patient carries not only a current risk but a motion through the worklist. That motion should be shaped by evidence, moderated by confidence, adjusted for intervention saturation, and smooth enough to be operationally useful. The next section turns to how such a system should be learned when positive events are rare and when utility is concentrated in the extreme upper tail rather than across the full set of patient-time windows.

5. Learning Ranked Worklists Under Rare Events

Once the surveillance system is understood as a dynamic prioritizer, the shortcomings of standard learning objectives become more visible. Binary cross-entropy is proper for probability estimation under independent examples and stable class balance, but ICU deterioration windows are neither independent nor balanced. Positive windows are rare, adjacent windows are heavily dependent, and the practical value of the model is concentrated in a small fraction of high-score cases. In such a setting, learning should be guided by the geometry of ranked worklists rather than by average fit alone.

Let $y_i(t) \in \{0,1\}$ denote the deterioration label within the chosen horizon, and let $\hat{p}_i(t) = \sigma(\eta_i(t))$ be the predicted probability. A class-imbalance-aware starting point is focal loss:

$$\begin{aligned}\mathcal{L}_{\text{focal}} = &-\sum_{i,t} \alpha y_i(t)(1-\hat{p}_i(t))^\gamma \log \hat{p}_i(t) \\ &-\sum_{i,t}(1-\alpha)(1-y_i(t))\hat{p}_i(t)^\gamma \log(1-\hat{p}_i(t)).\end{aligned} \tag{21}$$

This improves retrieval of rare positives, but it does not explicitly reflect the queue objective. A model can optimize focal loss while still producing an upper tail too broad or too unstable to support a useful top-K_t worklist. Therefore one needs ranking-sensitive terms aligned with queue use [16].

A natural addition is a pairwise or listwise ranking loss over matched strata. Let $\mathcal{P}$ be a set of positive-negative pairs sampled within comparable ICU-age and intervention contexts. A logistic ranking term is

$$\mathcal{L}_{\text{rank}} = \sum_{(a,b)\in\mathcal{P}} \log\Big(1+\exp(-(\eta_a-\eta_b))\Big). \tag{22}$$

However, even this does not fully reflect the queue. Pairwise ranking is agnostic to how much of the positive mass is pushed into the top few slots. For a queue-limited surveillance engine, the objective should encourage concentration of utility into the very top of the ranked list. One way to express this is with a differentiable tail-concentration term. Let $w_i(t)$ denote a soft top-K membership weight derived from the current minibatch scores. Then define

$$\mathcal{L}_{\text{tail}} = -\sum_{i,t} w_i(t)\, y_i(t), \tag{23}$$

where $w_i(t)$ approximates whether a case lies in the batch's high-score tail. This encourages true positives to migrate toward the top queue rather than merely above negatives in a broad average sense.

Because worklists are dynamic, training should also penalize queue-irrelevant volatility. Suppose $\eta_i(t)$ is the current logit and $\eta_i(t-1)$ the previous one. A volatility term can be weighted by novelty of new evidence $n_i(t)$, defined from the amount and relevance of information added since the previous step. Then

$$\mathcal{L}_{\text{vol}} = \sum_{i,t}(1-n_i(t))\,\big(\eta_i(t)-\eta_i(t-1)\big)^2. \tag{24}$$

This does not demand smoothness universally. It demands smoothness when the record has changed little. By doing so, it regularizes the worklist toward stability without blunting responsiveness to genuinely new clinical information.

Another issue is that positives are heterogeneous in action value. A window one hour before deterioration and a window twenty-three hours before deterioration may share the same binary label. Yet from a queue perspective their utility differs. To encode this, one can introduce a lead-time-sensitive target [17]. Let $d_i(t)$ be time until the next deterioration event when one occurs within the horizon. Define a soft urgency weight

$$u_i(t) = y_i(t)\,\exp\Big(-\frac{d_i(t)}{\tau_u}\Big), \tag{25}$$

for some scale τ_u. One may then train an auxiliary head to predict $u_i(t)$ or use $u_i(t)$ to reweight the positive term in the loss. This shapes the score surface so that higher positions in the queue are preferentially occupied by windows with not just positive labels, but high immediate action value.

