Dynamic Risk-Set Geometry and Eligibility Drift in Hourly ICU Deterioration Forecasting Under Treatment and Discharge Censoring

Nguyen Duc Anh[1], Tran Minh Quang[2], and Le Hoang Phuc[3]

[1]Vinh University, Faculty of Information Technology, 182 Le Duan Street, Ben Thuy Ward, Vinh City 460000, Vietnam
[2]Quy Nhon University, Department of Computer Science, 170 An Duong Vuong Street, Nguyen Van Cu Ward, Quy Nhon 551000, Vietnam
[3]Can Tho University, College of Information and Communication Technology, Campus II, 3/2 Street, Ninh Kieu District, Can Tho 900000, Vietnam

Abstract

Intensive care warning systems are commonly evaluated as though they solve a straightforward repeated classification problem over patient-time windows. At each hour, the model reads the chart and estimates whether deterioration will occur within a future horizon. This setup is practical and has enabled broad methodological progress, yet it conceals a basic structural complication. The denominator of the task is moving. Patients enter and leave the set of clinically eligible windows through discharge, death, prior intervention, endpoint occurrence, and benchmark-specific exclusion rules. Consequently, a score assigned at one hour is not directly comparable to the same score assigned later in the stay unless the changing risk set is modeled explicitly. This paper develops a formal account of hourly deterioration forecasting as a dynamic risk-set problem rather than a static sequence of binary labels. The central argument is that much of the apparent difficulty, calibration drift, and transport instability of ICU warning systems arises from eligibility dynamics and policy-dependent censoring that are currently hidden inside benchmark construction. A multistate framework is introduced in which event onset, rescue intervention, ineligibility, and discharge are competing transitions over a latent severity trajectory. The paper then studies how repeated-window labeling distorts prevalence, how intervention-sensitive censoring reshapes observed risk, and how thresholding policies become unstable when the denominator contracts in a state-dependent manner. It concludes by arguing that warning models should be selected and interpreted through eligibility-aware likelihoods, risk-set-normalized ranking metrics, and counterfactual policy analyses rather than through pooled window-level discrimination alone.

1 The Moving Denominator

Most early warning studies in critical care begin with a table of windows [1]. Each row corresponds to a patient at a clock time, each row receives a label according to whether an adverse event happens within the next several hours, and the model is asked to estimate the probability of that label from the chart up to the current time [2]. The convenience of this representation is undeniable [3]. It turns a messy longitudinal record into a supervised learning problem with familiar losses and metrics. Yet the table format quietly hides a structural asymmetry. The rows are not independent samples from one stable population. They are samples from a population whose eligibility changes from hour to hour in ways that are tightly linked to latent severity, intervention policy, and local benchmark construction [4].

The simplest way to see the issue is to ask what it means for a patient to be at risk. A patient already receiving one component of the endpoint, such as vasopressor support or invasive ventilation, may no longer be eligible for some future event definitions. A patient discharged from the ICU is removed from the denominator entirely [5]. A patient who has crossed the endpoint may either disappear from future windows or remain under a modified label regime depending on the benchmark design. A patient who dies exits through a terminal absorbing state rather than through the ordinary continuation of the stay. These are not technical details appended after the fact. They define who is even counted when the model is judged [6]. The denominator is therefore not fixed. It is reconstituted at every prediction time by a collection of state-dependent rules [7].

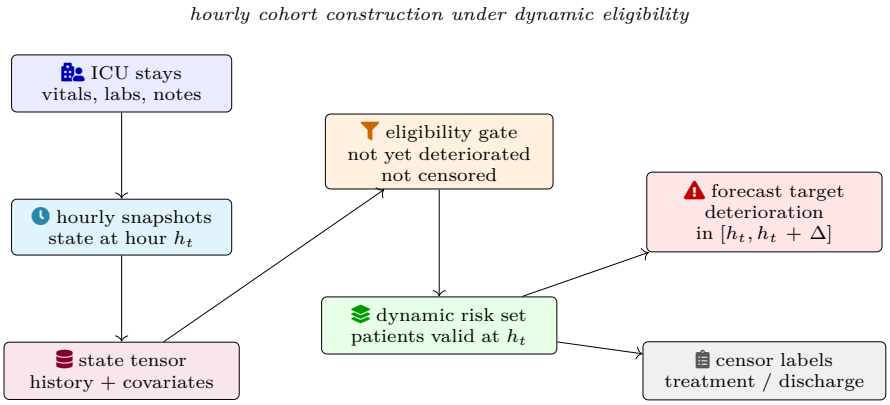

Figure 1: Cohort assembly for hourly ICU deterioration forecasting. Raw ICU trajectories are discretized into hourly states, screened for hour-specific eligibility, and only then admitted to the dynamic risk set from which deterioration labels and censoring labels are constructed.

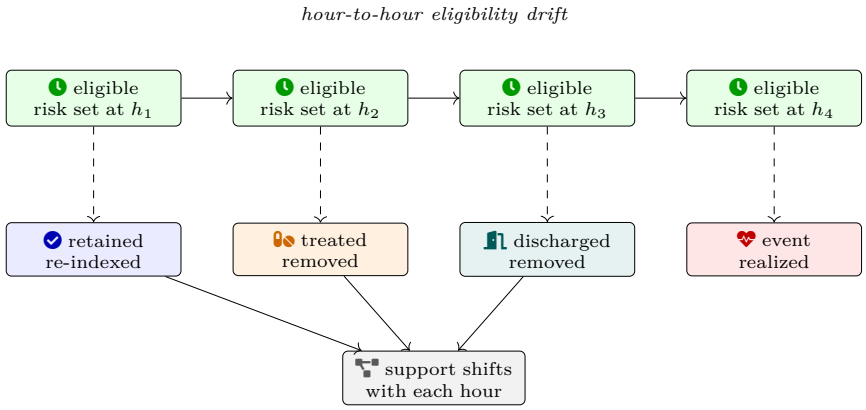

Figure 2: Eligibility is not static across hourly prediction times. As the clock advances, some patients remain in the analyzable set, whereas others leave because treatment has started, discharge has occurred, or deterioration has already happened; the geometry of the risk set therefore drifts even when the nominal prediction task remains the same.

