Scidata Consortium
An international open-access platform

SDC SciDataConsortium

Research Article

OPEN ACCESS

Service-Aware Continuous Prioritization for ICU Deterioration Under Queue and Attention Constraints

Krzysztof Mazur[1], Lukasz Dabrowski[2]

[1] Opole University of Technology, Department of Computer Science, Proszkowska 76 Street, 45-758 Opole, Poland;
[2] University of Zielona Gora, Department of Information Engineering, Licealna 9 Street, 65-417 Zielona Gora, Poland

Abstract

Intensive care prediction is typically framed as binary event forecasting, yet bedside surveillance is rarely practiced as a sequence of isolated yes-or-no decisions. What clinicians actually manage is a continuously changing allocation problem: which patients deserve immediate review, which require intensified monitoring, which are rising in urgency, and which can be safely deprioritized for the moment. This distinction matters because the statistical object most models produce is not the object the unit consumes. The practical output of a deterioration system is a ranked and repeatedly updated worklist constrained by finite review capacity, uncertain evidence quality, intervention saturation, and queue spillover effects. This paper develops a data-centered treatment of ICU deterioration modeling as service-aware continuous prioritization rather than simple binary alarm generation. The framework formalizes the active census as a coupled collection of partially observed stochastic processes whose scores compete for a limited number of review slots. It then derives a dynamic priority model that integrates irregular physiologic streams, time-varying intervention context, note refresh dynamics, and support-aware uncertainty into a population-level queueing formulation. Particular attention is paid to the geometry of the actionable tail, the distinction between risk and review value, and the way calibration, score smoothness, and uncertainty interact with constrained service capacity. The resulting perspective connects sequence modeling, ranking theory, state estimation, queue control, and human-in-the-loop governance. Rather than treating deterioration prediction as a narrow classification task, the paper argues that clinically useful systems should be designed as streaming prioritization engines whose outputs remain stable, interpretable, and capacity-aware under the operational realities of intensive care.

1. Introduction

The modern intensive care unit is one of the few settings in which machine learning appears to possess almost ideal raw material [1]. Patients are observed repeatedly, large numbers of physiologic variables are stored, interventions are timestamped, and clinical documentation accumulates around the clock. There are clear adverse events of interest, substantial incentives for early recognition, and electronic systems capable of carrying predictive outputs back into workflow. Yet these favorable conditions conceal a more difficult computational truth. The problem faced by an actual ICU is not the same as the one typically encoded in a supervised benchmark. A benchmark asks whether deterioration will occur within some future horizon. A unit asks which patients should be reviewed first, how that ordering should change when new evidence arrives, how many cases can be escalated without overwhelming staff, and how uncertainty should modulate scarce attention. The first problem is statistical classification. The second is dynamic priority allocation under partial observability and limited service capacity.

That distinction is not semantic. It changes what the model should optimize, what its output should look like, and what it means for the system to work. If a prediction engine emits an hourly risk for each patient, clinicians do not act on all of those risks independently. They compare them, informally or formally [2]. A high-scoring patient displaces someone else from the top of the worklist. A noisy upper tail creates more competing "urgent" cases than can be reviewed. A brittle ranking that oscillates from hour to hour causes attention churn even when overall discrimination is respectable. Thus, the clinically consumed

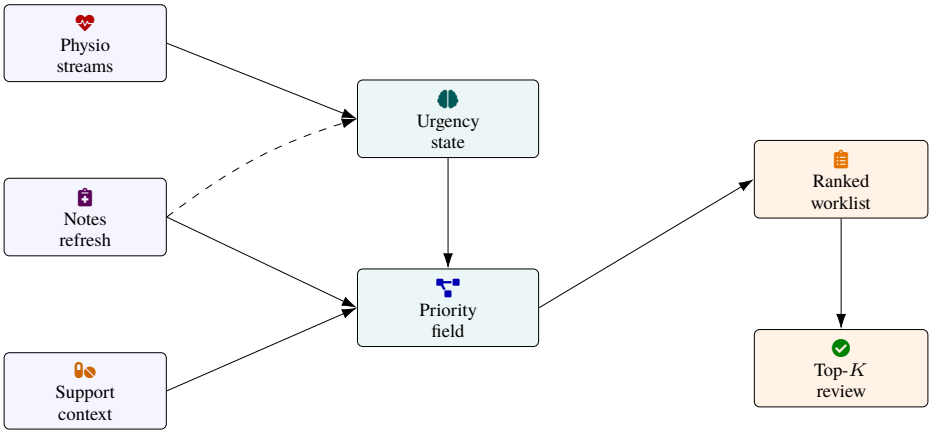

Figure 1: This overview figure frames the ICU system as a streaming prioritization pipeline. Physiologic signals, note updates, and support context are fused into a latent urgency state and a census-level priority field, which is then converted into a ranked worklist. Only a small top segment can realistically enter review, so the figure emphasizes the path from heterogeneous evidence to scarce attention allocation rather than binary alert firing.

object is not the marginal probability of deterioration for one patient-time window in isolation. It is a continuously changing ordering over the census, produced by the model and filtered through the capacity of humans and protocols to respond.

This means that many familiar modeling choices are answering the wrong question. An ICU surveillance system should not be judged only by whether positives receive higher scores than negatives on average. It should be judged by whether the top of the ranked list is sufficiently enriched, sufficiently stable, and sufficiently interpretable to justify consuming one more review slot. That criterion immediately reveals why the usual separation between prediction and operations is artificial. Score calibration, alert thresholds, queue length, service capacity, review cost, and score volatility all determine the practical value of the same model. In other words, the statistical model and the service process are coupled. A deterioration predictor is not merely deployed into a workflow after the fact. It helps define the workflow by deciding who enters the review queue, when, and how often [3].

The low prevalence of new adverse events within short horizons intensifies the problem. When only a small fraction of hourly windows contain events within the next day, a predictor can achieve decent global discrimination while still presenting far too many patients as moderately urgent. Such a model is acceptable as a scoring rule in a paper and troublesome as a surveillance engine in a busy unit. The clinically useful region of the score distribution is not the full interval from zero to one. It is the extreme right tail, where queue slots are contested and where the difference between the tenth and eleventh ranked patients can determine whether someone receives timely attention. A model that is "good" in the statistical sense but diffuse in that tail may add less value than a model with slightly lower AUROC and far better top-of-list concentration.

Another problem emerges once one admits that not all high-risk patients benefit equally from the same action. A patient who is already receiving maximal bedside attention and high-intensity support may be correctly assigned high deterioration risk, but further queue elevation may have little marginal value. Another patient with only slightly lower event probability but limited current attention and a modifiable trajectory may be the more important review target. This means the system should not optimize raw risk alone. It should optimize review value, a quantity that depends on risk, lead time, intervention opportunity, and the current state of care saturation. A pure event forecaster has no direct place to encode that distinction unless it is rebuilt as a service-aware prioritizer.

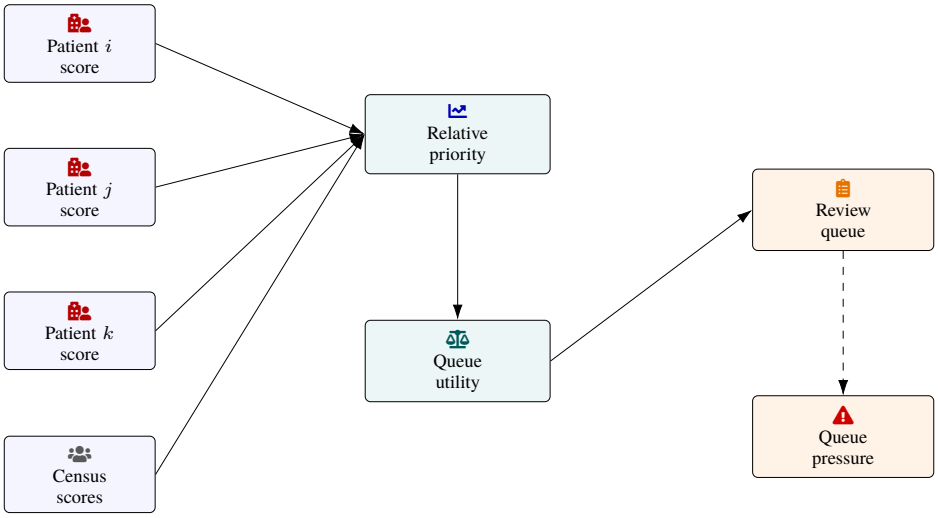

Figure 2: This diagram shifts the view from isolated patient prediction to census-level competition. Individual scores are first interpreted relative to the active unit and then translated into queue utility, which determines which patients actually occupy finite review capacity. The lower block highlights the operational consequence of score inflation or tail crowding: when too many cases cluster near the top, queue pressure rises and the worklist becomes harder to use effectively.

