Scientific Journals
An international open-access platform

Hanlin Letters

Article

OPEN ACCESS

Calibration-Aware Multimodal Forecasting of ICU Deterioration Under Irregular Sampling and Documentation Sparsity

Nguyen Duc Huy,[1] Pham Quang Minh[2] Le Thanh Nam[3]

[1] Faculty of Information Technology, Hung Yen University of Technology and Education, Đường Nguyễn Văn Linh 555, Phường Hiến Nam, Hưng Yên, Vietnam;
[2] Department of Computer Engineering, Vinh University of Technology Education, Đường Lê Duẩn 117, Phường Bến Thủy, Vinh, Vietnam;
[3] School of Computing and Information Systems, Thai Nguyen University of Information and Communication Technology, Đường Z115, Phường Tân Thịnh, Thái Nguyên, Vietnam;

Abstract

Intensive care units produce high-velocity streams of structured physiologic measurements and lower-frequency clinical documentation, both of which can contribute to early warning systems for impending deterioration. While modern deep sequence models often improve ranking performance for near-term adverse events, they frequently output probabilities that are poorly calibrated, especially when trained under extreme class imbalance and heterogeneous missingness. In operational settings, calibration is not a cosmetic property: it governs alarm thresholds, resource allocation, and downstream decision policies that interpret scores as risk. This paper develops a calibration-aware, multimodal forecasting framework for ICU deterioration prediction that explicitly couples representation learning with probability quality constraints. We formalize hourly prediction as conditional risk estimation over partially observed multivariate histories augmented by asynchronous text context, and we analyze how common optimization choices for imbalanced learning can distort probabilistic semantics even when discrimination remains strong. We then propose a structured methodology that separates discrimination learning from calibration recovery through principled post-hoc maps, while maintaining strict temporal availability to avoid leakage from documentation latency. The approach combines mask- and staleness-aware temporal encoders, modality-stable fusion operators, and calibration layers designed to preserve ordering while repairing probability scale. We further outline validation protocols that measure calibration under prevalence shift, quantify reliability under missingness stratification, and translate probability error into operational utility under alerting constraints. The result is a technically grounded blueprint for building multimodal ICU risk models whose outputs can be meaningfully used as probabilities rather than merely as scores.

1. Introduction

Clinical deterioration forecasting in the intensive care unit is typically posed as a sequence prediction problem in which a model observes a patient's evolving history and outputs a probability that an adverse event will occur within a fixed horizon [1]. Compared with classical early warning scores, deep learning approaches can exploit multivariate trends, non-linear interactions, and temporal dependencies that are otherwise inaccessible to additive scoring rules. However, the ICU is not a clean streaming environment in the computer-science sense. The observation process is event-driven rather than clock-driven, clinicians measure more when they are worried, and documentation pipelines introduce delays and batching effects. These properties create three intertwined technical challenges: the data are irregular and missingness is informative, the unstructured modality arrives asynchronously and can be temporally ambiguous, and the outcomes are low prevalence when framed as hourly "new event within horizon" labels.

A further challenge that becomes central when moving from retrospective benchmarks to real systems is calibration. Discrimination metrics such as AUROC and AUPRC quantify ranking quality, but they do not guarantee that a score of 0.40 corresponds to a 40% empirical

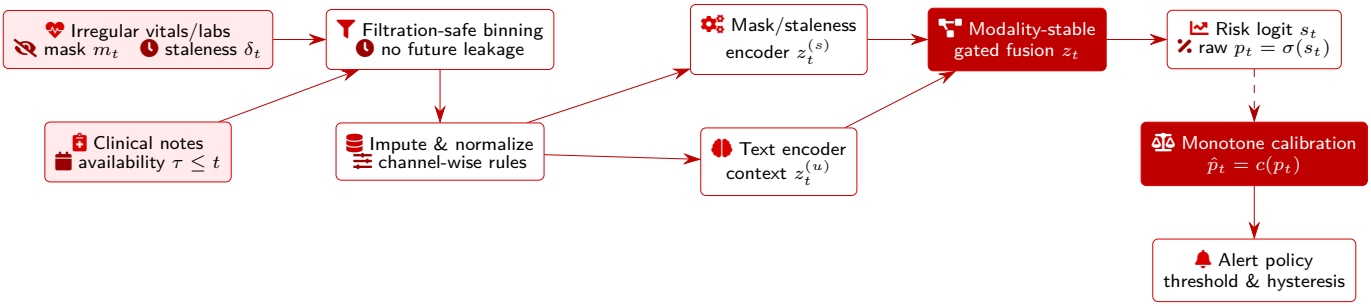

Figure 1: End-to-end calibration-aware multimodal ICU forecasting pipeline: filtration-safe preprocessing, mask/staleness-aware structured encoding and asynchronous text encoding, modality-stable gated fusion for risk scoring, followed by a monotone post-hoc calibration map and deployment-facing alert policy.

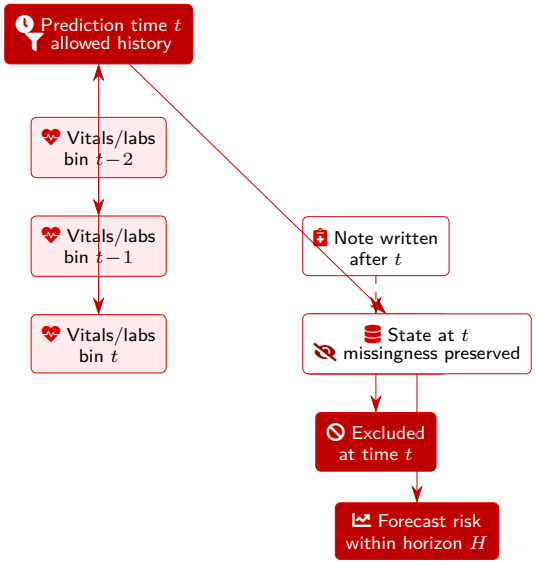

Figure 2: Filtration-correct conditioning under documentation latency: structured bins up to time t are usable, while notes that become available after t are excluded to prevent converting a forecaster into a detector of clinician recognition.