This point is not merely conceptual. In the benchmark configuration described by Sadanandan (2025) [8], risk is refreshed on an hourly basis across ICU trajectories, which means that the prediction problem is inherently one of changing eligibility rather than one-time classification. Once that is recognized, the usual language of "window-level prediction" starts to look incomplete. The model is not simply asking whether an event will happen in the future. It is asking whether an event will happen while the patient remains in an eligible state and before competing transitions remove the patient from the forecast denominator. This changes the meaning of a score in subtle but important ways. A risk estimate is always conditional not only on observed severity but on continued membership in the benchmark's at-risk set.

If this conditioning were random or weakly related to the underlying disease process, the resulting bias might be modest. In the ICU it is not. Discharge is informative. Treatment onset is informative. Endpoint crossover is informative. Even the sampling rhythm of the chart often intensifies as clinicians approach a decision boundary. The patient who remains in the risk set at hour 60 is not exchangeable with the patient who left it at hour 18, and the remaining windows at hour 60 are not an unbiased continuation of the baseline cohort. They are the survivors of a sequence of state-dependent transitions [9]. A model evaluated on those windows is therefore evaluated on a moving denominator shaped by both patient evolution and institutional response.

The central contention of this paper is that many persistent puzzles in ICU warning systems become more intelligible once this moving denominator is brought to the center of the formulation. Calibration often drifts across stay time because the pool of eligible windows changes composition. Top-of-list precision can fluctuate unpredictably because the risk set contracts around a subset of patients with different event and intervention intensities. Models can appear to generalize within a retrospective dataset while actually learning denominators and censoring patterns specific to that environment. Thresholds that look clinically sensible in one unit can become erratic in another because the local timing of ineligibility differs even if latent severity patterns are similar. These are not independent problems. They are downstream consequences of treating a dynamic risk-set problem as though it were an ordinary repeated binary classification problem.

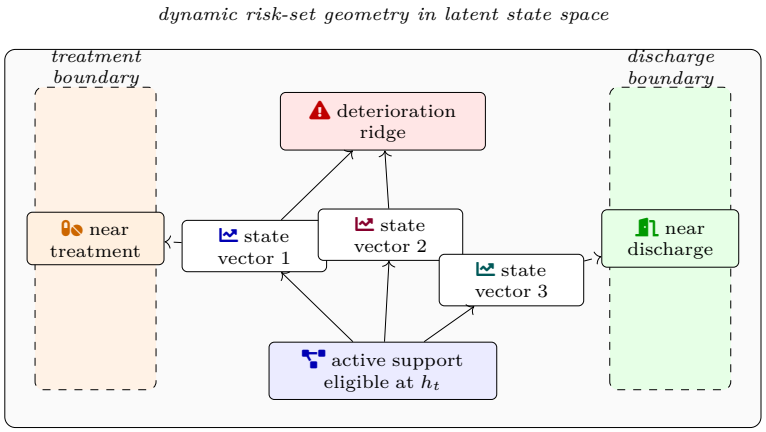

Figure 3: A geometric view of the hourly risk set. At any hour, eligible patient states occupy an active support region whose upper direction points toward imminent deterioration, while lateral erosion occurs near treatment and discharge boundaries; as patients cross those boundaries, the observed support changes shape before the next forecasting step.

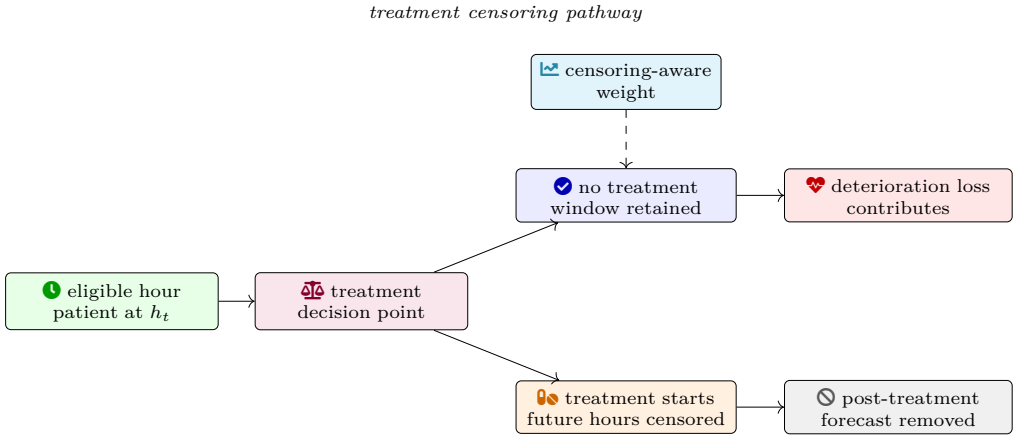

Figure 4: Treatment introduces informative censoring because the future trajectory after intervention is no longer comparable to the untreated deterioration process being forecast. The retained path can be weighted or otherwise adjusted for treatment propensity, while the post-treatment path is masked from subsequent untreated-risk supervision.

To make that argument precise, the following sections develop a sequence of increasingly explicit models. The next section defines hourly ICU forecasting through eligibility indicators and time-varying denominators. The subsequent sections examine how horizon labels are altered by treatment-sensitive censoring, how multistate formulations recover the hidden geometry of risk-set contraction, how estimation and thresholding should be redesigned under this view, and how deployment claims must be rewritten when pooled window metrics are no longer treated as self-explanatory. The conclusion then returns to the broader point: a warning model that ignores how patients leave the denominator is not just incomplete. It is solving the wrong statistical object.

2 Windows Are Not the Sample Space

Let patient i occupy the ICU over discrete times $t = 1, \ldots, T_i$, where one unit of time may be an hour. In the standard benchmark table, every eligible pair (i, t) becomes a data point [10]. A label $y_{i,t}^{(H)}$ is assigned according to whether the deterioration endpoint occurs within horizon H. A model produces $\hat{p}_{i,t}$, and a loss is computed row by row. What this representation suppresses is the fact that the sample space of valid rows is already filtered through a time-varying eligibility process.