Dependence among repeated windows creates another problem. A single event can generate many positive windows, causing the model to overfit to long plateaus of pre-event similarity. To reduce this, the learning process should not weight every window equally. One approach is novelty-aware sampling. If $n_i(t)$ measures how much the state or evidence changed since the previous window, then the training probability for (i,t) can be set proportional to both class importance and novelty:

$$p_{\text{sample}}(i,t) \propto \omega(y_i(t))\,(\epsilon + n_i(t)). \tag{26}$$

This favors windows that contribute new information rather than repeated near-duplicates. The effect is to sharpen the model's sensitivity to transitions in priority rather than to reward persistence alone.

The queue interpretation also argues for loss terms that approximate expected utility under capacity constraints. Let K be a chosen queue size in the training surrogate, and let $\tilde{Q}$ be a soft top-K set induced by batch scores. A utility-driven term can be written as

$$\mathcal{L}_{\text{queue}} = -\sum_{(i,t)\in\tilde{Q}} \Big(u_i(t) - \lambda_{\text{fp}}(1-y_i(t))\Big), \tag{27}$$

where λ_{fp} prices queue slots occupied by false positives. This directly pushes the optimization toward a more valuable upper tail. Although the soft queue in minibatches is only an approximation to the true unit-level queue, it better reflects deployment than objectives indifferent to capacity.

Calibration complicates learning because most ranking-enhancing strategies distort raw probability scale. This is acceptable so long as calibration is recovered later, but it is wise to include at least a modest penalty discouraging extreme misalignment of scores and observed frequencies in comparable support regimes. Let B_m denote bins over score and confidence strata [18]. A calibration penalty is

$$\mathcal{L}_{\text{cal}} = \sum_m \omega_m \left(\frac{1}{|B_m|} \sum_{(i,t)\in B_m} \hat{p}_i(t) - \frac{1}{|B_m|} \sum_{(i,t)\in B_m} y_i(t) \right)^2. \tag{28}$$

This does not replace post-hoc calibration, but it prevents the optimization from producing a score field so distorted that later calibration becomes unstable or context insensitive.

Because the deployed worklist depends on support as well as risk, the learning objective should include consistency between predicted uncertainty and evidence quality. Let $c_i(t)$ be the confidence estimate. If evidence is sparse, stale, or contradictory, then the model should learn not only to moderate score changes but also to lower confidence. A simple support-consistency penalty is

$$\mathcal{L}_{\text{conf}} = \sum_{i,t} \Big(c_i(t) - \rho_i(t)\Big)^2, \tag{29}$$

where $\rho_i(t)$ is a target or proxy derived from evidence freshness, missingness structure, and support-set familiarity. This helps the later queue-admission rule treat uncertainty explicitly instead of forcing all ambiguity into the risk logit.

Putting these components together yields

$$\begin{aligned}\mathcal{L} = {} & \mathcal{L}_{\text{focal}} + \lambda_r \mathcal{L}_{\text{rank}} + \lambda_t \mathcal{L}_{\text{tail}} \\ & + \lambda_q \mathcal{L}_{\text{queue}} + \lambda_v \mathcal{L}_{\text{vol}} + \lambda_c \mathcal{L}_{\text{cal}} \\ & + \lambda_f \mathcal{L}_{\text{conf}} + \lambda_w \|\Theta\|_2^2.\end{aligned} \tag{30}$$

The point is not that every implementation must use all these terms exactly. The point is that learning should respect the actual object being optimized: a ranked, capacity-constrained worklist with stability and confidence requirements.

The value of this formulation becomes apparent when considering what kinds of errors it discourages. It discourages broad inflation of the high-score region because queue-based penalties make false occupancy of top slots costly. It discourages brittle hour-to-hour reordering because volatility penalties respond to unchanged evidence. It encourages concentration of true urgent cases in the top queue rather than only pairwise separation over all windows. It supports confidence estimates that can be consumed by selective action. In short, it makes the loss function answer to the surveillance system rather than only to the statistical label.

The next challenge is how these learned scores should be turned into actual decisions. Since the queue is small and the upper tail is precious, one cannot simply present all probabilities to clinicians as if they were equally actionable. Selective escalation, calibration, and explicit pricing of attention become central [19]. That is the focus of the next section.