frequency, nor that a threshold chosen today will remain meaningful tomorrow as patient mix and practice patterns change. In an ICU alerting workflow, probability quality matters because many downstream operations implicitly assume calibrated probabilities. For example, a hospital may want a policy that triggers an alert only when the predicted probability exceeds a value corresponding to an expected number-needed-to-evaluate, or it may want to allocate limited resources to the top set of patients whose risk exceeds a clinically defined target. If the model is overconfident, alert fatigue increases; if it is underconfident, true deterioration is missed or delayed. Calibration error therefore translates directly into system-level harm even when ranking remains strong [2].

Deep neural networks trained for rare-event forecasting are particularly prone to miscalibration. Two drivers are common in practice. First, imbalance-aware losses, while beneficial for recall and precision, may violate strict proper scoring properties that would otherwise encourage calibrated probabilities. Second, architectural features such as attention, gating, and large embedding spaces can yield sharp decision boundaries and brittle confidence when the label is noisy or heterogeneous. In the ICU, label heterogeneity is not an edge case; composite deterioration definitions conflate several physiologic and intervention pathways, and documentation patterns can encode future interventions in a way that alters the semantics of the outcome. The result is a model that can appear highly competent under ranking metrics yet unreliable as a probabilistic estimator.

This paper presents a calibration-aware multimodal framework with an explicit goal: the output should be

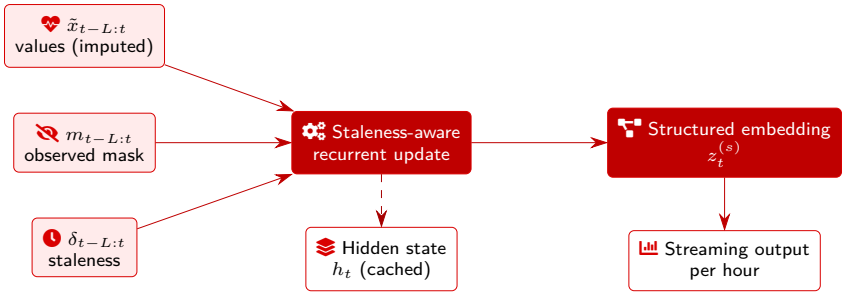

Figure 3: Structured-path encoder for irregular sampling: imputed values are paired with explicit masks and per-channel staleness, enabling a recurrent update that treats missingness as information while supporting cached, low-latency streaming inference.

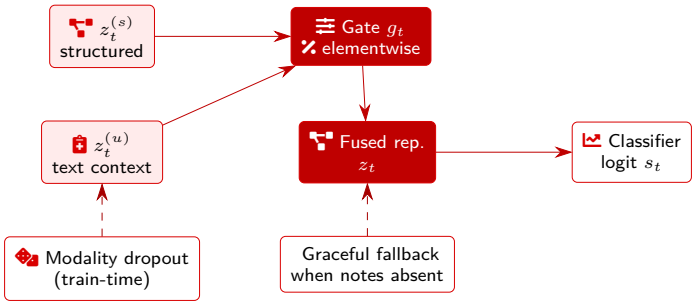

Figure 4: Modality-stable fusion: a learned gate blends structured and text embeddings in a way that remains well-defined when notes are missing, supported by modality dropout during training to prevent brittle reliance on documentation availability.

usable as a probability under realistic operational constraints, not merely as a score. The design is intentionally computer-science leaning, focusing on representational stability under missingness, modality fusion under asynchronous text, and algorithmic calibration recovery mechanisms that preserve discrimination. We treat the prediction function as a mapping from a temporally valid history to a probability, then analyze where and why common training pipelines distort that mapping. We propose a layered approach: learn a discriminative representation that is robust to irregular sampling and modality absence, then repair probability scale through post-hoc calibration maps that are monotone and therefore do not degrade ranking [3]. We also emphasize evaluation protocols that are sensitive to prevalence shift and missingness stratification, because calibration assessed only on a single held-out set can be misleading when deployment conditions differ from the training distribution.

The remainder of the paper is organized as a technical development. We first formalize the forecasting problem with explicit missingness and text availability constraints and define calibration metrics that align with decision-making. We then describe a multimodal representation-learning architecture that remains well-defined when notes are absent and when structured channels are sparsely sampled. Next, we analyze how imbalance-aware optimization affects probability semantics and derive post-hoc calibration operators that preserve ordering while reducing reliability error. We then discuss uncertainty quantification and robustness under distribution shift, framing monitoring as an online systems problem rather than a one-time evaluation. Finally, we propose an experimental protocol and error-analysis methodology that treat calibration as a first-class target, culminating in a deployment-oriented view of calibrated risk forecasting in critical care.