The nature of clinical text further complicates the picture [4]. Notes do not simply add more predictors. They alter how risk should be updated over time because they summarize state, concern, and trajectory in ways structured variables cannot. Yet they are also temporally problematic. Their semantic signal can persist longer than an individual measurement but age less predictably. The need to model document refresh explicitly is highlighted by Sadanandan (2025) [5], who notes that reusing the same radiology-note embedding across all hourly predictions removes temporal granularity and helps explain why text alone is insufficient for near-term forecasting. That observation is not just about note modeling. It exposes a broader systems issue: a prioritization engine must know how evidence ages if it is to update rankings coherently. A stale but semantically rich note and a fresh but narrow measurement should not influence the queue in the same way.

This paper develops an alternative formulation in which ICU deterioration modeling is treated as service-aware continuous prioritization. The system is described as maintaining a population-level priority field over the active census rather than a collection of independent window-level predictions. Every patient carries a latent urgency state, an uncertainty state, and a marginal review value conditioned on current context. Scores are consumed through a constrained queue rather than through static thresholding. Notes, measurements, and interventions contribute to this process not only as features, but as evolving evidence streams whose influence on queue position should depend on freshness, support, and actionability [6]. The resulting problem sits at the intersection of temporal machine learning, ranking theory, queue control, and human oversight.

The layout of the paper follows that reframing. The next section identifies the real computational object of ICU surveillance once one stops pretending that the deployed task is a binary alarm. After that, the discussion turns to a census-level state model in which patients compete for limited review capacity and influence one another only through the service process that consumes their scores. The paper then develops a treatment of textual refresh and documentary half-life, because a continuous prioritizer must know when semantic evidence remains relevant enough to affect rank. Later sections address learning objectives tailored to the top of the worklist, selective escalation and review economics, and finally the operational governance of such systems under drift, feedback, and bounded human authority. The goal throughout is not to exaggerate the novelty of any one model component. It is to make clear that once the object of prediction is correctly specified, many design choices that look optional in ordinary supervised learning become unavoidable.

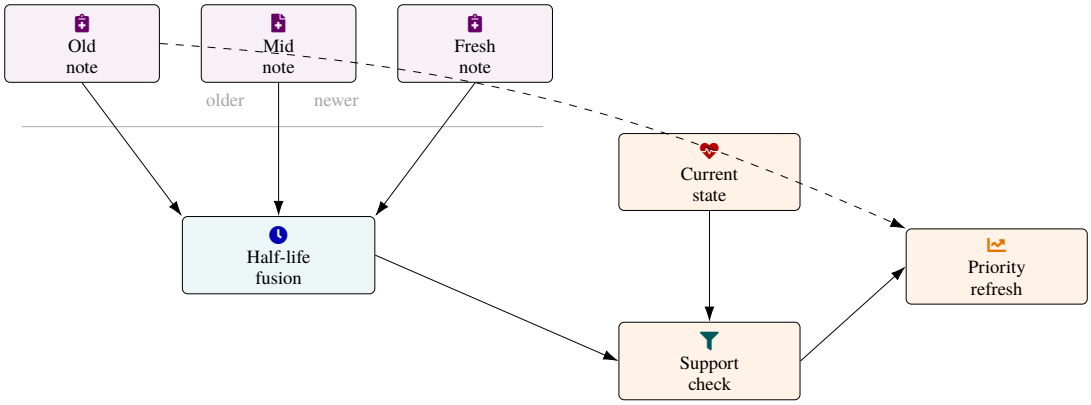

Figure 3: This figure encodes documentary half-life. Notes are treated as time-sensitive evidence packets whose influence depends on both recency and semantic persistence, rather than as static embeddings reused forever. After temporal fusion, note-derived signal is checked against the current structured state before it can contribute to priority refresh, allowing fresh concern to matter while stale unsupported documentation naturally fades.

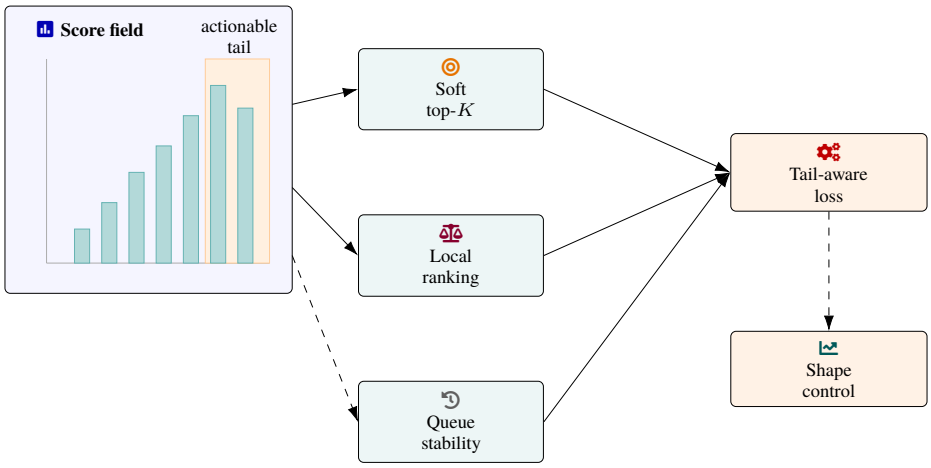

Figure 4: The training view concentrates modeling effort on the actionable tail instead of the full score distribution. Soft top-K weighting encourages future high-value events to rise into the review region, local ranking emphasizes competition among simultaneously present patients, and queue-stability regularization discourages unnecessary churn. A separate shape-control component keeps the upper tail sufficiently separated for prioritization without letting it become erratic or unusably compressed.

2. The Output Is a Worklist, Not an Alarm

The easiest way to mis-specify ICU deterioration modeling is to confuse the label with the deployed object. The label may well be binary, such as deterioration within twenty-four hours. The deployed object is not. The deployed object is a ranked worklist. At any given time, the system receives a census of patients and assigns each one a place in an ordering that competes for limited attention [7]. The practical question is therefore not "Will this patient deteriorate?" in isolation. It is "Where should this patient sit relative to all others who also compete for review, and how should that position change as evidence arrives?" A surveillance system that fails to produce a coherent answer to that second question can be impressive on paper and underwhelming in use.

Let the active census at time t be $\mathcal{I}(t)$, and let $s_i(t)$ be the model score for patient $i \in \mathcal{I}(t)$. The ordering induced by the model is

$$\pi_t = \operatorname{argsort}\Big(\{s_i(t)\}_{i\in\mathcal{I}(t)}\Big). \tag{1}$$

This ordering, not the isolated score vector, is what clinicians effectively consume. The top few positions receive review, heightened surveillance, or immediate escalation. The middle region may shape situational awareness without action. The lower region is usually ignored. Conse-

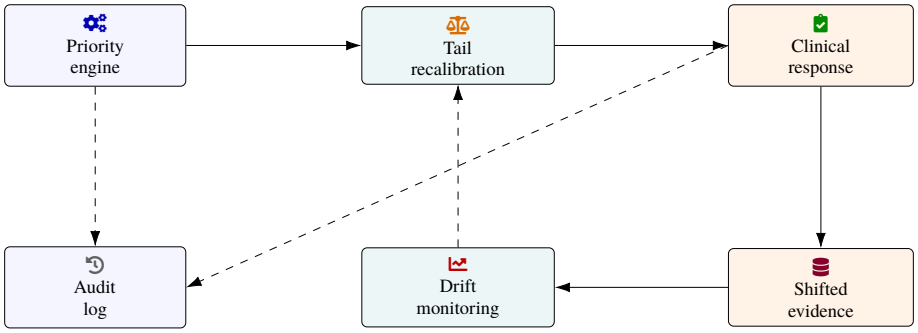

Figure 5: The deployment loop is governed as a feedback system. Model outputs influence clinical responses, those responses change the downstream evidence pattern, and the changed evidence in turn affects future ranking behavior. To keep the worklist trustworthy over time, the architecture includes audit logging, drift monitoring, and constrained recalibration focused on preserving the operational meaning of the upper tail rather than allowing uncontrolled shifts in queue behavior.

Aspect	Binary event prediction	Service-aware prioritization
Unit of decision	Isolated patient-time window	Full census at time t competing for limited review slots
Output object	Risk score or alarm per window	Ranked worklist π_t and queue $Q_t \subseteq \mathcal{I}(t)$
Error notion	Misclassified window (FP/FN)	Misallocated queue position and displaced patients
Capacity awareness	Typically ignored or added post hoc	Built-in via K_t, queue pressure, and service cost
Temporal behavior	Independent predictions across time	Continuously updated ordering with controlled churn
Evaluation focus	AUROC, AUPRC over all windows	Tail enrichment, queue stability, utility per slot

Table 1: Conceptual contrast between standard binary deterioration prediction and the worklist-centric formulation.

quently, errors should be understood relationally. If one patient is ranked too high, another patient is ranked too low. A false positive at the top is not simply one extra alarm. It is a displaced queue slot. A false negative is not merely a missed label. It is a patient whose priority remained too low to receive marginally useful attention [8].