Introduce an eligibility indicator $a_{i,t} \in \{0, 1\}$, where $a_{i,t} = 1$ if patient i is considered at risk and actionable for the endpoint at time t. Eligibility can fail for multiple reasons: prior endpoint occurrence, prior support onset, discharge, transfer, death, or benchmark-specific exclusions. The effective risk set at time t is then

$$\mathcal{R}_t = \{i : a_{i,t} = 1\}. \tag{1}$$

A patient-time row exists in the prediction table only if $i \in \mathcal{R}_t$. This means the data-generating process for windows is not the same as the data-generating process for patients. Windows are a censored and filtered projection of

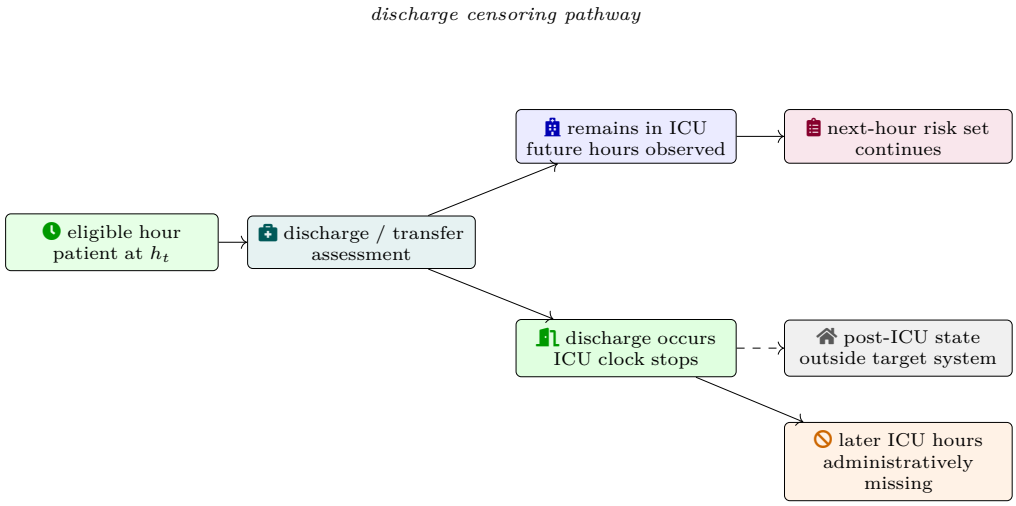

Figure 5: Discharge censoring differs from treatment censoring because the patient exits the ICU observation system altogether rather than undergoing a competing clinical intervention inside it. Once discharge occurs, later hourly windows are absent from the ICU risk process, which changes both the observable follow-up horizon and the composition of future risk sets.

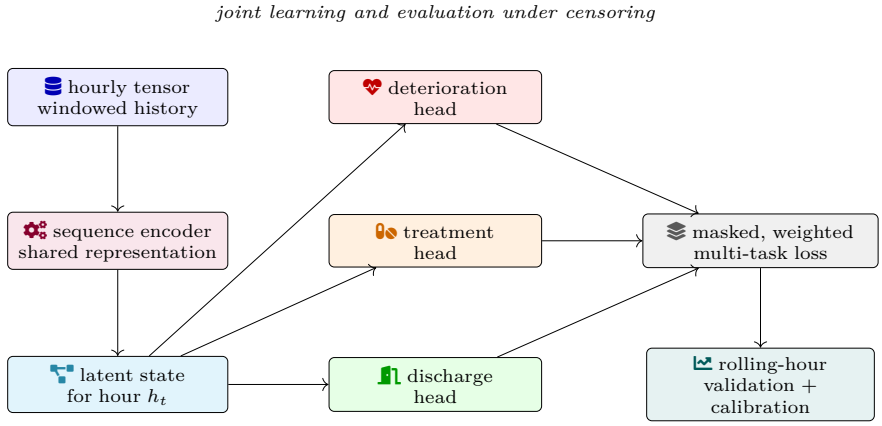

Figure 6: One practical modeling strategy is to learn a shared hourly state representation with separate heads for deterioration, treatment onset, and discharge. The loss can then be masked by eligibility and reweighted for censoring, while evaluation proceeds on rolling hourly risk sets using calibration-aware validation rather than a single static cohort split.

trajectories onto times when the benchmark declares them still meaningful.

The usual label can therefore be written more carefully as

$$y_{i,t}^{(H)} = a_{i,t}\,\mathbb{I}\Big(\exists u \in (t, t+H] : E_i(u) = 1\Big), \tag{2}$$

where $E_i(u)$ denotes endpoint onset. This expression already reveals a problem. The label is numerically binary, but semantically conditional. It only has meaning under ongoing eligibility at time t, and its event interpretation is entangled with whether the patient remains in the benchmark's universe of concern. Once $a_{i,t} = 0$, the model is no longer asked to reason about the future at all, even though the patient's latent severity may remain clinically important in a broader sense.

A more revealing decomposition introduces competing transitions. Let $E_i(t)$ denote the target endpoint process, $U_i(t)$ an ineligibility-inducing support or treatment process, $D_i(t)$ discharge or ICU exit, and $M_i(t)$ mortality if it is not already one endpoint component. Eligibility can then be written recursively as

$$a_{i,t+1} = a_{i,t}\,\mathbb{I}\big(E_i(t) = 0\big)\,\mathbb{I}\big(U_i(t) = 0\big)\,\mathbb{I}\big(D_i(t) = 0\big)\,\mathbb{I}\big(M_i(t) = 0\big). \tag{3}$$

This expression makes clear that the risk-set trajectory is itself a stochastic process. Once one patient receives an intervention earlier than another, or is discharged earlier, or dies earlier, the future sample space of windows changes [11]. Pooled metrics over all rows therefore average over a denominator that has been repeatedly thinned by informative transitions.

The difference between a patient-indexed and a window-indexed view can be quantified by considering the prevalence at each time. Let

$$\pi_t^{(H)} = \Pr\left(y_{i,t}^{(H)} = 1 \mid a_{i,t} = 1\right) \tag{4}$$

denote the time-specific prevalence inside the current risk set. There is no reason for $\pi_t^{(H)}$ to remain stable over t. Indeed, if sicker patients are removed early by endpoints or interventions and stable patients are discharged, then the remaining denominator can become enriched in ambiguous, prolonged, or policy-sensitive trajectories. The global prevalence reported for the whole benchmark is therefore a weighted average over many different underlying risk-set prevalences:

$$\bar{\pi}^{(H)} = \frac{\sum_t |\mathcal{R}_t|\, \pi_t^{(H)}}{\sum_t |\mathcal{R}_t|}. \tag{5}$$

This scalar is often cited as though it describes class imbalance in a simple sense. In fact it conflates many local prevalences arising in different eligibility contexts.