2. Formal Framework for Calibrated Risk Under Partial Observability

Let an ICU stay be indexed by i with discrete prediction times $t \in \{1, \dots, T_i\}$ at a fixed bin width Δ. Structured measurements are represented by a vector $x_{i,t} \in \mathbb{R}^d$ with a corresponding mask $m_{i,t} \in \{0,1\}^d$ indicating observed components and a staleness vector $\delta_{i,t} \in \mathbb{R}_+^d$ measuring time since last observation per channel. Unstructured documentation is a set of notes $\{(n_{i,j}, \tau_{i,j})\}_{j=1}^{J_i}$ where $\tau_{i,j}$ denotes availability time. The binary target $y_{i,t} \in \{0,1\}$ indicates whether a defined deterioration event occurs within a horizon of H bins after time t and is constrained to represent new events rather than continuing interven-

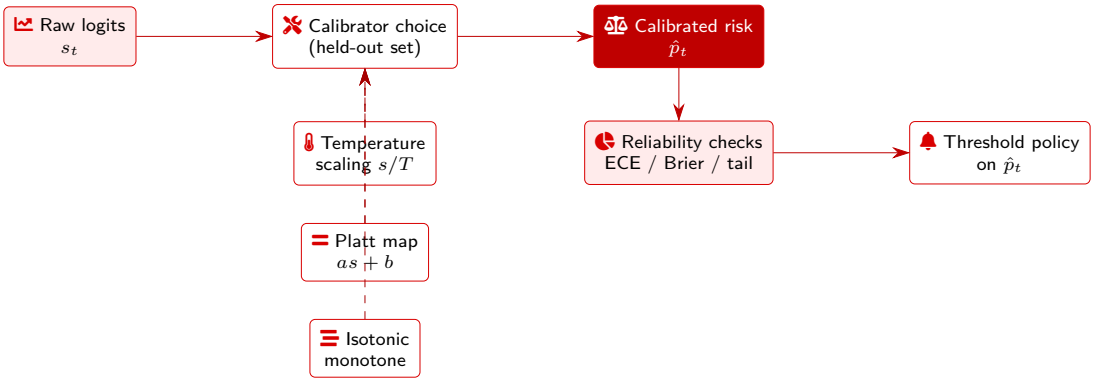

Figure 5: Post-hoc calibration layer: logits are mapped through a monotone calibrator (e.g., temperature scaling, affine Platt map, or isotonic regression) fit on held-out data to repair probability scale while preserving ordering; reliability is validated with global and tail-focused checks before threshold-based alerting.

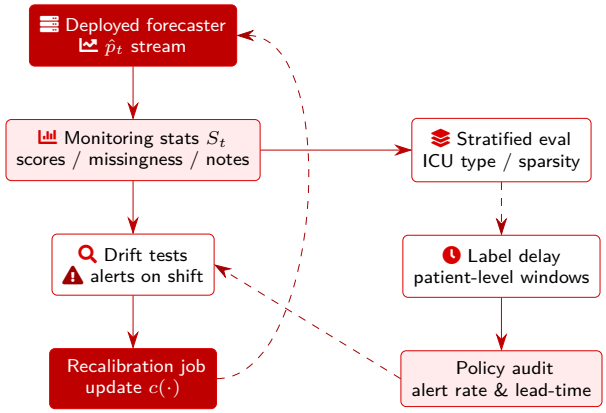

Figure 6: Deployment as a maintained system: calibrated risk streams are monitored via score/missingness/documentation summaries and stratified evaluations; detected drift triggers lightweight recalibration (updating only the calibration map) while policy auditing tracks operational impact under label delay and prevalence shift.

tions. The prediction is a conditional probability $r_\theta(i,t)$ parameterized by θ.

The first technical requirement is filtration correctness: the model must only use information available at or before time t. This can be stated as measurability with respect to the history sigma-algebra generated by $\{(x_{i,s}, m_{i,s}, \delta_{i,s})\}_{s \le t}$ and $\{n_{i,j} : \tau_{i,j} \le t\}$. In practice, filtration correctness is enforced through preprocessing rules that define bin endpoints, exclude events with timestamps after the query, and map note availability to conservative timestamps when charting delays are suspected [4]. This requirement is not only about avoiding blatant leakage. It also determines the semantics of the risk estimate, because including notes that were written after the clinician already recognized deterioration effectively turns the model into a detector of clinician recognition rather than a forecaster of physiologic change.

The second requirement is a computational representation of partial observability that does not equate missingness with normality. Because measurement is event-driven, $m_{i,t}$ carries predictive information about clinician concern and care intensity. Treating missing values as zeros can be acceptable only if the model is given $m_{i,t}$ and $\delta_{i,t}$ so it can learn to distinguish an imputed placeholder from an observed low value. A representation is therefore a triple $(\tilde{x}_{i,t}, m_{i,t}, \delta_{i,t})$, where $\tilde{x}_{i,t}$ is an imputed and normalized value vector. The imputation rule is part of the problem definition. Forward-fill produces dense tensors but extends old information into the present, which can create spurious stability. Zero-fill after normalization is computationally simple but demands explicit masks to avoid conflating missingness with typicality.

Calibration is defined relative to the probability output. A predictor is perfectly calibrated if, for any score value p, the empirical frequency of positives among sam-

Aspect	Notation	Description
ICU stay	$i, t \in \{1, \ldots, T_i\}$	Discrete prediction times on a fixed grid of width Δ within a single ICU admission.
Structured measurements	$x_{i,t} \in \mathbb{R}^d$	Multivariate physiologic and laboratory values aligned to prediction bins.
Missingness and staleness	$m_{i,t}, \delta_{i,t}$	Binary masks and time-since-last-observation vectors capturing informative observation patterns.
Unstructured notes	$\{(n_{i,j}, \tau_{i,j})\}$	Clinical documentation with content and availability timestamp, potentially delayed or batched.
Target label	$y_{i,t} \in \{0, 1\}$	Indicator of a new deterioration event occurring within horizon H after time t.
Risk estimator	$r_\theta(i, t)$	Conditional probability of deterioration given history, parameterized by model parameters θ.

Table 1: Core elements of the ICU deterioration forecasting setup with partial observability and asynchronous text.

ples with predicted probability p equals p. In finite data, calibration is evaluated through binning or kernel smoothing. Two widely used scalar summaries are the Brier score, which is a proper scoring rule capturing squared error of probabilities, and the expected calibration error (ECE), which approximates the average absolute deviation between predicted probabilities and empirical frequencies across bins [5]. These metrics have different failure modes. Brier score conflates calibration and discrimination because it penalizes both wrong ranking and wrong scaling. ECE isolates scaling but depends on bin choice and can hide tail miscalibration where operational thresholds often lie.