This point changes the meaning of utility. In conventional binary decision theory, a prediction is evaluated according to whether it crosses a threshold and whether the future event occurs. In worklist form, the immediate utility of a score depends on what other scores exist at the same time, how many patients can be acted upon, and whether a patient already under active management gains anything from additional priority. The model's output is therefore census-relative. The urgency of a score is contextualized by the distribution of all other scores and by current capacity. A probability that would look high in a quiet unit may not merit active review in a crowded high-acuity shift if several other patients have still higher and more actionable values.

A simple formalization is to define a review queue $Q_t \subseteq \mathcal{I}(t)$ of patients selected for active attention. If the service capacity at time t is K_t, then the simplest policy is

$$Q_t = \left\{ i \in \mathcal{I}(t) : \mathrm{rank}_{\pi_t}(i) \leq K_t \right\}. \tag{2}$$

This top-K_t rule is crude but reveals the essential structure. The model does not simply classify each patient independently. It allocates scarce queue membership. Under this interpretation, the key quantity is no longer average classification accuracy. It is the composition of Q_t, its stability over time, and the degree to which its members genuinely warrant review.

The geometry of the right tail of the score distribution becomes the dominant concern once the system is viewed this way [9]. Suppose the model produces many

Symbol	Level	Informal meaning
$\mathcal{I}(t)$	Census	Set of patients present in the unit at time t
$s_i(t)$	Patient	Raw model score / risk for patient i at time t
π_t	Census	Ordering induced by scores at time t (argsort over $\{s_i(t)\}$)
Q_t	Queue	Subset of $\mathcal{I}(t)$ admitted for active review at time t
K_t	Capacity	Available review capacity (max $\lvert Q_t \rvert$) at time t
$z_i(t)$	Patient	Latent urgency state summarizing trajectory and context
$r_i(t)$	Census	Standardized priority relative to census, based on $s_i(t)$

Table 2: Key notation for the census-level state and queue formulation.

Component	Notation	Source	Role in queue utility
Raw risk	$\hat{p}_i(t)$ or $s_i(t)$	Temporal sequence model	Estimates event probability within horizon
Standardized priority	$r_i(t)$	Cross-sectional normalization	Positions patient relative to current census
Confidence / support	$c_i(t)$	Evidence freshness, agreement	Modulates aggressiveness of queue admission
Marginal review value	$g_i(t)$ or $m_i(t)$	Lead time, modifiability	Encodes expected benefit of spending a slot now
Action saturation	$a_i(t)$	Current bedside activity	Discounts patients already under maximal attention
Queue utility	$u_i(t)$	Combination $h(\cdot)$	Scalar used to construct $Q_t^\star$ under capacity K_t

Table 3: Decomposition of the operational utility driving queue membership.

moderately high scores spread across the census. Then the boundary between patients ranked K_t and $K_t + 1$ becomes noisy and unstable. A small change in a lab value or note may cause a large change in queue membership, not because the patient changed meaningfully, but because the upper tail is too flat. By contrast, a model with a sharply concentrated tail can support a more decisive queue. The top few patients are clearly separated from the rest, and queue membership reflects a stronger expected difference in utility. It follows that tail concentration is more important than whole-distribution discrimination once capacity enters the problem.

This observation also clarifies why ordinary metrics only partially characterize a surveillance engine. AUROC evaluates pairwise ranking over all thresholds and all score regions. AUPRC focuses more on the rare positive class but still averages over operating points. Neither metric measures queue occupancy, queue turnover, or enrichment in the exact part of the score distribution that is consumed by staff. For a service-aware system, one also wants to know how many true future events are represented among the top K_t, how much earlier those queued cases are captured relative to outcomes, and how often the queue changes membership in ways that appear clinically meaningful rather than noisy. These are not cosmetic add-ons. They are core properties of the deployed object [10].

A worklist view also suggests that ranking should not be tied too tightly to static thresholds. A fixed threshold can make sense in deployment, but it hides the fact that queue size is constrained and variable. Units differ by census size, staffing, escalation infrastructure, and tolerance for alert burden. Therefore a score should ideally support both score-based and capacity-based interpretation. In one regime, a patient enters active review because the probability exceeds a validated threshold. In another, a patient enters because they are among the top few patients in a temporary surge of acuity. Capacity-aware ranking accommodates both cases.

Evidence source	Typical cadence	Aging behavior	Effect on priority
Bedside vitals	Minutes to hours	Rapidly stale; frequent refresh	Drives short-term urgency and score slope
Laboratory results	Hours	Moderate half-life; context dependent	Anchors medium-horizon risk, responds to treatment
Clinical notes	Hours to daily	Semantic signal persists but with uncertain decay	Provide documentary lift and trajectory hints
Interventions	Event-driven	Effect depends on intensity and timing	Alter marginal review value via saturation $a_i(t)$
Background data	Admission/once	Very slow decay	Shape baseline risk but rarely affect queue dynamics alone

Table 4: Qualitative temporal behavior of different evidence streams and their influence on the worklist.

Mechanism	Notation / form	Priority effect	Comment
Recency weighting	$\omega_{i,k}(t)$	Favors more recent documents in $c_i^{(d)}(t)$	Basic time-aware attention over notes
Document survival	$h_{i,k}(t) = \exp(-\lambda(\cdot)(t-\tau_{i,k}))$	Learns half-life per note type and state	Distinguishes persistent from fleeting claims
Documentary lift	$\Delta_i^{(d)}(t) = w_d^\top c_i^{(d)}(t)$	Shifts risk level $s_i(t)$ via text	Allows notes to move patients up or down the list
Trajectory influence	$\dot{s}_i(t) = \phi_s(z_i^{(0)}(t), c_i^{(d)}(t), u_i^{\text{clin}}(t))$	Changes score slope, not just level	Captures rising concern before structured changes
Document confidence	$c_i^{\text{doc}}(t) = \psi_c(\cdot)$	Gates how strongly text affects queue utility	Downweights stale or weakly supported notes

Table 5: Elements of the documentary half-life model and how they shape queue behavior.

Another underappreciated implication is that score magnitude and score position are not the same object. A patient can have a stable absolute score yet drift upward in the worklist because the rest of the census stabilizes. Conversely, a patient's score can rise but their rank can remain almost unchanged if multiple other patients worsen more. Since review capacity is tied to rank, not only to score, the surveillance engine should be analyzed at both levels. This motivates keeping track of census-relative urgency, which will later be formalized as standardized priority [11].

The worklist object also makes it possible to speak clearly about operational stability. Clinicians do not merely need a model whose top scores correspond to future events. They need a model whose top-of-list behavior is coherent enough to support action. A worklist that churns constantly under minor documentation noise is hard to trust even if it is statistically respectable. A worklist that remains overly static despite new physiologic deterioration is equally poor. The correct target is controlled mobility: queue membership should change when evidence changes in important ways and should resist change when the evidence is effectively the same. This is a temporal systems property rather than a simple predictive one.

Finally, framing the output as a worklist emphasizes that the model participates in a social and organizational process. High placement can trigger chart review, bedside reassessment, escalation huddles, additional tests, or more frequent monitoring. In each case, the score becomes an instrument of labor allocation. This practical role imposes ethical and governance requirements that are harder to see when the system is imagined as a neutral classifier. If the queue systematically favors well-documented patients, for example, then surveillance capacity is being distributed unevenly. If the queue is driven heavily by intervention artifacts, then staff may be repeatedly directed

Loss term	Primary target	Queue-facing effect	Section reference
$\mathcal{L}_{\text{focal}}$	Rare-event discrimination	Maintains sensitivity under class imbalance	Learning the actionable tail
$\mathcal{L}_{\text{tail}}^{\text{val}}$	High-value positives near top	Enriches actionable tail with modifiable cases	Learning the actionable tail
$\mathcal{L}_{\text{queue\_tv}}$	Soft queue membership over time	Reduces unnecessary churn of Q_t	Learning the actionable tail
$\mathcal{L}_{\text{pair}}$	Within-census comparisons	Preferentially ranks valuable positives above competitors	Learning the actionable tail
$\mathcal{L}_{\text{trans}}$	State transitions	Encourages upward trajectories before events	Learning the actionable tail
$\mathcal{L}_{\text{shape}}$	Upper-tail geometry	Controls separation among highest scores	Learning the actionable tail
$\mathcal{L}_{\text{conf}}$	Confidence calibration	Aligns $c_i(t)$ with evidence quality and reliability	Learning the actionable tail

Table 6: Tail-focused and queue-aware learning terms used to train the prioritization model.

Metric	Informal definition	Depends on	Operational question
Top-K event recall	Fraction of future events with patient in top K_t	π_t, Q_t	Are scarce slots spent on true deteriorations?
Lead time among queued	Time between first queue entry and outcome	$Q_t, y_i(t)$	How early does the worklist surface deteriorating patients?
Queue turnover $T_K(t)$	Jaccard change in top-K_t set between steps	Q_t, Q_{t-1}	Is the worklist stable enough to act on?
Queue pressure	Relationship of arrival λ_t to service μ_t	Induced service process	Does the score distribution overload reviewers?
Utility per slot	Realized benefit per reviewed patient	Actions, outcomes	Is each review tier economically justified?