A second implication concerns repeated windows around the same trajectory. Suppose a patient remains eligible for many hours before an endpoint. Then the benchmark includes many rows from that one latent progression. Another patient may leave the denominator early by discharge or support onset and therefore contribute few rows. Standard training treats these rows as separate supervised instances, perhaps with some patient-level split discipline but usually without explicit adjustment for the geometry of repeated eligibility. The model can therefore receive disproportionate gradient from long eligible stretches, not because they are more clinically informative, but because the benchmark retained them longer.

The risk-set view also alters the interpretation of "false positives." A high score on a window that never becomes positive before the patient exits eligibility may not indicate that the model misread severity. It may indicate that the benchmark's event path was interrupted by a competing transition, such as discharge, rescue treatment, or policy-dependent support initiation that the label design handled differently [12]. Without an explicit model of the denominator, one cannot distinguish ordinary score error from eligibility-path mismatch. Both are absorbed into the same window-level metric.

The broader point is that the natural sample space of ICU warning is not the table of windows. It is the trajectory of patient membership in the at-risk set. Once the windows are recognized as observations emitted from that moving sample space, a different modeling agenda becomes necessary. The next section studies how fixed-horizon labels interact with this denominator geometry and why treatment-sensitive censoring can substantially reshape the object the model is actually learning.

3 Horizon Labels Under Informative Exit

A fixed future horizon is often justified as the operational definition of early warning. One wants to know whether something serious will happen in the next six, twelve, or twenty-four hours. This is reasonable. Yet once eligibility is dynamic, the horizon label does not simply indicate future event occurrence. It indicates future event occurrence before the patient exits the benchmark's risk set through competing transitions. That distinction is easy to overlook because it is rarely written into the definition of the label, but it is statistically central.

Let T_i^E be the time of endpoint onset, T_i^U the time of ineligibility-inducing support or treatment, T_i^D the time of discharge, and T_i^M the time of death when treated separately. Define the effective benchmark endpoint time as the earliest relevant transition:

$$T_i^\star = \min\{T_i^E, T_i^U, T_i^D, T_i^M\}. \tag{6}$$

The usual label at time t is then not purely about T_i^E [13]. It is really about whether T_i^E occurs before the horizon closes and before competing transitions remove the patient from the benchmark's notion of future:

$$y_{i,t}^{(H)} = \mathbb{I}\big(t < T_i^\star\big)\, \mathbb{I}\big(T_i^E - t \le H\big). \tag{7}$$

This appears innocuous until one notices that T_i^U and T_i^D are themselves informative about the latent severity process. A patient who receives rescue therapy early may be more unstable, not less. A patient discharged early may be truly improving. Thus, the failure of a window to become positive before exiting eligibility can mix benign and non-benign reasons.

The standard binary framework usually handles this by silence. It simply excludes post-exit windows and treats the pre-exit window labels as ground truth. But the score learned from these labels is then an estimate of a composite conditional probability:

$$p_{i,t}^{\text{obs}} = \Pr\left(T_i^E - t \le H,\, T_i^E < T_i^U,\, T_i^E < T_i^D,\, T_i^E < T_i^M \mid \mathcal{F}_{i,t}\right), \tag{8}$$

where $\mathcal{F}_{i,t}$ denotes the admissible history. This is not the same as the pure event probability

$$p_{i,t}^{E} = \Pr\left(T_i^E - t \le H \mid \mathcal{F}_{i,t}\right). \tag{9}$$

The benchmark target is therefore a cause-specific horizon probability modulated by competing exits. A model trained on $p_{i,t}^{\text{obs}}$ may be clinically useful, but its meaning is narrower than commonly assumed.

One can express this difference through discrete-time hazards. Let $\lambda_{i,t}^{E}$, $\lambda_{i,t}^{U}$, $\lambda_{i,t}^{D}$, and $\lambda_{i,t}^{M}$ be the one-step hazards of endpoint, support exit, discharge, and mortality conditional on current eligibility. Then the observed horizon probability becomes

$$\begin{aligned} p_{i,t}^{\text{obs}} = \sum_{\ell=1}^{H} \left[\prod_{r=1}^{\ell-1} \left(1-\lambda_{i,t+r-1}^{E}\right)\left(1-\lambda_{i,t+r-1}^{U}\right)\left(1-\lambda_{i,t+r-1}^{D}\right)\left(1-\lambda_{i,t+r-1}^{M}\right) \right] \\ \times \lambda_{i,t+\ell-1}^{E}. \end{aligned} \tag{10}$$

This expression makes explicit that the observed label depends on the entire competing-risk structure of the next H steps. If the support hazard rises under certain latent states, the observed endpoint risk may fall even while underlying clinical instability rises, simply because those states are more likely to leave the denominator through treatment first.

This has two important consequences [14]. First, a model can appear poorly calibrated with respect to pure event risk when in fact it is well calibrated for the benchmark's competing-risk target. Second, a model can appear to overpredict deterioration in regimes where timely rescue is common, because the benchmark counts rescue as censoring rather than as evidence that deterioration was clinically relevant. The model is then penalized for recognizing severity in a state space where the label definition chooses a different exit path.

Horizon length interacts strongly with this phenomenon. For short horizons, the distinction between event risk and observed benchmark risk may be moderate if competing exits are relatively rare within the next few hours. For long horizons, competing exits accumulate. Then the score required to optimize the binary loss is increasingly shaped by who remains eligible rather than only by who is close to the endpoint. A longer horizon can therefore cause the model to internalize policy-specific exit patterns as if they were part of the endpoint's natural anticipatory signal.

This is one reason thresholds often behave unpredictably across institutions. A threshold fitted to one unit's observed horizon probabilities is fitted to one unit's balance of event and competing-exit hazards. If another unit discharges earlier, initiates support earlier, or has different composite endpoint structure, the same threshold no longer corresponds to the same clinical state. What appears as threshold miscalibration may be denominator misalignment.