To make calibration operationally meaningful, one must connect it to decision policies. Consider a threshold policy that triggers an alert when $r_\theta(i, t) \geq \lambda$. If the model is overconfident, then at a given λ the realized positive rate among alerted samples will be lower than expected, increasing false alerts. If the model is underconfident, then to reach an acceptable alert precision one may need an unrealistically high threshold, missing events. A calibration-aware design therefore treats probability quality as a constraint on deployment. The objective is not necessarily to produce globally calibrated probabilities for all $p \in (0, 1)$, but to be reliable in the high-risk region where interventions are triggered.

These considerations motivate an approach that separates representation learning, which primarily affects ranking, from calibration recovery, which primarily affects probability scaling. The separation is computationally attractive because many calibration operators can be applied as lightweight post-processing layers that preserve ordering and thus keep AUROC and AUPRC unchanged [6]. It is also statistically attractive because it allows calibration to be adapted to local deployment conditions without retraining a high-capacity multimodal model, provided that distribution shift is not so severe that ranking fails.

3. Multimodal Temporal Representation Learning with Modality-Stable Fusion

A multimodal ICU forecaster can be constructed from a structured temporal encoder, a text encoder, and a fusion operator that yields a joint representation for risk estimation. The structured temporal encoder ingests a lookback window of length L bins ending at time t, represented as sequences $\{\tilde{x}_{t-\ell}, m_{t-\ell}, \delta_{t-\ell}\}_{\ell=0}^{L-1}$. The text encoder produces a time-indexed textual context u_t by embedding notes with availability time at or before t. Because notes are sparse, u_t is often piecewise constant, changing only when a new note arrives. The fusion operator must remain

Component	Representation choice	Rationale
Irregular sampling	Forward- or zero-filled $\tilde{x}_{i,t}$ with explicit $m_{i,t}$, $\delta_{i,t}$	Separates imputed placeholders from true values and encodes observation intensity.
Missingness signal	Mask channels $m_{i,t}$	Allows the model to leverage informative missingness without conflating it with normality.
Staleness	Per-feature $\delta_{i,t}$	Captures aging of measurements so old values can be discounted in temporal encoders.
Leakage control	Filtration based on $\tau_{i,j} \leq t$	Enforces temporal validity of notes and prevents post-recognition information from entering forecasts.
Imputation policy	Explicit rule for $\tilde{x}_{i,t}$	Makes the observation process part of the formal problem definition instead of an implementation detail.

Table 2: Design choices for encoding irregular sampling and informative missingness in structured ICU data.

well-defined when text is absent and must not allow note availability to become a shortcut proxy for acuity.

For structured sequences, a mask-aware recurrent encoder is a pragmatic baseline because it can update state sequentially and naturally supports streaming inference. A recurrent update can incorporate staleness by discounting hidden states when observations are old. A simplified formulation is to define a decay gate based on δ_t and combine it with a standard gated recurrent unit update, so that hidden state evolution reflects both observed values and their freshness. Temporal convolutional encoders can be competitive when the window length is fixed and parallelization is desired, but recurrent encoders remain attractive for low-latency deployment because they support incremental updates without recomputing the entire window.

Text encoding can be performed with a domain-pretrained transformer producing an embedding vector for each note. Given sparsity, one can define u_t as the embedding of the most recent available note, or as a weighted pool over all available notes with a decay kernel that discounts older documentation [7]. The key is temporal validity. If note timestamps are uncertain, using only radiology reports or only specific note categories can reduce the risk that text encodes post-event knowledge, but it can also reduce signal. In any case, the embedding should be treated as a noisy semantic observation that complements structured measurements rather than replacing them.

Fusion is where multimodal systems often fail due to instability under missing text. Naive concatenation of structured and text embeddings allows the classifier to infer note availability from systematic patterns in the concatenated vector, potentially learning that the presence of a note correlates with higher risk. This correlation may be an artifact of workflow rather than physiology. A modality-stable fusion operator instead treats text as optional evidence and provides a continuous interpolation between modalities. A gating mechanism is a simple and effective primitive: it learns a gate value based on both modalities and produces a convex combination in representation space. This yields a stable limit where the model can fall back to structured data when text is absent.

Let $z_t^{(s)} \in \mathbb{R}^h$ be the structured representation and $z_t^{(u)} \in \mathbb{R}^h$ the text representation projected into the same dimension. A gated fusion can be expressed as [8]

$$\begin{aligned} g_t &= \sigma\Big(W_g z_t^{(s)}\Big) \\ &\quad + \sigma\Big(U_g z_t^{(u)}\Big) \\ &\quad + \sigma(b_g) \\ z_t &= g_t \odot z_t^{(s)} \\ &\quad + (1 - g_t) \odot z_t^{(u)}, \end{aligned} \tag{1}$$

where σ is the logistic function and $\odot$ is elementwise product. The decomposition above emphasizes that the gate can be influenced by each modality separately, reducing reliance on a brittle concatenation pattern. In implementations, one typically uses an affine map of the concatenated representations, but separating terms clarifies that the gate

Source of text	Timing characteristics	Modeling decision
Progress notes	Often delayed, summarizing prior hours	Encode with transformers and align using conservative availability timestamps.
Nursing notes	Frequent but variable granularity	Treat as sparse context and optionally decay influence over time.
Radiology reports	Episodic, tied to imaging events	Restrict to reports whose timing reliably precedes potential deterioration.
Operational artifacts	Templates, copy-forward content	Rely on domain-pretrained encoders and regularization to avoid spurious lexical shortcuts.
Text context vector	Time-indexed u_t	Use last-available or time-decayed pooling over prior notes to form a streaming-compatible representation.