Table 7: Example metrics that evaluate the worklist as the primary deployed object.

toward patients already receiving maximal attention [12]. Such issues belong to the core computational specification because they arise from how scores create rankings, not from superficial user-interface choices.

The main point of this section is that the deployed object of ICU deterioration prediction is a dynamically updated worklist. The next step is to model that worklist at the level of the census, not merely patient by patient. Doing so reveals how queue capacity, cross-sectional score distributions, and patient-specific state evolution jointly determine who receives attention.

3. A Census-Level State Model

Most clinical sequence models are patient-centric. They take one patient at one time, perhaps with a look-back window, and output a risk estimate. This is sensible as a building block, but a prioritization engine exists at the census level. At time t, the model observes not one patient but all patients currently present. Those patients do not interact causally in the medical sense, yet they interact operationally because they compete for queue slots, reviewer bandwidth, and escalation resources. A census-level model does not need to make one patient's physiology depend on another's. It does need to make one patient's operational priority relative to the entire field of scores visible and actionable.

Let $z_i(t)$ be the latent urgency state of patient i, summarizing the information relevant to deterioration forecasting, intervention interpretation, and evidential support. A patient-centric model would map $z_i(t)$ to a scalar score $s_i(t)$ [13]. The census-level perspective adds a normal-

Tier a	Typical action	Capacity symbol	Usage in policy
0: background	No immediate action; routine monitoring	–	Default state for low-priority patients
1: passive watchlist	Highlight in dashboard; optional review	$K_t^{(1)}$	Keeps medium-risk or low-confidence cases visible
2: chart review	Structured review by clinician	$K_t^{(2)}$	Primary consumption of queue slots under $K_t^{(2)}$
3: active escalation	Bedside reassessment, huddle, protocolized response	$K_t^{(3)}$	Reserved for highest utility and confidence cases

Table 8: Illustrative action tiers built on top of the continuous worklist and capacity constraints.

Drift type	Definition	Example manifestation	Governance signal
Score drift	Change in distribution of $\tilde{p}_i(t)$ or $u_i(t)$	Global inflation of risk after documentation change	Time-varying calibration curves, tail quantiles
Queue drift	Change in composition or turnover of Q_t	Different case mix in top-K_t despite similar AUROC	Shift in queue-state embedding $D_Q(t), T_K(t)$
Policy drift	Change in clinician response to same scores	New protocol alters when escalation is triggered	Divergence between proposed and realized actions
Confidence drift	Systematic shift in $c_i(t)$ for similar states	More low-confidence queues after template redesign	Joint distribution of $(\tilde{p}, c)$ in the tail
Subgroup drift	Differential behavior across patient groups	One subgroup gains fewer tier-2 slots over time	Tier-level fairness and access monitoring

Table 9: Major forms of drift for a deployed prioritization engine and how they can appear in practice.

ization layer over the active set $\mathcal{I}(t)$. Define the cross-sectional mean and variance

$$\mu_s(t) = \frac{1}{|\mathcal{I}(t)|} \sum_{i \in \mathcal{I}(t)} s_i(t),$$
$$\sigma_s^2(t) = \frac{1}{|\mathcal{I}(t)|} \sum_{i \in \mathcal{I}(t)} \big(s_i(t) - \mu_s(t)\big)^2. \tag{3}$$

Then define standardized priority

$$r_i(t) = \frac{s_i(t) - \mu_s(t)}{\sigma_s(t) + \varepsilon}. \tag{4}$$

The quantity $r_i(t)$ is not simply a rescaled score. It is the patient's urgency relative to the present census. Since the queue admits only a small portion of the unit, $r_i(t)$ is often more operationally meaningful than $s_i(t)$ itself.

A census-level system should not necessarily threshold $s_i(t)$ directly. It can instead compute a queue utility

$$u_i(t) = h\big(s_i(t), r_i(t), c_i(t), m_i(t)\big), \tag{5}$$

where $c_i(t)$ is confidence or support quality and $m_i(t)$ is an estimate of marginal benefit from further review given current treatment intensity. Queue admission then solves

$$Q_t^\star = \arg\max_{Q \subseteq \mathcal{I}(t)} \sum_{i \in Q} u_i(t)$$
$$\text{s.t.} \quad |Q| \le K_t. \tag{6}$$

This formulation departs from the common assumption that the top-K_t by predicted event probability is always the optimal queue. If some patients are already under maximal attention or if their trajectories are less modifiable, then expected review value may differ from event proba-

bility. The model should therefore learn not only who is risky, but who is worth a scarce slot now.

A useful dynamical interpretation is to treat $z_i(t)$ as evolving under evidence and interventions, while $u_i(t)$ is the control-oriented projection of that latent state into the queue [14]. The state evolution can be written as

$$z_i(t+\Delta t) = F\big(z_i(t), o_i(t), u_i^{\text{clin}}(t)\big) + \epsilon_i(t), \quad (7)$$

where $o_i(t)$ are new observations and $u_i^{\text{clin}}(t)$ are clinical interventions. The queue utility is then

$$u_i(t) = G\big(z_i(t), \kappa_i(t)\big), \quad (8)$$

with $\kappa_i(t)$ representing service-related variables such as confidence, action saturation, or even unit-level load. Separating F and G is important. The first governs how the patient evolves. The second governs how the system should allocate attention.

Intervention saturation deserves special treatment because it is one of the clearest places where prediction and operations diverge. A patient can be extremely high risk while offering little additional utility for queue elevation if the team is already fully mobilized around them. Let $a_i(t)$ summarize the intensity of ongoing review and intervention. Then one may define a saturation-adjusted utility

$$u_i(t) = \tilde{u}_i(t) - \lambda_a \, \phi_a\big(a_i(t)\big), \quad (9)$$

where $\tilde{u}_i(t)$ is the baseline review utility derived from risk and lead time, and ϕ_a is increasing in current action saturation. This does not mean the patient is no longer important. It means the marginal value of yet another queue slot assigned to that patient is smaller. The queue thus becomes better aligned with opportunity cost [15].

A census-level model also exposes an important source of instability: global score inflation. Suppose documentation density increases across the entire unit, or a calibration shift causes all scores to rise. In a patient-centric evaluation this may look like a tolerable change in absolute scale. In a queue system it can produce either too many queue candidates or artificially flattened relative priorities if normalization is not handled carefully. Monitoring $\mu_s(t)$, $\sigma_s(t)$, and the empirical distribution of $r_i(t)$ provides a more direct view of operational drift than monitoring isolated calibration errors alone.

Another advantage of the census view is that it allows explicit modeling of queue pressure. Let λ_t denote the arrival rate into the review queue and μ_t the service rate. Under persistent overload, waiting time grows rapidly. A stylized penalty on queue pressure can be written as

$$\mathcal{C}_{\text{queue}}(t) = \zeta \Big(\max(0, \lambda_t - \mu_t) \Big)^2. \quad (10)$$

This penalty need not be used exactly in deployment, but it captures the fact that a model should be judged partly by the service process it induces. A score distribution that leads to chronic overflow is not benign simply because its AUROC is good. By making queue pressure explicit, the census-level model connects predictive scoring to capacity planning.

To make the queue process more concrete, define soft queue membership weights [16]

$$w_i(t) = \frac{\exp(\beta_u u_i(t))}{\sum_{j \in \mathcal{I}(t)} \exp(\beta_u u_j(t))}, \quad (11)$$

where β_u controls sharpness. When β_u is large, the highest-utility patients dominate. These weights can be used during training to approximate the top-K_t queue in a differentiable way. More importantly, they provide an operational interpretation of marginal queue probability. A patient with high $w_i(t)$ is not just high risk. They are likely to occupy one of the scarce attention slots.

Temporal coherence at the census level is also different from temporal coherence at the patient level. Even if every patient's score is smooth, the queue can still fluctuate if the top tail is crowded. Conversely, patient-level volatility may matter little if it occurs deep in the queue where no action is taken. Therefore a service-aware system should focus its regularization and evaluation on the upper ranked region. One may define top-tail turnover

$$T_K(t) = 1 - \frac{|Q_t \cap Q_{t-1}|}{|Q_t \cup Q_{t-1}|}, \quad (12)$$

and explicitly penalize unnecessary changes in Q_t during training or threshold tuning. Doing so shifts the objective from abstract score stability to the stability of the actual object consumed by staff [17].

The census-level formulation also helps explain why support and confidence matter. In a patient-centric model, uncertainty is typically interpreted as epistemic or aleatoric error around one score. In a queue, uncertainty modifies how aggressively one should spend scarce slots. A high-risk low-confidence patient can be kept visible without displacing a high-risk high-confidence patient. Thus, confidence should enter queue utility as a multiplicative or gating factor, not as an afterthought. This issue will re-

turn in the sections on escalation and governance, but it belongs already at the level of the census state.

The main gain of the census perspective is conceptual clarity. It makes clear that the ICU surveillance system is a population process. Scores are not merely private estimates attached to individuals. They are competitive claims on shared attention. Once that is acknowledged, the next challenge is to determine how individual priority states should evolve through time, including how text and other delayed or persistent evidence should refresh or decay within a continuously updated queue.