The natural remedy is not to avoid horizons altogether. It is to model the horizon target as a derived functional of multiple exit hazards rather than as a primitive binary fact. That move is developed next through a multistate view of ICU stays, where eligibility transitions are modeled jointly with deterioration [15].

4 A Multistate View of ICU Warning

A more faithful statistical object for hourly ICU warning is a multistate process rather than a sequence of independent binary windows. In a multistate formulation, the patient occupies one of several latent or observed states and moves among them according to transition intensities that may depend on clinical history, interventions, and environment. Warning is then derived from the chance of entering clinically relevant states before competing departures, rather than being trained directly on a pooled table of future labels.

At minimum, one can define an eligible state R for "at risk and under monitoring," an endpoint state E for the target deterioration event, a support-induced ineligibility state U, a discharge state D, and a mortality state M. The patient begins in R and transitions to one of the others. The one-step transition matrix at time t can be written as

$$P_{i,t} = \begin{pmatrix} 1-\lambda_{i,t}^{E}-\lambda_{i,t}^{U}-\lambda_{i,t}^{D}-\lambda_{i,t}^{M} & \lambda_{i,t}^{E} & \lambda_{i,t}^{U} & \lambda_{i,t}^{D} & \lambda_{i,t}^{M} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{11}$$

with the absorbing states E, U, D, M shown for simplicity. More elaborate formulations could allow return from temporary support, step-down transfer, or partial endpoint states, but even this minimal process already captures the denominator geometry hidden by the benchmark table.

The hourly warning score can now be defined as a functional of this multistate system. If the clinical intention is to warn about endpoint E before other exits, then the score is the finite-horizon transition probability

$$p_{i,t}^{R\to E}(H) = \Pr\big(X_{i,t+\ell} = E \text{ for some } 1 \le \ell \le H \mid X_{i,t} = R,\, \mathcal{F}_{i,t}\big), \tag{12}$$

which is exactly the competing-risk probability described earlier. If the intention is instead to quantify total impending instability, then one might define a broader risk

$$p_{i,t}^{R\to\{E,U,M\}}(H) = \Pr\big(X_{i,t+\ell} \in \{E,U,M\} \text{ for some } 1 \le \ell \le H \mid X_{i,t} = R,\, \mathcal{F}_{i,t}\big). \tag{13}$$

This distinction is clinically important. The former corresponds to benchmark-defined endpoint occurrence. The latter corresponds more closely to "the patient is on a serious trajectory even if rescue or death intervenes before the benchmark event." A standard binary label conflates these possibilities depending on how endpoints and censoring are designed [16].

The multistate view also supports richer latent structure. The eligible state R can be split into substates indexed by severity or stage of instability. Rescue-sensitive substates can be added to distinguish temporary improvement under intervention from genuine recovery. Support states can be typed by intervention channel. The score shown to clinicians can remain simple, but the hidden representation no longer has to pretend that all future pathways are equivalent.

One can further define transition intensities through a shared latent severity process $z_{i,t}$. For example,

$$\begin{aligned}
\lambda_{i,t}^{E} &= \exp\big(\alpha_E^\top z_{i,t} + \beta_E^\top u_{i,t}\big),\\
\lambda_{i,t}^{U} &= \exp\big(\alpha_U^\top z_{i,t} + \beta_U^\top u_{i,t}\big),\\
\lambda_{i,t}^{D} &= \exp\big(\alpha_D^\top z_{i,t} + \beta_D^\top u_{i,t}\big),\\
\lambda_{i,t}^{M} &= \exp\big(\alpha_M^\top z_{i,t} + \beta_M^\top u_{i,t}\big).
\end{aligned} \tag{14}$$

The same latent severity may therefore increase endpoint risk and support risk simultaneously. A direct binary model cannot represent this without ambiguity. The multistate model can. It can say that the patient is highly unstable, but the most likely next transition in this institution may be support onset rather than the benchmark endpoint. This is often exactly the information a clinician would want.

Another advantage concerns explanation. In the standard setup, a high score is simply "high risk." In a multistate model, the high score can be decomposed into hazard contributions. One patient may be high because endpoint hazard dominates. Another may be high because support hazard is large and endpoint hazard moderate, indicating a rescue-prone but clinically serious state [17]. This decomposition can clarify whether the model is recognizing latent severity or merely chasing the precise benchmark event boundary.

The multistate formulation also changes how missingness and note timing should be used. Observation density may affect support initiation hazard differently from discharge hazard. Textual concern may be more predictive of an upcoming intervention than of death. Rather than forcing all evidence channels into one scalar pathway, the model can let them contribute differently to different transitions. This reduces the temptation to treat documentation or treatment history as generic shortcuts to the one binary label.

Most importantly, the multistate view makes clear that a benchmark label is a projection of a richer process. The label is not wrong. It is incomplete. Once the richer process is modeled, both estimation and evaluation need to change. That is the topic of the next section.

5 Estimation, Calibration, and the Upper Tail Under Eligibility Drift

The prevailing estimation strategy in ICU deterioration work is to optimize a window-level loss such as binary cross-entropy or focal loss over all eligible rows. This is computationally attractive and often yields strong discrimination. Yet once the outcome is understood as a multistate transition probability, the usual loss is seen to be only a surrogate for a richer likelihood. Moreover, because the denominator changes across time and trajectories, standard calibration procedures can become misleading if they ignore the composition of the underlying risk set [18].

The natural likelihood under the multistate model is a product over transition events rather than over binary window labels. For the simple absorbing-state process above, the contribution of patient i at time t while still in state R is

$$\begin{aligned}
\mathcal{L}_{i,t} = &\big(\lambda_{i,t}^{E}\big)^{\delta_{i,t}^{E}} \big(\lambda_{i,t}^{U}\big)^{\delta_{i,t}^{U}} \big(\lambda_{i,t}^{D}\big)^{\delta_{i,t}^{D}} \big(\lambda_{i,t}^{M}\big)^{\delta_{i,t}^{M}}\\
&\times \big(1 - \lambda_{i,t}^{E} - \lambda_{i,t}^{U} - \lambda_{i,t}^{D} - \lambda_{i,t}^{M}\big)^{1-\delta_{i,t}^{\text{any}}},
\end{aligned} \tag{15}$$

where $\delta_{i,t}^{E}$ etc. indicate which transition occurred in the next step and $\delta_{i,t}^{\text{any}} = \delta_{i,t}^{E} + \delta_{i,t}^{U} + \delta_{i,t}^{D} + \delta_{i,t}^{M}$. The full negative log-likelihood then aligns estimation with actual transition structure rather than with the collapsed binary horizon label. Horizon probabilities can still be derived afterward.