Table 3: Characteristics of clinical text and their implications for temporally valid integration into forecasting models.

Metric	Quantity measured	Limitations in this setting
AUROC	Global ranking between positive and negative samples	Insensitive to calibration and to prevalence, may look strong under severe imbalance.
AUPRC	Ranking quality focused on positives	Depends on base rate and does not constrain probability scale.
Brier score	Squared error of predicted probabilities	Conflates discrimination and calibration into a single scalar.
ECE	Average absolute deviation between predicted and empirical frequencies	Sensitive to binning and may hide tail miscalibration at high-risk thresholds.
Tail event rate	Empirical event rate in top risk quantiles	Focused on operational region but can be noisy when positives are rare.

Table 4: Discrimination and calibration metrics used to assess ICU risk models and their respective failure modes.

can be regularized to avoid degenerate behavior such as always selecting one modality.

Cross-modal attention can enrich this by allowing the structured sequence to be summarized in a text-conditioned way. However, cross-modal attention increases the number of parameters and the risk of overfitting under low prevalence. If used, it should be paired with conservative temporal alignment and with modality dropout so the model remains competent without text. Modality dropout can be implemented by randomly zeroing the text embedding during training with a fixed probability, encouraging the structured pathway to remain predictive. This is a form of regularization that targets the multimodal failure mode where the model becomes overly dependent on documentation availability.

The output layer maps z_t to a scalar logit s_t and then to a probability $p_t = \sigma(s_t)$. The architecture up to s_t is primarily responsible for representation and ranking [9]. Calibration is addressed by subsequent mapping of s_t or p_t through a calibration operator. This separation is central to the calibration-aware framework: it allows the high-capacity multimodal encoder to focus on extracting discriminative information while deferring probability scale correction to a smaller, more controllable component.

Objective	Modification to log loss	Effect on probability semantics	Typical motivation
Unweighted cross-entropy	None	Yields calibrated probabilities under correct model and i.i.d. data	Baseline, theoretically proper scoring rule.
Class-weighted loss	Per-class weights w_1, w_0	Shifts effective prior odds, producing biased probabilities under true prevalence	Improves recall under severe imbalance.
Focal loss	Factor $(1-p)^\gamma$	Emphasizes hard examples, often producing overconfident tails	Targets rare but important positives.
Label smoothing	Blurs targets away from $\{0, 1\}$	Reduces overfitting, but distorts exact conditional probabilities	Stabilizes training with noisy labels.
Sampling strategies	Over-/under-sampling	Changes training distribution relative to deployment	Balances batches to improve optimization.

Table 5: Imbalance-aware optimization choices and their consequences for the probabilistic interpretation of model outputs.

4. Imbalanced Optimization and Its Effects on Probability Semantics

Rare-event forecasting induces optimization behavior that differs sharply from balanced classification. When the positive rate is a few percent or lower, standard cross-entropy training can yield a model that minimizes loss by focusing on negative examples, producing conservative outputs that rank poorly for the minority class. Practitioners therefore introduce class weights, focal loss, label smoothing, or sampling strategies. These modifications often improve AUPRC and thresholded recall, but they can distort probability semantics. The result is a score that is useful for ranking but misinterpreted if treated as a probability.

The technical reason is that probability calibration is tied to proper scoring rules under the data-generating distribution. Unweighted log loss is strictly proper: it is minimized when the predicted probability equals the true conditional probability. When class weights are introduced, the objective corresponds to a different target distribution in which positives and negatives are artificially rebalanced. The minimizer is no longer the true conditional probability, but a transformed version [10]. If the deployment decision uses the raw output as a probability under the original prevalence, miscalibration is expected even in the infinite-data limit. Focal loss introduces an additional non-linear reweighting that emphasizes hard examples and further breaks strict properness, often yielding sharper probabilities that can be overconfident in certain regions.

These distortions can be analyzed by considering the relationship between the learned logit and the Bayes log-odds under the weighted objective. For a class-weighted log loss with positive weight w_1 and negative weight w_0, the optimal logit differs from the true log-odds by an additive constant that depends on the weight ratio. In effect, weighting changes the implicit prior odds seen by the model. This can be corrected post-hoc by shifting logits, but focal loss and label smoothing complicate the relationship because the weighting depends on the predicted probability itself.

To formalize the role of post-hoc correction, let $s(x)$ denote the learned logit from the representation model and let $p(x) = \sigma(s(x))$ be the raw probability. A calibration map c produces a calibrated probability $\hat{p}(x) = c(p(x))$ or, equivalently, applies a transformation to logits $\hat{s}(x) = h(s(x))$ followed by σ. If h is monotone increasing, ranking is preserved and AUROC is unchanged. Many post-hoc methods are therefore designed as monotone functions, either parametric such as temperature scaling or non-parametric such as isotonic regression. The key is that calibration can often be repaired without retraining the encoder, provided the encoder learned a reasonable ordering [11].

However, calibration must also consider heterogeneous label noise in ICU deterioration definitions. Com-

Method	Functional form	Monotonicity	Strengths and caveats
Temperature scaling	$p = \sigma(s/T)$	Monotone in logit s	Simple, preserves ranking, best for global over/underconfidence.
Platt scaling	$p = \sigma(as + b)$	Monotone if $a > 0$	Corrects both slope and intercept; can overfit on tiny calibration sets.
Isotonic regression	Piecewise-constant monotone map	Enforced by construction	Flexible, nonparametric; may be unstable in low-data tails.
Spline-based maps	Piecewise-linear monotone functions	Enforced via constrained knots	Balances flexibility and smoothness; needs careful regularization.
Ensemble-based calibration	Map applied to ensemble scores or logits	Depends on chosen map	Can combine with epistemic uncertainty estimates in deployment.