4. Documentary Half-Life and Priority Refresh

Not all evidence ages in the same way. Bedside vital signs lose freshness quickly but may update often [18]. Laboratory results can remain informative for hours, though their interpretation depends on time since measurement and on subsequent treatment. Clinical notes occupy an even stranger temporal regime. They are semantically dense, often highly contextual, and not naturally synchronized to hourly prediction windows. In a service-aware prioritization system, notes matter because they can change a patient's queue trajectory even when raw physiologic measurements look similar. Yet they can also distort that trajectory if treated as permanently current. A continuous prioritizer therefore needs a model of documentary half-life: how long a piece of documentation should continue to influence queue position and under what circumstances it should be refreshed, reinforced, or allowed to decay.

The relevance of this issue is emphasized by Sadanandan (2025) [5], who notes that reusing the same radiology-note embedding across all hourly predictions strips text of temporal granularity. Once a note is encoded once and propagated unchanged across the whole future sequence, it behaves like an immortal feature rather than a time-sensitive document. This can be acceptable for stable background context, but it is poorly matched to a queue whose meaning changes hour by hour. A note that meaningfully elevates concern when filed should not necessarily continue to carry the same priority weight twelve hours later if structured evidence has changed and no new textual support has appeared.

To model documentary half-life, consider each textual unit $d_{i,k}$ with filing time $\tau_{i,k}$, semantic representation $v_{i,k}$, and document type $l_{i,k}$. Let the current patient state before text refresh be $z_i^{(0)}(t)$. A refresh-aware note contribution can be written as

$$\omega_{i,k}(t) = \frac{\exp\Big(g(z_i^{(0)}(t), v_{i,k}, l_{i,k}, t-\tau_{i,k})\Big)}{\sum_{j:\tau_{i,j}\le t}\exp\Big(g(z_i^{(0)}(t), v_{i,j}, l_{i,j}, t-\tau_{i,j})\Big)},$$
$$c_i^{(d)}(t) = \sum_{k:\tau_{i,k}\le t}\omega_{i,k}(t)v_{i,k}. \tag{13}$$

This is not yet enough. A pure recency-weighted attention still treats time decay as a generic function rather than as a document-state interaction [19]. Some documents should decay quickly unless reinforced by new measurements. Others describe persistent state and can remain relevant longer. Therefore the decay law itself should be conditioned on content and evolving patient state.

A more expressive model defines a document survival factor

$$h_{i,k}(t) = \exp\Big(-\lambda(v_{i,k}, l_{i,k}, z_i^{(0)}(t))\,(t-\tau_{i,k})\Big), \tag{14}$$

where $\lambda(\cdot)$ is learned. A persistent radiologic burden or chronic ventilator dependence may produce a small λ, while a fleeting bedside impression may produce a larger one. Then the note contribution becomes

$$c_i^{(d)}(t) = \sum_{k:\tau_{i,k}\le t}\tilde{\omega}_{i,k}(t)\,h_{i,k}(t)\,v_{i,k}, \tag{15}$$

with $\tilde{\omega}$ encoding semantic relevance. This allows the same note to decay differently depending on how the rest of the patient state evolves. In queue terms, a note is not a one-time additive feature. It is a reservoir of concern whose remaining priority value depends on whether the current trajectory continues to support its claims.

This perspective naturally connects text to risk refresh. Let $s_i^{(0)}(t)$ be the structured-only or pre-text score and $s_i(t)$ the post-text score. Then

$$s_i(t) = s_i^{(0)}(t) + \Delta_i^{(d)}(t), \quad \Delta_i^{(d)}(t) = w_d^\top c_i^{(d)}(t). \tag{16}$$

The quantity $\Delta_i^{(d)}(t)$ is the documentary lift. A fresh and relevant note can move a patient upward in the worklist [20]. The same note should lift less as its half-life expires. If new structured evidence reinforces it, the lift can persist or increase. This makes documentary influence a dynamic contribution to rank rather than a static embedding merged once and then forgotten.

An important refinement is to allow documents to affect not only score level but score slope. A new note

may not immediately imply high absolute risk, but it may indicate that a patient's trajectory is shifting in a way that warrants closer watch. To capture that, let the short-term expected score change $\dot{s}_i(t)$ depend on document refresh:

$$\dot{s}_i(t) = \phi_s\big(z_i^{(0)}(t), c_i^{(d)}(t), u_i^{\text{clin}}(t)\big). \tag{17}$$

In this formulation, text acts partly as a momentum signal. A note documenting rising concern can accelerate a patient upward in the queue before structured variables alone would place them high enough for review. This is particularly useful in service-aware surveillance, where the benefit of the model lies partly in capturing rising-risk patients early enough to justify a scarce queue slot.

Because notes are not equally trustworthy, documentary refresh should also produce a confidence component. If only one stale note is available, then the model should not be as confident about documentary lift as it would be when multiple recent notes and measurements tell the same story. A document confidence variable can be written as

$$c_i^{\text{doc}}(t) = \psi_c\big(\{\omega_{i,k}(t)\}_k, \{h_{i,k}(t)\}_k, \Gamma_i^{(sd)}(t)\big), \tag{18}$$

where $\Gamma_i^{(sd)}(t)$ measures agreement between document-derived and structured-derived latent states. This confidence can later gate queue admission or influence how strongly a document-induced score increase is allowed to move a patient upward.

A crucial failure mode arises when documentation frequency differs across units or shifts over time [21]. In such cases, a model that has learned a strong documentary lift may begin overprioritizing note-rich settings or underprioritizing note-poor ones. To reduce this, one can apply documentary dropout during training, randomly suppressing or delaying some note contributions. This forces the model to learn that notes are supportive and context-enhancing, not mandatory scaffolding for every queue movement. The goal is graceful degradation rather than invariance. When text disappears, the score should usually become less confident, not collapse or remain falsely precise.

One can also model document refresh at the sentence or claim level rather than at the whole-note level. This is especially useful because a single note often mixes stable context with highly perishable state descriptions. In that case, each claim $q_{i,k,m}$ within note k receives its own half-life parameter and support interaction with the current structured state. The resulting representation is more computationally demanding but conceptually cleaner. The system can keep long-lived background claims alive while allowing acute observations to decay quickly. Such granularity is particularly aligned with a worklist, since the top of the queue is most sensitive to short-lived, high-salience textual observations.

The broader point is that documentary evidence in ICU surveillance should be treated as refreshable fuel for queue position, not as a static modality. A prioritization engine that ignores documentary half-life risks making text either too weak to matter or too permanent to trust. By contrast, a system that models textual lift, decay, and confidence can use documentation to update rankings in a way that is both more sensitive and more disciplined [22]. This sets the stage for learning objectives that reward not just event discrimination, but the formation of a useful and stable top-of-list. Those objectives are the focus of the next section.

5. Learning the Actionable Tail

Once the output is recognized as a dynamic worklist, the learning problem changes. A model trained only to separate positives from negatives across all windows may not produce a useful worklist because the statistical bulk of the data is dominated by negative windows far from the decision boundary and by repeated windows within the same stays. In deployment, however, almost none of that bulk matters directly. The unit consumes only a small actionable tail of the ranking. Therefore the learning objective should pay disproportionate attention to where positive mass accumulates near the top, how that tail behaves over time, and whether queue slots are spent on cases with real marginal review value.

Let $\eta_i(t)$ be the raw logit, $\hat{p}_i(t) = \sigma(\eta_i(t))$ the raw risk, and $y_i(t)$ the binary deterioration label. A sensible base term remains class-imbalance-aware cross-entropy or focal loss:

$$\begin{aligned}\mathcal{L}_{\text{focal}} = &-\sum_{i,t} \alpha y_i(t)(1-\hat{p}_i(t))^{\gamma} \log \hat{p}_i(t) \\ &-\sum_{i,t}(1-\alpha)(1-y_i(t))\hat{p}_i(t)^{\gamma}\log(1-\hat{p}_i(t)).\end{aligned} \tag{19}$$

This encourages sensitivity to rare positives, but it remains agnostic to queue structure. A patient barely above many negatives and a patient decisively at the very top are treated similarly if their contribution to the loss is comparable. That is not good enough for a prioritizer.

To shape the actionable tail, consider a differentiable approximation to top-K membership. Let $u_i(t)$ be the queue utility score, not necessarily identical to $\eta_i(t)$ [23]. Define soft queue weights

$$w_i(t) = \frac{\exp(\beta u_i(t))}{\sum_{j \in \mathcal{I}(t)} \exp(\beta u_j(t))}, \tag{20}$$

with large β making $w_i(t)$ concentrate on the top ranks. A tail-enrichment term can then be written as

$$\mathcal{L}_{\text{tail}} = -\sum_t \sum_{i \in \mathcal{I}(t)} w_i(t)\, y_i(t). \tag{21}$$

This objective rewards the model for placing positives into the soft queue. It is not a replacement for cross-entropy, because it does not supervise the full score field. It is a queue-facing correction that shapes the right tail.