6. Selective Escalation, Calibration, and the Economics of Attention

A score becomes clinically meaningful only when it is connected to an action policy. In the ICU, such policies face two constraints simultaneously. First, deterioration is rare enough that indiscriminate alerting quickly overwhelms reviewers. Second, missing a truly actionable case can be costly. Therefore the system must not only estimate risk. It must allocate scarce attention economically. This is why a continuous risk score is better understood as a prioritization signal than as a binary alarm. The

score should support graded action under uncertainty rather than force a brittle yes-or-no decision at a single threshold.

Suppose the model emits a calibrated risk $\tilde{p}_i(t)$ and a confidence $c_i(t)$. Confidence summarizes evidence quality, support-set familiarity, and perhaps disagreement across modalities or state components. The simplest policy already differs from ordinary thresholding:

$$a_i(t) = \mathbb{I}\big(\tilde{p}_i(t) \geq \tau_p\big)\, \mathbb{I}\big(c_i(t) \geq \tau_c\big). \tag{31}$$

This means that active escalation is reserved for patients who are both sufficiently high risk and sufficiently well supported. A patient with high risk but low confidence is not silently treated as low risk. Instead, that patient can enter a secondary review category, remain high on a dashboard without alerting, or trigger a request for additional evidence. This is a more faithful expression of what the model actually knows.

Why is this preferable to simply changing τ_p? Because risk and confidence are not the same quantity. A high but weakly supported score indicates suspicion under uncertainty, not reassurance. Lowering risk because evidence is sparse confounds epistemic uncertainty with benignity. Conversely, keeping risk high but requiring confidence for escalation allows the system to express two different kinds of statements: one about the possibility of deterioration, another about the reliability of the supporting record [20]. In operational settings this distinction is often exactly what clinicians need.

The queue interpretation suggests a multi-level action set $\mathcal{A} = \{\text{none}, \text{monitor}, \text{review}, \text{escalate}\}$. A utility function can then be defined as

$$U_i(a,t) = \mathbb{E}\Big[R_i(a,t) - C_i(a,t) \mid \mathcal{O}_i(t)\Big], \tag{32}$$

where R_i is expected benefit of acting at level a and C_i is the attention and workflow cost. The optimal action is

$$a_i^\star(t) = \arg\max_{a \in \mathcal{A}} U_i(a,t). \tag{33}$$

In practice, $\tilde{p}_i(t)$ and $c_i(t)$ are sufficient statistics for approximating U_i, but the framing matters because it makes clear that active alerting is just one possible action. In a tiered system, a large fraction of patients may only require passive ranking or continued monitoring, while only a few enter higher-cost escalation states.

Capacity constraints return here with full force. Let $K_t^{(r)}$ be the number of chart reviews the team can perform and $K_t^{(e)}$ the number of active escalations it can manage. The policy should satisfy

$$\begin{aligned} \sum_{i \in \mathcal{I}(t)} \mathbb{I}\big(a_i(t) = \text{review}\big) &\leq K_t^{(r)}, \\ \sum_{i \in \mathcal{I}(t)} \mathbb{I}\big(a_i(t) = \text{escalate}\big) &\leq K_t^{(e)}. \end{aligned} \tag{34}$$

This immediately shows why probability calibration matters. If the upper tail is not well shaped, then small changes in thresholds can greatly alter how many patients meet action criteria, making the system operationally unstable. Good calibration in the high-confidence upper tail translates into more predictable queue occupancy.

A convenient deployment strategy is to separate ranking from thresholding. First, the system computes a ranked list by $s_i(t)$ or $\tilde{p}_i(t)$. Then confidence and capacity determine how far down that list the unit can act. This yields dynamic thresholds based on current census and staffing rather than fixed scores alone. If the census is unusually sick or staffing is low, the boundary for active escalation rises. If the census is stable or staffing expands, more of the ranked list can receive attention. Such a policy is often more realistic than a fixed threshold carried unchanged across contexts.

Economic interpretation helps clarify false-alarm burden. A false alert is often treated as a statistical cost, but in surveillance it is better thought of as opportunity cost measured in queue slots and clinician time. If one false positive occupies a review slot, then another patient is not reviewed. Let c_{slot} denote the expected value of a review slot. Then the utility of assigning patient i to review becomes [21]

$$U_i(\text{review}, t) = \tilde{p}_i(t)\, g_i(t) - c_{\text{slot}} - c_{\text{noise}}\big(i,t\big), \tag{35}$$

where $g_i(t)$ is the expected intervention value conditional on deterioration and c_{noise} is any added cost from alarm fatigue or repeated low-yield alerts. This formulation explains why the top-K queue is such a useful abstraction: each slot has value, and assigning a patient to it should reflect expected benefit minus slot cost.