Table 6: Post-hoc calibration operators used to repair probability scale while preserving ranking.

posite outcomes combine mortality, initiation of vasopressors, and initiation of mechanical ventilation, each of which can occur for different reasons, including planned procedures. This introduces a mixture distribution where the same observed history may map to different event mechanisms. The true conditional probability is therefore a mixture probability that can be difficult to learn. Even under a strictly proper loss, a high-capacity model can overfit to spurious correlates that separate submechanisms but do not generalize. This appears as an increase in discrimination on the training set with degraded calibration and often a widening gap between training and validation loss. Regularization helps, but calibration-aware validation is needed because a model can improve AUROC while worsening reliability.

These considerations suggest a practical principle: optimize for discrimination under imbalance, but treat calibration as a separate engineering target to be repaired and monitored. The representation model is trained with an imbalance-aware loss chosen for operational ranking performance, and a calibration layer is learned on held-out data to map scores to probabilities under the target prevalence. This reduces the temptation to force the encoder to satisfy conflicting objectives and yields a modular pipeline in which calibration can be updated as conditions shift [12].

5. Post-Hoc Calibration, Reliability, and Uncertainty Quantification

Post-hoc calibration aims to transform raw model outputs into probabilities whose empirical frequencies match predicted values. The simplest and most widely used parametric method is temperature scaling, which divides logits by a positive scalar T before applying the logistic function. Temperature scaling preserves ranking because s/T is monotone in s for $T > 0$, and it can substantially improve calibration when miscalibration is primarily due to overconfidence or underconfidence rather than complex non-monotone distortions.

Let s_k be logits on a calibration set with labels y_k. Temperature scaling solves

$$\hat{p}_k(T) = \sigma\Big(\frac{s_k}{T}\Big)$$
$$T^\star = \arg\min_{T>0} \sum_k \Big(- y_k \log \hat{p}_k(T) - (1 - y_k) \log(1 - \hat{p}_k(T)) \Big), \qquad (2)$$

which is a one-dimensional convex optimization problem in many practical regimes. Because it uses log loss on a held-out set, it directly targets probability quality. Temperature scaling cannot fix all calibration issues. If the model is miscalibrated differently in different score regions, a single scalar may be insufficient. Nonetheless, it is computa-

Monitoring signal	Level	Purpose
Risk score distribution	Global, per-unit	Detects shifts in overall acuity or model output scale that threaten calibration.
Positive prevalence	Global, subgroup	Tracks base-rate drift that can invalidate previously fitted calibration maps.
Missingness rates	Global, feature-specific	Identifies changes in measurement policy that alter informative missingness patterns.
Note availability pattern	Global, modality-specific	Reveals documentation workflow shifts affecting the utility of text encoders.
Calibration metrics over time	Windowed, patient-level	Monitors degradation in reliability under evolving deployment conditions.
Alert volume and precision	System-level	Links calibration and thresholds to operational workload and clinical utility.

Table 7: Monitoring signals for treating calibration and robustness as ongoing systems-level responsibilities.

tionally attractive and stable, which matters when calibration must be refreshed periodically in deployment [13].

A distinct advantage of temperature scaling in ICU forecasting is that it cleanly separates discrimination learning from calibration recovery under imbalance-aware training. Even if focal loss produces overconfident probabilities, a fitted temperature can soften outputs without changing ordering. A concrete empirical illustration is that Sadanandan (2025) reports temperature scaling with a fitted scalar that reduced expected calibration error from 0.2129 to 0.0240 and improved Brier score while leaving AUROC and AUPRC unchanged [14], which is consistent with the theoretical expectation that monotone logit scaling preserves ranking but repairs probability scale when confidence is systematically distorted.

More expressive calibration maps include Platt scaling with an affine transform $as+b$, beta calibration, Dirichlet calibration for multiclass, and non-parametric isotonic regression. In binary settings, Platt scaling is often sufficient when class weighting shifts the intercept. Isotonic regression can fit arbitrary monotone maps but may overfit when the calibration set is small or when the positive rate is extremely low. A compromise is to use a piecewise-linear monotone spline with a small number of knots, which offers more flexibility than temperature scaling while retaining smoothness and reducing overfitting risk.

Calibration in the ICU is further complicated by prevalence shift and missingness shift. The positive rate can vary by ICU type, by season, by changes in admission patterns, and by protocol updates that alter the frequency of interventions. Missingness patterns can also shift if measurement devices change or if documentation templates are updated. A calibration map fit on one distribution may therefore become stale even if ranking remains stable. This motivates an online monitoring view in which calibration is periodically re-estimated using recent labeled data, potentially with a moving window and with safeguards against overreacting to short-term noise [15].

Uncertainty quantification extends beyond calibration. A calibrated probability reflects aleatoric uncertainty under the learned conditional distribution, but it does not capture epistemic uncertainty due to model uncertainty or distribution shift. In deployment, epistemic uncertainty can be as important as risk because it indicates when the model is operating outside its competence. A practical computational approach is to approximate epistemic uncertainty through ensembles or Monte Carlo dropout, producing a distribution over logits. One can then compute the mean probability and a dispersion measure, using dispersion as a confidence signal for triage. However, ensemble dispersion is not automatically calibrated, and it may correlate with missingness or note availability rather than with true out-of-distribution behavior.