Because not all positives are equally useful to capture, the tail objective should ideally weight them by review value. Let $g_i(t)$ summarize expected lead-time-adjusted benefit. Then

$$\mathcal{L}^{\text{val}}_{\text{tail}} = -\sum_t \sum_{i \in \mathcal{I}(t)} w_i(t)\, y_i(t)\, g_i(t). \tag{22}$$

This encourages the top queue to contain not just positives, but positives with favorable intervention opportunity. A patient with a tiny remaining actionable window does not contribute as much to the worklist's utility as one with slightly lower immediate risk but more room for useful response.

A second problem is queue instability caused by minor score perturbations. Since adjacent windows are highly overlapping, the model can produce jittery scores that lead to unnecessary worklist churn. A queue-aware stability term should focus on the upper ranks rather than the whole distribution [24]. Let $\widetilde{Q}_t$ be the soft queue induced by $w_i(t)$. One can penalize changes in queue membership mass:

$$\mathcal{L}_{\text{queue\_tv}} = \sum_t \sum_{i \in \mathcal{I}(t)} \big|w_i(t) - w_i(t-1)\big|\,(1 - n_i(t)), \tag{23}$$

where $n_i(t)$ measures the novelty of new evidence since the previous step. If nothing substantial has changed for a patient, then movement in and out of the soft queue is penalized. This does not enforce a static queue. It enforces that queue change be evidence justified.

Another challenge is repeated-window dependence. One deterioration episode produces many positive windows, often with similar labels and similar contexts. If all these windows are weighted equally, the model may learn broad positive plateaus instead of sharp prioritization. A useful response is to train on state transitions rather than raw windows alone. Let $\Delta y_i(t)$ indicate whether the patient moved into a more urgent pre-event regime relative to previous time. Then an auxiliary transition objective can encourage the score to anticipate upward priority motion:

$$\mathcal{L}_{\text{trans}} = -\sum_{i,t} \Delta y_i(t) \log \sigma(\eta_i(t) - \eta_i(t-1)). \tag{24}$$

This promotes rising trajectories before events rather than merely elevated plateaus. It is particularly suited to a prioritization system, where upward motion toward the queue boundary is often more useful than static high risk in patients already being watched [25].

A further refinement is pairwise ranking within the active census. For each time t, define pairs (i, j) with i a future positive of high review value and j a negative or low-value case currently competing for the same queue region. Then

$$\mathcal{L}_{\text{pair}} = \sum_t \sum_{(i,j) \in \mathcal{P}_t} \log\Big(1 + \exp(-(u_i(t) - u_j(t)))\Big). \tag{25}$$

This differs from generic ranking because the pairs are drawn from the same census-time context. The model is asked to solve the actual operational comparison it will face in deployment: which of these simultaneously present patients deserves the scarce slot?

The shape of the upper tail must also remain calibratable. Aggressive ranking objectives and focal weighting tend to distort probability scale, especially in the presence of severe imbalance. Since later sections require calibrated or at least controllably calibratable scores, one should include a mild score-shape regularizer. A useful choice is to penalize excessive compression or expansion of the upper tail. Let $q_\alpha(t)$ be the empirical α-quantile of scores at time t. Then a tail-shape control term can be written as

$$\mathcal{L}_{\text{shape}} = \sum_t \Big(q_{0.99}(t) - q_{0.95}(t) - m^\star\Big)^2, \tag{26}$$

where $m^\star$ is a target separation learned from validation utility. The point is not that one universal margin is correct, but that the model should not be allowed to flatten or overexplode the score gap among the highest-ranked patients arbitrarily.

Confidence should enter learning as well [26]. A queue slot spent on a weakly supported patient is more costly than one spent on a well-supported patient, because the chance of wasted review is higher. Therefore the confidence head $c_i(t)$ should be aligned with evidence quality and with future score reliability. One can use a support-consistency term

$$\mathcal{L}_{\text{conf}} = \sum_{i,t} \Big(c_i(t) - \rho_i(t) \Big)^2, \tag{27}$$

where $\rho_i(t)$ is a proxy built from freshness, documentation support, intervention saturation, and support-set familiarity. A second term can encourage confidence to be inversely related to realized calibration error or score instability in held-out strata. The net effect is that the model learns not only where a patient belongs in the worklist but how assertively that placement should be trusted.

The full objective then becomes explicitly service-aware:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_t \mathcal{L}_{\text{tail}}^{\text{val}} + \lambda_q \mathcal{L}_{\text{queue\_tv}} \\ + \lambda_p \mathcal{L}_{\text{pair}} + \lambda_r \mathcal{L}_{\text{trans}} + \lambda_s \mathcal{L}_{\text{shape}} \\ + \lambda_c \mathcal{L}_{\text{conf}} + \lambda_w \|\Theta\|_2^2. \end{aligned} \tag{28}$$

Such an objective is more complex than standard classification loss, but that complexity is justified by the complexity of the deployed task. The model is being asked not just to classify windows but to create a high-value, low-noise, confidence-aware worklist under resource constraints.

A final point concerns evaluation during training. Since the objective is queue facing, model selection should not rely on AUROC alone. Validation should examine event capture among the top K, mean lead time among queued positives, utility per queue slot, and queue turnover under realistic update frequencies. It is entirely plausible that the model with the highest AUROC is not the model with the highest expected worklist value [27]. This is not a paradox. It is the expected consequence of optimizing the wrong object if one uses ordinary metrics alone.

The main point of this section is that learning for ICU surveillance should be explicitly tail oriented and queue aware. The next step is to turn those learned scores into policies that spend attention economically, using calibration and selective escalation to translate continuous prioritization into actions that a real unit can sustain.

6. Selective Escalation and Review Economics

A worklist becomes useful only when it is coupled to a policy for spending attention. In intensive care that policy is constrained by two opposing facts. First, deteriorating patients benefit from timely recognition, which argues for sensitivity. Second, every chart review, bedside reassessment, and escalation consumes effort and contributes to fatigue, which argues for restraint. A continuous prioritization engine sits directly inside this tension. It should not force the entire burden of decision onto clinicians by showing an undifferentiated list, nor should it reduce everything to one brittle threshold. Instead it should support graded action under capacity and uncertainty.

Let the model produce a calibrated risk $\tilde{p}_i(t)$ and a confidence $c_i(t)$. The simplest graded policy uses both:

$$a_i(t) = \mathbb{I}\big(\tilde{p}_i(t) \geq \tau_p\big)\, \mathbb{I}\big(c_i(t) \geq \tau_c\big). \tag{29}$$

This captures only one action tier, but it already improves on ordinary thresholding because it distinguishes low confidence from low risk [28]. A more realistic policy defines several actions: no action, passive watchlisting, chart review, and active escalation. Denote these by $\mathcal{A} = \{0, 1, 2, 3\}$. A policy score can then be built as

$$\psi_i(t) = \tilde{p}_i(t)\, g_i(t)\, c_i(t) - \lambda_a \phi_a\big(a_i^{\text{ongoing}}(t)\big), \tag{30}$$

where $g_i(t)$ reflects expected intervention value and ϕ_a penalizes current attention saturation. Higher tiers correspond to higher $\psi_i(t)$ regions. In effect, the system ranks patients by expected marginal value of action rather than by event probability alone.

Review economics enters when one prices attention explicitly. Let c_{rev} be the cost of a chart review, c_{esc} the higher cost of escalation, and $b_i(t)$ the expected benefit of acting on patient i now. Then

$$\begin{aligned} U_i^{\text{rev}}(t) &= b_i^{\text{rev}}(t) - c_{\text{rev}}, \\ U_i^{\text{esc}}(t) &= b_i^{\text{esc}}(t) - c_{\text{esc}}. \end{aligned} \tag{31}$$

The system should choose the highest nonnegative utility tier subject to capacity. This framing helps explain why all high-risk cases should not receive identical treatment. A patient whose risk is high but whose marginal review benefit is low because the team is already deeply engaged may have $U_i^{\text{rev}}(t) \leq 0$. Another patient with slightly lower raw risk but large expected gain from early attention may have strongly positive utility. In a binary alarm regime these two cases are easily conflated.

Selective escalation also allows the model to express "review but not alert" states. These are especially valuable when confidence is intermediate [29]. A low-confidence but high-risk case can be kept near the top of the passive worklist without consuming the strongest escalation channel. This is more nuanced than suppressing the alert outright, because it preserves situational awareness while acknowledging evidence limitations. It also helps reduce false-alert burden in environments where not all top-ranked cases can be addressed immediately.

Capacity constraints can be encoded tier by tier:

$$\sum_i \mathbb{I}\big(a_i(t)=2\big) \le K_t^{(2)}, \qquad \sum_i \mathbb{I}\big(a_i(t)=3\big) \le K_t^{(3)}, \tag{32}$$

where $K_t^{(2)}$ and $K_t^{(3)}$ are review and escalation capacities. These capacities need not be fixed. They can vary by staffing, time of day, or concurrent unit burden. Therefore the thresholds separating action tiers should be dynamic or at least capacity aware. A unit under surge conditions may tighten the escalation boundary while preserving passive ranking across a broader set of patients.