Lead time can also be priced explicitly. If $d_i(t)$ is expected time to event conditional on occurrence, then a utility-modulated priority can be

$$\psi_i(t) = \tilde{p}_i(t)\, h\big(d_i(t)\big)\, c_i(t), \tag{36}$$

where h rewards actionable lead times. Cases with very short expected lead time may still merit attention, but perhaps more for immediate escalation than for standard review. Cases with moderate lead time and high confidence may be ideal for early preventive intervention. By encoding such distinctions into $\psi_i(t)$, the queue becomes more than a probability sorter. It becomes an action-value ranking.

Calibration itself should be treated as conditional on confidence and perhaps on subgroup or unit context. A stratified temperature scaling approach defines

$$\tilde{p}_i(t) = \sigma\Big(\eta_i(t) / T_{b(i,t)}\Big), \tag{37}$$

where $b(i,t)$ indexes reliability strata or unit-level contexts. This preserves within-stratum ranking and allows the numerical scale of scores to adapt across evidence regimes. Such conditioning is especially important if the same nominal score can arise from radically different observational support patterns. Without it, the action policy may overcommit to fragile scores and underuse stable ones.

Selective escalation also offers a principled alternative to forcing complete automation coverage. Some windows should

not produce strong action recommendations because the evidence is too poor or too unfamiliar. Coverage can be summarized by

$$\mathrm{Cov}(\tau_c) = \Pr\left(c_i(t) \geq \tau_c\right), \tag{38}$$

and the error among covered cases can be tracked as a risk-coverage curve. In many safety-sensitive applications, it is preferable to accept lower coverage in exchange for more trustworthy high-impact actions [22]. In the ICU, where every alert competes with human workload, this tradeoff is particularly salient. A model that insists on acting everywhere may create more burden than benefit.

Temporal behavior of actions matters too. If a patient repeatedly crosses and recrosses an escalation threshold, the system becomes noisy and hard to trust. Hysteresis provides a remedy. Let (τ_p^+, τ_c^+) be activation thresholds and (τ_p^-, τ_c^-) be lower release thresholds. Then once a patient enters an active state, the system does not release that state until both risk and confidence fall below the lower pair. This dampens oscillation without muting true recovery or new decline. Hysteresis is especially effective when coupled with the smooth state dynamics discussed earlier.

The broader point of this section is that a continuous risk score acquires meaning through an action economy. Scarce attention, queue slots, confidence, and lead time determine how scores should be consumed. A model that is excellent statistically but indifferent to this economy can generate poor action policies. A model that supports selective escalation, stratified calibration, and capacity-aware ranking is far more likely to remain useful when translated into real workflow. The final major question is how such a system behaves under drift and how it should be governed once deployed continuously. That is the focus of the next section.

7. Operational Drift, Robustness, and Human Oversight

No ICU surveillance system operates in a static environment. Patient populations change across seasons and services. Documentation practice changes after new templates or policy updates. Monitoring intensity changes with staffing and unit culture. Once the model is deployed, clinician responses to its outputs can themselves alter the downstream data distribution [23]. For a system whose value depends on ranked worklists and high-tail calibration, such drift is not a peripheral concern. It strikes at the heart of what the score means and how the queue behaves. Therefore robustness and governance are integral parts of the computational design.

A useful distinction is between three kinds of drift. The first is patient-state drift, in which the distribution of underlying clinical trajectories changes. The second is observation-process drift, in which measurement frequency, note density, or intervention practice changes without a commensurate change in latent disease states. The third is action-feedback drift, in which deployment of the model changes behavior in ways that subsequently alter both outcomes and observations. These forms of drift affect prioritization differently. Patient-state drift may change who should truly be near the top of the queue. Observation-process drift may change who appears to deserve a top slot under the learned mapping. Action-feedback drift may make retrospective evaluation misleading because the model begins influencing the very events it predicts.