This likelihood has several desirable properties. It avoids backward-smearing the same positive label across many hours. It directly models the competing transitions that thin the denominator. And it encourages the model to allocate probability mass among clinically distinct exits rather than forcing all of them into a generic notion of future badness. The price is that one needs reliable transition-time data and, ideally, a clear operational definition of each absorbing or semi-absorbing state. In many ICU corpora those ingredients are available to a greater extent than the literature's common binary framing suggests.

Calibration becomes more interesting under this view. Suppose a model produces $\hat{p}_{i,t}^{R\to E}(H)$ and this score is numerically calibrated against observed event frequencies in the benchmark world. That does not imply that the same score is calibrated for the broader notion of clinical deterioration or for another unit with different competing-risk structure. A calibrated score in one denominator can be miscalibrated in another even when the underlying latent severity mapping is stable. The relevant notion of calibration is therefore denominator-aware. One should ask not only whether predicted probabilities match observed frequencies overall, but whether they do so conditional on time in stay, current eligibility pattern, and competing-exit intensity [19].

Let $\mathcal{B}_g$ denote a subgroup of windows defined by a risk-set feature g, such as late-stay membership, high support hazard, or imminent discharge risk. A denominator-aware calibration error is

$$\mathrm{CE}_g = \left| \frac{1}{|\mathcal{B}_g|} \sum_{(i,t)\in\mathcal{B}_g} \hat{p}_{i,t}^{R\to E}(H) - \frac{1}{|\mathcal{B}_g|} \sum_{(i,t)\in\mathcal{B}_g} y_{i,t}^{(H)} \right|. \tag{16}$$

Such slice-wise calibration can reveal that a model is well calibrated overall but systematically overestimates endpoint risk in high-rescue regimes or underestimates it in prolonged eligible trajectories that survive many hours without event. Global calibration curves can easily average these distortions away.

The upper tail is also affected by eligibility drift. In low-prevalence tasks, most operational utility lies in a small number of top-ranked windows. But those top-ranked windows are drawn from a denominator whose composition changes throughout the stay. A model that assigns high scores broadly to persistent eligible trajectories may occupy the upper tail with long plateau windows rather than with imminent transitions. Another model that sharply estimates transition intensities may place fewer but more action-relevant windows at the top. Global AUROC may not distinguish these regimes sharply. Top-K queue utility does.

A useful upper-tail metric under the multistate formulation is transition capture among the active risk set:

$$\mathrm{TC}_K = \frac{\sum_t \sum_{i\in Q_K(t)} \delta_{i,t}^{E}}{\sum_t \sum_{i\in\mathcal{R}_t} \delta_{i,t}^{E}}, \tag{17}$$

where $Q_K(t)$ is the set of top-K patients at time t ranked by the chosen score. One can analogously define capture for support or mortality transitions. This makes visible whether the queue concentrates endpoint-relevant imminence or merely general severity. The distinction is crucial when staffing can support only a small number of reviews [20].

Threshold selection should likewise be made conditional on denominator dynamics. A score threshold τ that yields ten alerts per shift in one unit may yield very different alert burden in another if the size and composition of $\mathcal{R}_t$ differ. The obvious remedy is to use quantile-based or capacity-aware thresholds. But even then, one needs to know what occupies the top quantiles as eligibility changes. A multistate model can support this by exposing which transition hazards dominate the score. A direct binary model typically cannot.

There is thus a strong argument that estimation and calibration should be driven by the multistate object, while binary horizon scores should be treated as downstream summaries. Doing so does not complicate deployment unnecessarily. It clarifies what the score means and why it may fail when institutional exit patterns shift. The next section addresses how this reframing alters scientific claims and deployment logic.

6 What a Score Means When the Denominator Moves

A score from a conventional benchmark-trained model is usually interpreted as a probability of deterioration within the next horizon. Under the dynamic risk-set view, that interpretation is too blunt. The score is a probability of a specific transition before competing exits remove the patient from the benchmark's notion of the future. That meaning is narrower and more conditional than the language of "imminent deterioration" suggests. Scientific claims, threshold decisions, and user interfaces should reflect that conditionality rather than conceal it.

One immediate implication concerns transfer across hospitals [21]. If two units have different discharge timing, different support thresholds, or different composite endpoint structure, then the same latent severity state may

map to different observed horizon labels. A score calibrated in one denominator may therefore mislead in another even when the underlying model still recognizes severe illness. This means that external validation should not ask only whether the global metric changed. It should ask whether the transition mapping or the latent severity mapping changed. If most of the drift lies in support and discharge hazards, then recalibrating the exit model may restore utility without altering the core severity representation.

A second implication concerns how to communicate risk. A clinician may reasonably hear "30% risk in the next 12 hours" as a claim about the patient, not about the benchmark denominator. A multistate system can be more honest. It can report, for example, a 30% probability of endpoint E before competing rescue or discharge, alongside a broader probability of any serious transition. The latter may better capture what clinicians often mean by "this patient is going bad" even when the exact benchmark endpoint never occurs. Such distinctions are not pedantic. They help align model outputs with clinical interpretation.

A third implication concerns evaluation fairness across slices. If some subpopulations are more likely to leave eligibility through support onset or discharge, then direct comparison of pooled binary performance across those subpopulations can be misleading. Apparent underperformance may partly reflect denominator structure rather than failure to recognize severity [22]. Conversely, apparent overperformance may reflect a benchmark whose label aligns unusually well with one group's exit pathways. Slice-aware evaluation under the multistate view can expose these effects, whereas ordinary subgroup AUROC may not.

A fourth implication concerns what should count as a false alarm. In a dynamic risk-set problem, a high score followed by rescue support rather than benchmark endpoint may not be evidence that the model was wrong. It may be evidence that the model recognized severe latent state whose first recorded consequence was a competing transition. Whether this should be operationally penalized depends on the deployment goal. If the goal is strict endpoint anticipation, then yes. If the goal is early identification of clinically serious trajectories, then the penalty should be softened. This cannot be decided coherently without acknowledging the multistate structure.