Calibration is central here because action economics depend on the upper tail. If the model is overconfident in the queue-admitting region, the review queue becomes too large or too noisy. If it is underconfident, potentially valuable cases remain buried. A reliability-stratified calibration rule is therefore preferable to a global one. Let $b(i,t)$ index evidence-support strata [30]. Then

$$\tilde{p}_i(t) = \sigma\Big(\eta_i(t)/T_{b(i,t)}\Big), \tag{33}$$

or, more generally, $\tilde{p}_i(t) = g_{\text{cal}}(\eta_i(t), c_i(t))$ with monotonicity in $\eta_i(t)$. The point is that the same raw logit should not necessarily imply the same clinical probability under radically different confidence regimes.

Another crucial design element is action hysteresis. Suppose a patient enters the escalation tier at time t. Releasing that patient from the tier at $t+1$ because the score dipped slightly wastes cognitive effort and can create distrust. Therefore activation and release should use different boundaries. Let (τ^+, ξ^+) be the activation pair and (τ^-, ξ^-) the release pair, with lower release requirements. This creates a stable control regime in which genuine changes in state matter more than small score oscillations. Hysteresis is not a superficial interface trick. It is a mathematically grounded way to align dynamic scores with bounded human attention.

Queue economics also motivates a different interpretation of false positives. In classification, false positives are often counted symmetrically. In prioritization, their cost depends on what they displaced and what level of action they consumed. A false chart review is cheaper than a false escalation [31]. A false review during a quiet shift is cheaper than the same review when the queue is already saturated. Hence the effective false-positive penalty is context dependent:

$$\lambda_{\text{fp}}(t) = \lambda_0 + \lambda_1 \left(\frac{\lambda_t}{\mu_t}\right), \tag{34}$$

where λ_t/μ_t reflects current queue pressure. Under heavy overload, the system should become more conservative because each erroneous slot is more costly. This suggests that queue pressure itself should influence the action boundary, not just the score.

One can go further and couple passive ranking to active measurement. A high-risk, low-confidence patient may not justify escalation, but may justify ordering or prioritizing certain observations. In that case the model's output feeds back into the information structure rather than directly into clinical intervention. This is particularly valuable in ICU data because uncertainty often reflects missingness or stale documentation. A service-aware system can therefore recommend not just "review this patient" but "this patient is near the queue boundary and would benefit from refreshed evidence." Such behavior is impossible to specify cleanly when the model is treated as a simple alarm generator.

The economic view also makes it easier to reason about subgroup disparities. If one subgroup consistently has lower confidence because of sparser documentation, then fewer members of that subgroup may enter higher-cost review tiers even at similar raw risk. This is not automatically unfair, since acting on unsupported evidence can be harmful [32]. But it is a distributional fact that must be measured and governed. The combination of risk, confidence, and capacity determines actual access to attention. Thus fairness-like analysis should examine tier membership and review allocation, not only score calibration by subgroup.

The essence of this section is that continuous prioritization becomes useful only when connected to a selective, economically informed action policy. Confidence, calibration, capacity, and hysteresis translate a ranked list into a manageable escalation process. The final step is to ask how such a system can remain reliable under drift, feedback, and the long-term realities of deployment.

7. Drift, Feedback, and Long-Run Governance

A prioritization engine embedded in an ICU does not live in a static world. Its own outputs can alter care, documentation, and observation intensity. Unit practices change. Note templates shift. Staffing varies. Patients arrive from different services and with different patterns of measurement density. All of this means that the score field, the queue, and the utility of queue membership can drift over time [33]. For a binary classifier, this is already a problem. For a queue-facing continuous prioritizer, it is especially serious because even modest score drift can change who receives scarce attention.

A useful distinction is between score drift, queue drift, and policy drift. Score drift means the distribution of $\tilde{p}_i(t)$ or $u_i(t)$ changes. Queue drift means the composition or turnover of the active worklist changes. Policy drift means the same queue position leads to different real actions because human behavior or staffing changed. These three are related but not identical. A model may maintain roughly stable AUROC while the queue becomes noisier, because the upper tail widened. A queue may remain the same size while its action consequences change because reviewers respond differently. Governance should therefore monitor all three levels.

At the model level, one can track drift in the latent urgency state $z_i(t)$, in confidence $c_i(t)$, and in the joint distribution of top-ranked cases. Let μ_Q^{ref} be the reference mean embedding of queued patients from a stable development or calibration period. Then

$$D_Q(t) = \left\| \frac{1}{|Q_t|} \sum_{i \in Q_t} z_i(t) - \mu_Q^{\mathrm{ref}} \right\|_2 \tag{35}$$

gives a simple queue-state drift signal. A large value does not prove degradation, but it indicates that the system is prioritizing a different kind of case than before. Since the queue is the action-relevant subset, such a signal is often more useful than whole-census drift [34].

Confidence drift is equally important. If average confidence among queued patients drops while average risk stays similar, then the system is effectively spending more top slots on weaker evidence. This can happen under documentation changes, measurement sparsity shifts, or calibration decay. Monitoring the joint distribution of $(\tilde{p}, c)$ for the top ranks therefore provides an immediate sense of whether the queue is becoming less trustworthy even before enough outcome labels accumulate for full recalibration.

Feedback loops complicate every part of this picture. Once staff begin using the worklist, patients near the top may be measured more often, discussed more in notes, or treated earlier. This changes the very evidence distribution used by the model, especially for those cases most likely to drive evaluation. Without logging these responses, the system may be retrained on data partly generated by its own prior actions, creating a hidden policy dependence. A queue-aware system must therefore record not only scores and outcomes but downstream review events, measurement increases, note activity, and escalation responses. Otherwise one cannot distinguish a model failure from a successful early warning whose downstream intervention altered the observed outcome.

Recalibration under drift should respect the prioritization role of the system. It is often safer to recalibrate the upper tail and confidence-conditioned score map more frequently than to retrain the entire state model. Since the queue is sensitive to the top of the distribution, tail-local recalibration can restore operational meaning with less disruption to worklist behavior. More aggressive retraining should be bounded so that queue semantics do not change abruptly [35]. A constrained update rule

$$\begin{aligned} \theta_{t+1}^{\star} &= \arg\min_{\theta} \mathcal{L}_{\mathrm{recent}}(\theta) \\ \text{s.t.} \quad & \mathbb{E}\big[\|f_\theta(\mathcal{O}) - f_{\theta_t}(\mathcal{O})\|_2\big] \le \epsilon \end{aligned} \tag{36}$$

on a protected set limits sudden behavioral shifts. In worklist terms, this protects clinicians from waking up to a system whose score scale and queue occupancy changed substantially overnight without explanation.

Human oversight should be designed around the same principles. A ranked list is easier to govern than a black-box alarm if it is accompanied by evidence freshness, score trend, confidence, and perhaps action saturation. These quantities help users interpret why a patient is highly ranked and whether the rank should trigger aggressive action or watchful review. Importantly, such summaries should be framed as evidence support rather than causal explanation. The model is not proving that one measurement caused deterioration. It is indicating which current factors justify the patient's place in the queue.

A final governance issue concerns bounded autonomy. A system that manages scarce attention should not silently optimize away human judgment. It should propose, prioritize, and stabilize review effort while remaining transparent about confidence and support. This is particularly important because queue-based systems can feel more authoritative than ordinary alerts: a patient placed at the top of a list appears to have won a competition for

urgency [36]. The system should therefore make it easy to see when that placement is strong, when it is marginal, and when it is mostly a consequence of current census conditions rather than absolute certainty about deterioration.

The longer-term research implication is that surveillance engines should be benchmarked not only by out-of-sample prediction but by their behavior as socio-technical control elements. Queue occupancy, queue turnover, subgroup access to review tiers, response-dependent drift, and calibration under service constraints are all part of what the system becomes in practice. Ignoring them leads to a familiar outcome: a model that is locally impressive and globally awkward.

8. Conclusion

A deterioration system in the ICU is not most fundamentally a detector of binary events. It is a device for spending limited attention over a moving population of partially observed patients. Once that fact is taken seriously, the design problem changes shape. The most important score is not the one that wins an average benchmark comparison, but the one that creates a top-of-list whose members are genuinely worth a scarce review slot, whose ranking is stable enough to guide action, whose upper tail is calibrated enough to control workload, and whose uncertainty is explicit enough to prevent automation from overrunning evidence. This paper has treated that problem as one of service-aware continuous prioritization: a census-level queueing problem in which patient trajectories, documentary refresh, intervention saturation, confidence, and service capacity all influence what the model should emit and how the unit should consume it. Under this view, the value of continuous risk is not that it can be thresholded later if desired. Its value is that it naturally supports worklist formation, graded escalation, and capacity-aware control. That is why a queue-facing formulation is not a cosmetic rewrite of binary prediction but a more faithful account of what ICU surveillance systems are actually for. The most useful next generation of deterioration models will likely be those that optimize not just for identifying future events, but for creating ranked, support-aware, operationally stable streams of urgency that can coexist with human workload, adapt under drift, and remain intelligible as part of a living clinical system [37].