Because prioritization relies heavily on the upper tail, drift detection should focus not only on global score distributions but on queue-relevant regions. Let Z_t denote a latent representation of active patients and let Q_t be the current review queue. A simple drift measure compares reference and current queue-state distributions:

$$D_t = \left\| \frac{1}{|Q_t|} \sum_{i \in Q_t} z_i(t) - \mu_Q^{\mathrm{ref}} \right\|_2, \tag{39}$$

where μ_Q^{ref} is the reference mean queue embedding from a stable development period. If D_t rises substantially, it suggests that the kinds of cases entering the queue have changed, even if the overall score distribution has not. This is often more informative than whole-population drift for operational surveillance because the queue, not the full census, is where human attention is spent.

One should also monitor queue turnover and occupancy relative to capacity. A sudden increase in T_t, the turnover metric defined earlier, may indicate that the model is responding too strongly to a new documentation pattern or a change in observation frequency. Similarly, a sustained rise in queue occupancy above intended levels signals score inflation or calibration drift [24]. Since the worklist is a queue-facing object, these operational statistics are first-class indicators of model health.

Support-aware confidence helps manage drift in a principled way. If the model includes a support or familiarity estimate $c_i(t)$, then it can abstain or downgrade actions when states move away from training support. A latent support distance can be defined as

$$d_i^{\mathrm{sup}}(t) = \min_k \|z_i(t) - \mu_k\|_2, \tag{40}$$

where $\{\mu_k\}$ are cluster centroids or prototypes from training. Confidence can then fall as $d_i^{\mathrm{sup}}(t)$ increases. This does not guarantee correctness, but it does prevent strong automation in clearly unfamiliar regions. In a prioritization system, this can prevent the queue from filling with cases whose high scores arise from extrapolation rather than supported inference.

Human oversight should be shaped by the same logic. A continuous ranking system need not always translate directly into an automated alert. Instead, low-confidence high-risk patients can be marked for discretionary review or for evidence acquisition rather than for immediate escalation. This preserves the value of the ranking while acknowledging uncertainty. It also allows the unit to learn from model behavior in unfamiliar regions before granting stronger automation authority there.

Governance becomes especially important when the model's predictions can influence care and thereby alter outcomes. Suppose high-ranked patients receive more monitoring and faster intervention. Some events may then be prevented or delayed. Naive retrospective evaluation after deployment might count

such cases as false positives even though the prediction contributed to harm reduction. This is not a reason to stop evaluating the model. It is a reason to log model exposure and follow-on actions so that feedback effects can be distinguished from raw predictive failure. In systems terms, the model and the ICU together form a closed loop after deployment. Closed-loop evaluation is more difficult, but it is also more honest.

A cautious adaptation strategy is therefore preferable to aggressive online retraining [25]. Calibration layers, queue thresholds, and confidence mappings can usually be updated more safely and more often than the core state model. If deeper retraining is needed, it should be constrained so that the ranking process does not change abruptly without careful review. One may formalize this with a bounded-update rule

$$\theta^{\star}_{t+1} = \arg\min_{\theta} \mathcal{L}_{\text{new}}(\theta)$$
$$\text{s.t.} \quad \mathbb{E}\left[\|f_{\theta}(\mathcal{O}) - f_{\theta_t}(\mathcal{O})\|_2\right] \le \epsilon \tag{41}$$

on a protected reference set. The point is not to freeze the model. It is to preserve behavioral continuity while allowing improvement. In a queueing system, abrupt score-scale changes can wreak havoc on admission and review policies even if standard metrics improve.

Subgroup robustness should also be analyzed through the lens of queue access. It is possible for a model to have similar overall discrimination across subgroups while offering different queue coverage or confidence levels. For example, units with sparser documentation may yield lower confidence, causing fewer patients from those units to enter active review despite comparable underlying risk. Such disparities are more visible when the surveillance system is evaluated in terms of queue membership and confidence-conditioned action rather than only pooled AUROC. Since the deployed object is a worklist, fairness-like concerns should be framed partly as equitable access to queue slots and review attention.

The user interface is another part of governance. A ranked list with no context is easy to implement but hard to trust. A better interface indicates why a patient is high in the queue, whether the score is rising or stable, and how confident the system is given the available evidence. This should be presented as evidence tracing rather than causal explanation. The system can say that a patient is high priority because of sustained hypotension, increasing oxygen requirement, and recent intervention context, with high or moderate confidence. Such summaries help clinicians use the queue appropriately without overinterpreting the model as omniscient.