A more principled approach is to decompose predictive uncertainty. If an ensemble produces logits $s^{(j)}$ for $j = 1, \ldots, M$, then probabilities are $p^{(j)} = \sigma(s^{(j)})$ and the mean $\bar{p}$ is an estimator of risk. The variability of $p^{(j)}$

Policy element	Description	Dependence on calibrated risk
Static alert threshold	Single probability cutoff for triggering alerts	Threshold meaning relies on calibrated mapping between risk and event rate.
Hysteresis band	Separate trigger and reset thresholds	Assumes stable probability scale to avoid oscillatory alert behavior.
Refractory period	Cooldown interval after an alert	Requires calibrated risk to tune trade-off between lead time and alert fatigue.
Persistent-high-risk rule	Alert only when risk stays high for multiple bins	Uses temporal patterns of calibrated probabilities rather than raw scores.
Resource-aware allocation	Prioritize top-k patients or those above target risk	Interprets probabilities in terms of expected number-needed-to-evaluate.

Table 8: Decision policy components that translate calibrated probabilities into actionable ICU alerts.

across ensemble members reflects epistemic components. One can define

$$\bar{p} = \frac{1}{M}\sum_{j=1}^{M} p^{(j)}$$
$$\mathrm{Var}_M(p) = \frac{1}{M}\sum_{j=1}^{M}\left(p^{(j)} - \bar{p}\right)^2, \tag{3}$$

and use $\mathrm{Var}_M(p)$ as a heuristic uncertainty score. If calibration is applied, it should be applied consistently across ensemble members or to the mean logit if the calibration map is defined in logit space. Monotone calibration preserves ranking but can compress probabilities, affecting dispersion; thus uncertainty monitoring must be interpreted in the calibrated space used for decisions [16].

Because ICU outcomes are rare, calibration and uncertainty estimation must be evaluated in the tail where alerts occur. Global ECE can look acceptable while the model remains overconfident among high predicted probabilities, because there are few samples in that region. Therefore, calibration evaluation should include tail-focused reliability checks, such as computing empirical positive rates within the top quantiles of predicted risk. These checks are not a replacement for ECE but a complement that aligns with operational decision thresholds.

Finally, calibration can be connected to decision policy optimization through expected utility. If the cost of a false alert is C_{FP} and the cost of missing a true deterioration is C_{FN}, then under calibrated probabilities the optimal threshold satisfies $\lambda = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}}+C_{\mathrm{FN}}}$. If probabilities are miscalibrated, this threshold is meaningless. This observation provides a pragmatic justification for calibration-aware modeling: it enables interpretable and stable decision thresholds grounded in explicit cost tradeoffs.

6. Distribution Shift, Missingness Stratification, and Monitoring as a Systems Problem

In ICU forecasting, distribution shift is not an exceptional event; it is a persistent property of the environment. Shifts arise from changes in patient demographics, evolving clinical protocols, new monitoring devices, and documentation practices that alter both structured measurements and text. A multimodal model can be more sensitive to shift than a structured-only model because text embeddings depend on language use and templates, while structured channels depend on measurement policy [17]. Calibration is especially vulnerable under shift, because small changes in prevalence or in score distribution can materially change probability reliability even when ranking remains stable.

A robust approach therefore treats monitoring as part of the algorithmic design. At minimum, one should track discrimination metrics and calibration metrics over time windows, stratified by clinically meaningful subpopulations. Because labels may be delayed, near-real-time monitoring can instead track proxy statistics such as score distributions, missingness patterns, and note availability rates, comparing them to training baselines. If the score distribution shifts substantially, probability calibration maps may need refresh. If missingness patterns shift, the model may be operating under a different observation

Protocol component	Choice	Motivation
Data splitting	Patient-level train / calibration / test	Prevents temporal leakage and reserves data for post-hoc calibration.
Evaluation granularity	Patient-level bootstrapping	Accounts for dependence among hourly samples within stays.
Metric set	AUROC, AUPRC, Brier, ECE, tail calibration	Separates ranking quality from probability reliability, especially in high-risk regions.
Prevalence shift analysis	Stratified subsets by unit or time	Probes robustness of calibration under realistic deployment shifts.
Modality ablations	Structured-only vs. text-only vs. fused	Diagnoses documentation shortcuts and multimodal instability.
Subgroup calibration	Stratified by demographics or ICU type	Detects systematic miscalibration affecting specific patient populations.

Table 9: Summary of an experimental protocol centered on discrimination, calibration, and robustness for ICU forecasting.

process, and both discrimination and calibration may degrade.

Missingness stratification is particularly important because in clinical data missingness can correlate with acuity in counterintuitive ways. High missingness can occur late in an ICU stay when fewer labs are ordered, but it can also occur during acute phases when a patient is moved or when documentation lags. A model that leverages masks may implicitly learn such correlations. This is acceptable for prediction, but it complicates portability: another institution may order labs differently, breaking the correlation between missingness and risk [18]. To diagnose this, one can evaluate performance separately across missingness strata defined by the fraction of observed structured features in the lookback window. If performance changes dramatically across strata, one should investigate whether the model is using missingness as a shortcut.

Text availability introduces similar issues. If notes are present for only a subset of stays, the model may treat note presence as an informative feature. A modality-stable fusion design reduces this risk, but it cannot remove it entirely. Monitoring should therefore include stratified evaluation by note availability and by note type, because radiology reports, nursing notes, and progress notes differ in both content and timing. If documentation policies change, the distribution of note types may shift, and the text pathway may become less useful or even harmful through spurious correlations.

From a systems perspective, monitoring can be cast as a drift detection problem on a set of summary statistics. Let S_t denote a vector of monitoring features at calendar time t, such as mean risk score, variance of risk score, fraction of samples above an alert threshold, missingness rate, and note availability rate. Drift detection then tests whether S_t differs from a baseline distribution. If drift is detected, one can trigger a recalibration procedure or, in more severe cases, a retraining procedure [19]. The key computational point is that recalibration is much cheaper than retraining and can often restore reliability when shift primarily affects prevalence or score scaling.