References

[1] S. Agrawal, V. Priya, R. George, S. V. Manirevu, T. Bhardwaj, R. Prajapati, S. Chakkrapani, M. S. Walker, V. P. Vaidya, and B. Narayanan, "Algorithm to derive progression-based lines of therapy from a real-world non-small cell lung cancer (nsclc) dataset.," *Journal of Clinical Oncology*, vol. 39, pp. e21172–e21172, 5 2021.

[2] T. Chomutare, K. Y. Yigzaw, A. Budrionis, A. Makhlysheva, F. Godtliebsen, and H. Dalianis, "Mie - de-identifying swedish ehr text using public resources in the general domain.," *Studies in health technology and informatics*, vol. 270, pp. 148–152, 6 2020.

[3] J. F. Vos, A. Boonstra, A. Kooistra, M. Seelen, and M. van Offenbeek, "The influence of electronic health record use on collaboration among medical specialties," 5 2020.

[4] S. E. Perlman, K. H. McVeigh, R. Newton-Dame, L. E. Thorpe, E. F. Snell, C. Chernov, J. Singer, and C. M. Greene, "Innovations in population health surveillance using electronic health record data," *Online Journal of Public Health Informatics*, vol. 6, 3 2014.

[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] K. Sharma, D. Parashar, O. Mengshetti, R. Ahmad, R. Mital, P. Singh, and M. Thawani, "Apache spark for analysis of electronic health records: A case study of diabetes management," *Revue d'Intelligence Artificielle*, vol. 37, pp. 1521–1526, 12 2023.

[7] H. Karami, A. Ionescu, and D. Atienza, "Point-process-based representation learning for electronic health records," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, vol. 30, pp. 1–4, IEEE, 10 2023.

[8] D. X. Yang and Y. A. Kim, "Patient-physician interactions and electronic health records," *JAMA*, vol. 310, pp. 1857–1857, 11 2013.

[9] E. Johnson and E. Roth, "Improving physician wellness through electronic health record education.," *International journal of psychiatry in medicine*, vol. 56, pp. 327–333, 7 2021.

[10] J. Longstaff, G. Capper, M. A. Lockyer, and M. G. Thick, "Ehr and epr confidentiality based on accountability and consent: tools for the caldicott guardian," *Health Informatics Journal*, vol. 6, pp. 45–52, 3 2000.

[11] E. Hatef, M. Rouhizadeh, I. Tia, E. Lasser, F. Hill-Briggs, J. Marsteller, and H. Kharrazi, "Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system (preprint)," 2 2019.

[12] A. Griesser and S. Bidmon, "A holistic view of facilitators and barriers of electronic health records usage from different perspectives: A qualitative content analysis approach.," *Health information management : journal of the Health Information Management Association of Australia*, vol. 53, pp. 227–236, 7 2023.

[13] A. M. Cogan, S. T. Rinne, M. Weiner, S. Simon, J. Davila, and E. M. Yano, "Using research to transform electronic health record modernization: Advancing a va partnered research agenda to increase research impacts.," *Journal of general internal medicine*, vol. 38, pp. 965–973, 10 2023.

[14] C. Liu, W. Zhang, B. C. Ooi, J. W. L. Yip, L. Zeng, and K. Zheng, "Toward cohort intelligence: A universal cohort representation learning framework for electronic health record analysis," 4 2023.

[15] A. Chou, J. K. Johnson, D. B. Jones, T. Euloth, B. A. Matcho, A. Bilderback, and J. K. Freburger, "Effects of an electronic health record-based mobility assessment and automated referral for inpatient physical therapy on patient outcomes: A quasi-experimental study.," *Health services research*, vol. 58 Suppl 1, pp. 51–62, 11 2022.

[16] M. Jane, "Towards the development of metamed: A blockchain-based ehr system integrated with a browser-based web application for interhospital communication and patient data ownership empowerment," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, pp. 360–372, IEOM Society International, 5 2023.

[17] M. C. Figgatt, J. Chen, G. Capper, S. Cohen, and R. Washington, "Chronic disease surveillance using electronic health records from health centers in a large urban setting," *Journal of public health management and practice : JPHMP*, vol. 27, pp. 186–192, 10 2019.

[18] S. Upadhyay and H.-F. Hu, "A qualitative analysis of the impact of electronic health records (ehr) on healthcare quality and safety: Clinicians' lived experiences.," *Health services insights*, vol. 15, pp. 11786329211070722–11786329211070722, 3 2022.

[19] E. Joukes, N. F. de Keizer, M. C. de Bruijne, A. Abu-Hanna, and R. Cornet, "Impact of electronic versus paper-based recording before ehr implementation on health care professionals' perceptions of ehr use, data quality, and data reuse," *Applied clinical informatics*, vol. 10, pp. 199–209, 3 2019.

[20] S. Seiedfarajollah, R. Safdari, M. Ghazisaeedi, and L. Keikha, "Key security and privacy issues from implementing the national electronic health record in the islamic republic of iran.," *Eastern Mediterranean health journal = La revue de sante de la Mediterranee orientale = al-Majallah al-sihhiyah li-sharq al-mutawassit*, vol. 25, pp. 656–659, 10 2019.

[21] J. Adler-Milstein, J. S. Adelman, M. Tai-Seale, V. L. Patel, and C. Dymek, "Ehr audit logs: A new goldmine for health services research?," *Journal of biomedical informatics*, vol. 101, pp. 103343–, 12 2019.

[22] 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.

[23] S. Singh, M. Rakhra, A. Malik, and D. Singh, "Blockchain-based ehr system for indian healthcare industry using aadhar," in *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pp. 997–1001, IEEE, 1 2023.

[24] L. Murali, G. Gopakumar, D. M. Viswanathan, and P. Nedungadi, "Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study.," *Journal of biomedical informatics*, vol. 143, pp. 104403–104403, 5 2023.

[25] V. ALCAN and C. DOĞRU, "Assessment of the health complaints among white-collar and blue-collar workers using the electronic health records," *Karaelmas İş Sağlığı ve Güvenliği Dergisi*, vol. 7, pp. 1–10, 4 2023.

[26] B. A. Goldstein, N. A. Bhavsar, M. Phelan, and M. J. Pencina, "Controlling for informed presence bias due to the number of health encounters in an electronic health record," *American journal of epidemiology*, vol. 184, pp. 847–855, 11 2016.

[27] M. Y. Garza, S. Myneni, S. H. Fenton, M. N. Zozus, M. Y. Garza, S. Fenton, and M. Zozus, "esource for standardized health information exchange in clinical research: A systematic review of progress in the last year," *Journal of the Society for Clinical Data Management*, vol. 1, 8 2021.

[28] D. C. Classen, C. A. Longhurst, T. Davis, J. A. Milstein, and D. W. Bates, "Inpatient ehr user experience and hospital ehr safety performance.," *JAMA network open*, vol. 6, pp. e2333152–e2333152, 9 2023.

[29] R. E. Jensen, K. S. Chan, J. P. Weiner, J. B. Fowles, and S. M. Neale, "Implementing electronic health record-based quality measures for developmental screening," *Pediatrics*, vol. 124, pp. e648–54, 10 2009.

[30] A. M. Davis, L. P. Hanrahan, A. F. Bokov, S. Schlachter, H. H. Laroche, and L. R. Waitman, "Patient engagement and attitudes toward using the electronic medical record for medical research: The 2015 greater plains collaborative health and medical research family survey (preprint)," 5 2018.

[31] 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.

[32] K. Honeyford, P. Expert, E. E. Mendelsohn, B. Post, A. A. Faisal, B. Glampson, E. K. Mayer, and C. E. Costelloe, "Challenges and recommendations for high quality research using electronic health records.," *Frontiers in digital health*, vol. 4, pp. 940330–940330, 8 2022.

[33] J. R. Blosnich and T. L. Boyer, "Concordance of data about sex from electronic health records and the national death index: Implications for transgender populations.," *Epidemiology (Cambridge, Mass.)*, vol. 33, pp. 383–385, 1 2022.

[34] Y. Huang, J. Guo, Z. Chen, J. Xu, W. T. Donahoo, O. Carasquillo, H. Adloori, J. Bian, and E. A. Shenkman, "The impact of electronic health records (ehr) data continuity on prediction model fairness and racial-ethnic disparities," 9 2023.

[35] M. D. Rozier, K. Patel, and D. A. Cross, "Electronic health records as biased tools or tools against bias: A conceptual model," *The Milbank quarterly*, vol. 100, pp. 134–150, 11 2021.

[36] R. Karmakar and A. Basu, *Implementation of a Reversible Watermarking Technique for Medical Images*, pp. 533–571. IGI Global, 6 2022.

[37] R. G. Mishuris, J. Yoder, D. Wilson, and D. M. Mann, "Integrating data from an online diabetes prevention program into an electronic health record and clinical workflow, a design phase usability study," *BMC medical informatics and decision making*, vol. 16, pp. 88–88, 7 2016.