These issues suggest a different philosophy of deployment. Rather than selecting one model and one threshold on the basis of pooled binary metrics, one should consider a family of scores and policies derived from the multistate system. One score may rank pure endpoint imminence. Another may rank total near-term instability including rescue transitions. One threshold may support interruptive alerts. Another may support passive review ranking [23]. Such differentiation is not needless complexity. It is a direct consequence of the fact that the denominator is moving and the endpoint is not the only clinically meaningful future.

A final implication is epistemic. Once the benchmark is seen as a dynamic risk-set problem, the field's familiar confidence in a few scalar metrics should soften. A good AUROC or AUPRC remains valuable, but it no longer has the same interpretive scope. It says the model ranks benchmark-defined futures well inside one eligibility geometry. It does not by itself say that the model has learned a portable notion of deterioration. That stronger claim requires understanding how the score would change if the denominator changed.

7 Conclusion

Much of ICU deterioration forecasting has been organized around a picture that is both elegant and too simple. The picture is this: at each hour the model reads the chart, predicts whether deterioration lies ahead, and is judged by how well that prediction matches the eventual binary label. This formulation has been productive because it standardizes comparison and makes difficult data tractable. Yet it also strips away one of the most consequential properties of the problem, namely that the population to which the model is applied is changing continuously through state-dependent exits. Patients are not merely accumulating toward an event or away from it. They are also leaving the benchmark's future through discharge, rescue, intervention onset, or other transitions that alter who still counts when the model is scored. That changing denominator is not a marginal technicality [24]. It is part of the task.

The paper has tried to show that once this denominator is modeled explicitly, several persistent sources of confusion become easier to understand. The first is calibration drift. A score can look well calibrated in aggregate and still mean different things at different stay times or under different local exit intensities because the composition of the at-risk set is changing. The second is transfer instability. Two hospitals with similar patient populations can still induce different score meanings if they differ in discharge timing, support initiation thresholds, or endpoint composition. The third is interpretive ambiguity. A supposedly "false positive" may in fact be a high-risk state diverted into a competing transition rather than an outright misread of patient severity. The fourth is queue behavior. Top-ranked windows can become dominated by persistent eligible plateaus rather than imminent transitions if the model is trained only on backward-smeared horizon labels.

These are not isolated annoyances. They all arise from the same hidden simplification: the assumption that the window table is the task. It is not. The table is a derived representation of a multistate process. Once that is acknowledged, a more faithful statistical language becomes available [25]. Eligibility indicators replace the fiction of a fixed sample space. Cause-specific hazards replace the flattening effect of one binary target. Rescue, discharge,

and endpoint onset become competing transitions rather than afterthoughts. Horizon probabilities become derived functionals of those transitions rather than primitive objects of learning. And calibration becomes an attribute of a particular denominator geometry rather than of one timeless risk scale.

This change in language matters because it changes what the field should treat as progress. A larger encoder or a more refined fusion mechanism may still be useful, but only if it is aligned with the right object. A model that wins on pooled binary metrics by fitting one benchmark's denominator structure more tightly has not necessarily learned more about impending clinical decline. It may have learned more about how one benchmark retains or removes windows. Conversely, a model that appears modestly weaker on ordinary metrics but correctly separates endpoint hazard from rescue hazard may be much more informative and portable. Once the problem is framed as dynamic risk-set geometry, those distinctions become visible.

There is also a practical lesson about what ought to be reported. A complete evaluation should not end with AUROC, AUPRC, and one calibration curve. It should report how the at-risk population evolves through the stay, how prevalence changes conditional on eligibility, how much of the model's score is driven by endpoint versus competing-exit hazard, and how thresholds behave under plausible changes in the denominator. It should make clear whether the system is built to predict a benchmark-defined endpoint, a broader instability process, or some combination of the two [26]. Without that clarity, users are asked to interpret one scalar score as though it referred to one coherent clinical future, when in fact the score may be conditional on several hidden exit mechanisms.

The broader methodological implication is that target design and denominator design deserve much more attention than they usually receive. In the current literature, endpoint definitions are scrutinized more than eligibility geometry. Yet who remains in the benchmark is as important as what event counts once they are there. A benchmark that removes patients aggressively after certain interventions is teaching one task. A benchmark that retains them is teaching another. A horizon that spans many likely competing exits is teaching a different risk object from a horizon confined to near-term transition windows. These are substantive design choices, not administrative ones.

If there is a single idea worth carrying forward, it is that ICU warning systems are not only predictors of future events. They are predictors over moving risk sets. A model that does not know this may still succeed statistically, but it cannot fully explain what it is predicting or why that prediction might fail elsewhere. A model that does know it can separate latent severity from exit timing, differentiate endpoint risk from rescue risk, and support operational policies that are explicit about which future they care about. That is a more demanding modeling program, but it is also a more truthful one. In critical care, where predictions are consumed under scarce attention and where the line between deterioration and intervention is often institutionally drawn, truthfulness about the denominator is not optional. It is part of what it means for a warning system to be scientifically and clinically interpretable [27].

References

[1] M. Z. Ahmed and C. Mahesh, "A weight based labeled classifier using machine learning technique for classification of medical data," *Revue d'Intelligence Artificielle*, vol. 35, no. 1, pp. 39–46, Feb. 28, 2021. DOI: 10.18280/ria.350104

[2] J. White et al., *Supplementary material for: Developing and testing electronic health record-derived caries indices*, Jan. 1, 2019. DOI: 10.6084/m9.figshare.8230535

[3] R. R. Aguirre, O. Suarez, M. Fuentes, and M. A. Sanchez-Gonzalez, "Electronic health record implementation: A review of resources and tools.," *Cureus*, vol. 11, no. 9, e5649–, Sep. 13, 2019. DOI: 10.7759/cureus.5649

[4] A. Naveed, T. Sigwele, Y. F. Hu, M. Kamala, and M. Susanto, "Addressing semantic interoperability, privacy and security concerns in electronic health records," *Journal of Engineering and Scientific Research*, vol. 2, no. 1, pp. 31–38, Dec. 30, 2020. DOI: 10.23960/jesr.v2i1.40

[5] A. K. Saiod, D. van Greunen, and A. Veldsman, *Handbook of Large-Scale Distributed Computing in Smart Healthcare - Electronic Health Records: Benefits and Challenges for Data Quality*. Springer International Publishing, Aug. 8, 2017. DOI: 10.1007/978-3-319-58280-1_6

[6] S. Chenthara, K. Ahmed, H. Wang, F. Whittaker, and Z. Chen, "Healthchain: A novel framework on privacy preservation of electronic health records using blockchain technology.," *PloS one*, vol. 15, no. 12, e0243043–, Dec. 9, 2020. DOI: 10.1371/journal.pone.0243043

[7] D. Leander, "It's in the ehr…but where?: Patient records," in Productivity Press, May 31, 2019, pp. 147–149. DOI: 10.4324/9780429031403-37

[8] 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.