Perhaps the most important governance principle is bounded autonomy [26]. The system should prioritize and suggest; it should not silently replace clinical judgment. That principle is not an admission of weakness. It is the rational design response to partial observability, shifting workflows, and closed-loop feedback. A prioritization engine embedded in a complex human system should act as a continuously updated aid to allocation, not as an unquestionable arbiter. Keeping the model within that role makes it easier to maintain trust, detect failure, and adapt safely.

The central message of this section is that robustness for ICU surveillance is not simply out-of-sample AUROC. It is the ability of a ranked, capacity-constrained worklist to retain meaning under changing data, changing behavior, and changing policy. Drift monitoring, support-aware confidence, bounded adaptation, and human oversight are therefore not afterthoughts. They are part of the computational specification of a reliable prioritization system.

8. Conclusion

ICU deterioration modeling is most usefully understood not as the generation of binary alarms, but as the maintenance of a continuous, queue-facing prioritization signal over a partially observed and dynamically changing patient population. Once the output is viewed this way, the central design questions shift. The model must still estimate future risk, but it must also shape a compact and stable upper tail, respect constraints on review capacity, distinguish strong evidence from weak evidence, and support graded actions rather than all-or-nothing thresholding. In that setting, ranking, calibration, uncertainty, and workload are no longer separate concerns. They are coupled properties of one surveillance system.

This paper developed that view by redefining the computational object of ICU prediction as a dynamic ranked worklist, then formulating surveillance as a queue-constrained ranking process. It proposed a dynamical account of patient priority as a flow variable, argued for learning objectives that optimize the geometry of the upper tail rather than only average discrimination, and showed how selective escalation and reliability-aware calibration can turn continuous scores into more meaningful actions. It also emphasized that drift, feedback, and subgroup access must be judged partly through queue behavior rather than only through aggregate predictive metrics.

Building predictors as prioritization engines for scarce human attention. A model that ranks slightly better but destabilizes the worklist may be worse than one that is modestly less discriminative but produces a coherent, confidence-aware queue. By making the queue the central object rather than the hidden consumer of a score, one can design, train, and evaluate deterioration systems in a way that is more faithful to operational reality and therefore more likely to yield durable clinical value [27].

Acknowledgments

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

References

[1] S. Sonsilphong, N. Arch-int, S. Arch-int, and C. Pattarapongsin, “A semantic interoperability approach to healthcare data: Resolving data-level conflicts,” *Expert Systems*, vol. 33, pp. 531–547, 7 2016.

[2] R. Gold, T. E. Burdick, H. Angier, L. S. Wallace, C. C. Nelson, S. Likumahuwa-Ackman, A. Sumic, and J. E. DeVoe, “Improve synergy between health information exchange and electronic health records to increase rates

of continuously insured patients.," *EGEMS (Washington, DC)*, vol. 3, pp. 1158–1158, 8 2015.

[3] S. Bakken, "What can you do with an electronic health record?," *Journal of the American Medical Informatics Association : JAMIA*, vol. 29, pp. 751–752, 4 2022.

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

[5] A. S. Malhan, K. Sadeghi-R, R. Pavur, and L. Pelton, "Healthcare information management and operational cost performance: empirical evidence.," *The European journal of health economics : HEPAC : health economics in prevention and care*, vol. 25, pp. 963–977, 11 2023.

[6] P. P. Ray, B. Chowhan, N. Kumar, and A. Almogren, "Biothr: Electronic health record servicing scheme in iot-blockchain ecosystem," *IEEE Internet of Things Journal*, vol. 8, pp. 10857–10872, 7 2021.

[7] S. Palojoki, T. Pajunen, K. Saranto, and L. Lehtonen, "Electronic health record-related safety concerns: A cross-sectional survey of electronic health record users.," *JMIR medical informatics*, vol. 4, pp. 92–102, 5 2016.

[8] K. S. Kaswan, L. Gaur, J. S. Dhatterwal, and R. Kumar, *AI-Based Natural Language Processing for the Generation of Meaningful Information Electronic Health Record (EHR) Data*, pp. 41–86. CRC Press, 8 2021.

[9] P. Plastiras and D. O'Sullivan, "Combining ontologies and open standards to derive a middle layer information model for interoperability of personal and electronic health records," *Journal of medical systems*, vol. 41, pp. 195–195, 10 2017.

[10] Y. Yamada, "The electronic health record as a primary source of clinical phenotype for genetic epidemiological studies," *Genomic medicine*, vol. 2, pp. 5–5, 7 2008.