[9] Z. Tang et al., "Icic (2) - embracing disease progression with a learning system for real world evidence discovery," in Germany: Springer International Publishing, Oct. 5, 2020, pp. 524–534. DOI: 10.1007/978-3-030-60802-6_46

[10] N. A. Bhavsar, A. Gao, M. Phelan, N. J. Pagidipati, and B. A. Goldstein, "Value of neighborhood socioeconomic status in predicting risk of outcomes in studies that use electronic health record data," *JAMA network open*, vol. 1, no. 5, e182716–, Sep. 7, 2018. DOI: 10.1001/jamanetworkopen.2018.2716

[11] C.-H. Ho, H.-C. Wene, C.-M. Chu, Y.-S. Wu, and J.-L. Wang, "Importance-satisfaction analysis for primary care physicians' perspective on ehrs in taiwan," *International journal of environmental research and public health*, vol. 11, no. 6, pp. 6037–6051, Jun. 6, 2014. DOI: 10.3390/ijerph110606037

[12] N. K. Chandra, A. Sarkar, J. F. de Groot, Y. Yuan, and P. Müller, "Bayesian nonparametric common atoms regression for generating synthetic controls in clinical trials," *Journal of the American Statistical Association*, vol. 118, no. 544, pp. 2301–2314, Jul. 26, 2023. DOI: 10.1080/01621459.2023.2231581

[13] C. Violán et al., "Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity.," *BMC public health*, vol. 13, no. 1, pp. 251–251, Mar. 21, 2013. DOI: 10.1186/1471-2458-13-251

[14] L. Timmins, L. M. Kern, A. S. O'Malley, C. Urato, A. Ghosh, and E. Rich, "Communication gaps persist between primary care and specialist physicians.," *Annals of family medicine*, vol. 20, no. 4, pp. 343–347, Jul. 25, 2022. DOI: 10.1370/afm.2781

[15] M. Tian, B. Chen, A. Guo, S. Jiang, and A. R. Zhang, *Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models*, Oct. 23, 2023. DOI: 10.48550/arxiv.2310.15290

[16] D. Yu et al., "Machine learning prediction of the adverse outcome for nontraumatic subarachnoid hemorrhage patients.," *Annals of clinical and translational neurology*, vol. 7, no. 11, pp. 2178–2185, Sep. 29, 2020. DOI: 10.1002/acn3.51208

[17] M. VanBeek et al., "The completeness and accuracy of dataderm™: The database of the american academy of dermatology (aad)," *Journal of the American Academy of Dermatology*, vol. 86, no. 2, pp. 394–398, Jun. 12, 2021. DOI: 10.1016/j.jaad.2021.06.024

[18] R. K. Saripalle, C. Runyan, and M. Russell, "Using hl7 fhir to achieve interoperability in patient health record.," *Journal of biomedical informatics*, vol. 94, pp. 103 188–103 188, May 4, 2019. DOI: 10.1016/j.jbi.2019.103188

[19] L. M. Kern and R. Kaushal, "Electronic health records and ambulatory quality," *Journal of general internal medicine*, vol. 28, no. 9, pp. 1133–1133, May 18, 2013. DOI: 10.1007/s11606-013-2475-4

[20] J. R. Robinson, W.-Q. Wei, D. M. Roden, and J. C. Denny, "Defining phenotypes from clinical data to drive genomic research.," *Annual review of biomedical data science*, vol. 1, no. 1, pp. 69–92, Apr. 25, 2018. DOI: 10.1146/annurev-biodatasci-080917-013335

[21] W. Zhou, D. Bitterman, M. Afshar, and T. A. Miller, *Considerations for health care institutions training large language models on electronic health records*, Aug. 24, 2023. DOI: 10.48550/arxiv.2309.12339

[22] M. Noaeen et al., "Unlocking the power of ehrs: Harnessing unstructured data for machine learning-based outcome predictions.," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2023, pp. 1–4, Jul. 24, 2023. DOI: 10.1109/embc40787.2023.10340232

[23] R. D. Wijaya and L. O. A. Rahman, "Implementasi electronic health records (ehrs) pada pelayanan kesehatan di komunitas: Literature review," *Jurnal Kesehatan*, vol. 8, no. 1, pp. 28–38, Feb. 6, 2020. DOI: 10.46815/jkanwvol8.v8i1.39

[24] A. Acharya et al., *Clinical risk prediction using language models: Benefits and considerations*, Nov. 29, 2023. DOI: 10.48550/arxiv.2312.03742

[25] C. Pawlowski, A. Venkatakrishnan, J. C. O'Horo, A. D. Badley, J. Halamka, and V. Soundararajan, *Covid-associated pediatric hospitalization and icu admission trends across a multi-state health system and the broader us population*, Apr. 5, 2021. DOI: 10.1101/2021.04.02.21254593

[26] C. Papoutsi, J. E. Reed, C. Marston, R. Lewis, A. Majeed, and D. Bell, "Patient and public views about the security and privacy of electronic health records (ehrs) in the uk: Results from a mixed methods study," *BMC medical informatics and decision making*, vol. 15, no. 1, pp. 86–86, Oct. 14, 2015. DOI: 10.1186/s12911-015-0202-2

[27] L. J. Beesley and B. Mukherjee, *Bias reduction and inference for electronic health record data under selection and phenotype misclassification: Three case studies.* Dec. 23, 2020. DOI: 10.1101/2020.12.21.20248644