[11] M. Phelan, N. A. Bhavsar, and B. A. Goldstein, "Illustrating informed presence bias in electronic health records data: How patient interactions with a health system can impact inference.," *EGEMS (Washington, DC)*, vol. 5, pp. 22–22, 12 2017.

[12] R. Eikstadt, H. Desmond, C. Lindner, L. Y. Chen, C. Courtlandt, S. F. Massengill, E. S. Kamil, R. A. Lafayette, A. Pesenson, M. Elliott, P. E. Gipson, and D. S. Gipson, "The development and use of an ehr-linked database for glomerular disease research and quality initiatives," *Glomerular diseases*, vol. 1, pp. 173–179, 7 2021.

[13] G. Verma, "Blockchain-based privacy preservation framework for healthcare data in cloud environment," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 36, pp. 147–160, 11 2022.

[14] S. Ge, R. J. Beechinor, C. P. Hornik, J. F. Standing, K. O. Zimmerman, M. Cohen-Wolkowiez, M. M. Laughon, R. H. Clark, and D. Gonzalez, "External evaluation of a gentamicin infant population pharmacokinetic model using data from a national electronic health record database.," *Antimicrobial agents and chemotherapy*, vol. 62, 8 2018.

[15] J. Perloff and S. Sobul, "Use of electronic health record systems in accountable care organizations.," *The American journal of managed care*, vol. 28, pp. e31–e34, 1 2022.

[16] G. T. Lapham, T. E. Matson, D. S. Carrell, J. F. Bobb, C. Luce, M. M. Oliver, U. E. Ghitza, C. Hsu, K. C. Browne, I. A. Binswanger, C. I. Campbell, A. J. Saxon, R. Vandrey, G. L. Schauer, R. L. Pacula, M. A. Horberg, S. R. Bailey, E. A. McClure, and K. A. Bradley, "Comparison of medical cannabis use reported on a confidential survey vs documented in the electronic health record among primary care patients.," *JAMA network open*, vol. 5, pp. e2211677–e2211677, 5 2022.

[17] Z. M. Nouns, S. Montagne, and S. Huwendiek, "Losing connectivity when using ehrs: a technological or an educational problem?," *Medical education*, vol. 49, pp. 449–451, 4 2015.

[18] J. K. Bower, S. Patel, J. E. Rudy, and A. S. Felix, "Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: Finding the signal through the noise.," *Current epidemiology reports*, vol. 4, pp. 346–352, 11 2017.

[19] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 25, pp. 1419–1428, 6 2018.

[20] V. M. Vemulakonda, R. A. Bush, and M. G. Kahn, ""minimally invasive research?" use of the electronic health record to facilitate research in pediatric urology.," *Journal of pediatric urology*, vol. 14, pp. 374–381, 6 2018.

[21] J. Espinoza, P. Shah, and J. K. Raymond, "Integrating continuous glucose monitor data directly into the electronic health record: Proof of concept.," *Diabetes technology & therapeutics*, vol. 22, pp. 570–576, 7 2020.

[22] S. Barron and S. Manhas, "Electronic health record (ehr) projects in canada: participation options for canadian health librarians," *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada*, vol. 32, pp. 137–143, 7 2014.

[23] J. Chen, X. Yin, and J. Ning, "A fine-grained and secure health data sharing scheme based on blockchain," *Transactions on Emerging Telecommunications Technologies*, vol. 33, 4 2022.

[24] D. Kimovski, S. Ristov, and R. Prodan, "Decentralized machine learning for intelligent health care systems on the computing continuum," 7 2022.

[25] L. M. Kern and R. Kaushal, "Electronic health records and the increasing complexity of medical practice," *Journal of general internal medicine*, vol. 28, pp. 1392–1392, 7 2013.

[26] J. M. Tan, J. O. Wasey, O. Nelson, V. Tam, A. F. Simpao, and J. A. Gálvez, "Leveraging electronic health records and administrative datasets to understand social determinants of health: Opportunities and challenges," *International Journal of Population Data Science*, vol. 4, 11 2019.

[27] K. Adane, M. Gizachew, and S. Kendie, "The role of medical data in efficient patient care delivery: a review.," *Risk management and healthcare policy*, vol. 12, pp. 67–73, 4 2019.