Recalibration must be performed carefully to avoid feedback loops. If clinicians respond to model alerts, the intervention changes the outcome distribution, potentially altering prevalence and causing the model's calibration to drift. This is a general issue in deployed predictive systems where predictions influence actions that influence outcomes. A practical approach is to recalibrate on labels defined in a way that is stable under intervention, or to use evaluation targets that reflect post-deployment semantics, acknowledging that the meaning of the predicted event may change. While fully causal correction is complex, a calibration-aware pipeline at least provides a mechanism to maintain probability reliability under evolving conditions.

An additional systems concern is that multimodal models may fail silently when one modality becomes unavailable due to pipeline outages. A gating-based fusion

architecture provides graceful degradation when text is missing, but only if the structured pathway remains strong. Therefore, system design should include health checks that detect modality outages and should log which modality contributed to each prediction in aggregate, enabling operators to notice when the model effectively runs in structured-only mode.

Finally, deployment requires clarity about the semantics of calibrated probabilities [20]. A calibrated probability is calibrated with respect to the deployed population, the label definition, and the prediction horizon. If any of these change, recalibration may be required. Calibration is therefore not a one-time property of a model; it is a maintained property of a model-system pair operating under a particular environment and label semantics. Treating calibration as a maintained systems artifact aligns better with real ICU workflows than treating it as a fixed metric in a retrospective benchmark.

7. Efficient Inference, Streaming Constraints, and Integration into Decision Policies

Even a well-calibrated model is not operationally useful unless it can run under the latency and reliability constraints of ICU systems. Structured data arrive continuously, labs arrive intermittently, and notes arrive in bursts. The prediction system must update risk scores without excessive recomputation and must handle late-arriving data. A streaming-friendly architecture therefore maintains incremental state for the structured encoder and updates text context only when new notes arrive. This favors recurrent encoders or causal convolutions with cached activations.

For a recurrent structured encoder, one can maintain a hidden state h_t updated as new bins arrive, using staleness and masks to interpret sparse observations [21]. If the prediction horizon is fixed, the model can output risk at each time step without reprocessing the full lookback window. If a fixed window is desired for reasons of stability, one can still maintain a rolling buffer and update state with a limited recomputation cost. The key algorithmic point is to avoid a transformer architecture that requires full recomputation over large windows unless key-value caching is carefully implemented.

Text integration in streaming settings is naturally event-driven. Let u_t denote the most recent note embedding available by time t. When a new note arrives at time τ, the system computes its embedding once and sets u_t to that embedding for all subsequent t until the next note arrives. The gating fusion then adapts to the updated context. This yields a low incremental cost relative to structured updates. However, it also introduces a potential for abrupt risk changes when a note arrives, which can be operationally confusing. Calibration and smoothing policies can mitigate this, for example by applying temporal smoothing to risk trajectories or by restricting alert triggers to persistently high risk across several bins [22].

Decision policies map a risk trajectory to actions. A naive policy triggers an alert when risk exceeds a threshold, but this often causes repeated alerts. A more stable policy uses hysteresis, requiring risk to exceed a high threshold to trigger and to fall below a lower threshold to reset. Another approach uses a refractory period that suppresses repeated alerts for a fixed time after an alert. These policy mechanisms are not incidental; they define how calibrated probabilities translate into real alert rates. Therefore, evaluation should assess the joint behavior of the model and policy, including measures such as alerts per patient-day, lead time before event, and fraction of events preceded by at least one alert.

Calibration interacts with policy design because thresholds are meaningful only when probabilities correspond to empirical frequencies. If the model is recalibrated, thresholds may need to be adjusted to maintain a desired alert rate or precision. This suggests treating threshold selection as part of a calibrated pipeline: recalibration updates probability scale, and then policy thresholds are either held fixed in probability space to preserve semantic meaning or adjusted to preserve operational characteristics. The choice depends on whether the institution values semantic thresholds or stable alert volume.

Computational reliability also requires handling missing data delays [23]. Labs can arrive late, and backfilled values can change the representation of prior bins. A strictly real-time system may choose to ignore backfilled values to preserve causality, but this can reduce accuracy. Alternatively, the system can allow retrospective updates to risk scores for audit while using forward-only data for alerts. The crucial design is to prevent backfilled data from influencing past alerts and to log the exact data used for each prediction to support reproducibility and incident analysis.

Finally, integration into clinical workflows demands interpretability signals that are compatible with calibrated probabilities. Rather than presenting raw feature attributions, which can be unstable, a calibrated system can present risk along with uncertainty indicators and trend summaries. Because this paper emphasizes computer-science aspects, the interpretability layer is framed as a user-interface and systems problem: present stable sum-

maries that do not imply causal explanations but support clinician sense-making, such as whether risk has been increasing over the last K hours and whether the model's uncertainty is high due to sparse observations.

8. Experimental Protocol and Error Analysis Focused on Calibration

A calibration-aware experimental protocol must address dependence among samples, prevalence effects, and tail reliability. Patient-level splitting is necessary to avoid leakage across time-indexed samples derived from the same patient. Beyond splitting, uncertainty estimation for metrics should be performed at the patient level, for example by bootstrapping patients rather than samples, because sample-level independence assumptions are violated in hourly forecasting [24].

Evaluation should report discrimination metrics such as AUROC and AUPRC, but it must also report calibration metrics such as Brier score and ECE. Because ECE depends on binning, it is advisable to compute multiple ECE variants, such as fixed-width bins and quantile bins, and to include reliability curves. Tail reliability should be explicitly assessed by computing empirical event rates among the top predicted risk quantiles, because these regions drive alerts. If calibration is repaired by a monotone map, discrimination metrics should remain unchanged; observing changes suggests either non-monotone calibration or numerical issues.

Calibration procedures must be fit on held-out data distinct from the test set to avoid optimistic bias. A three-way split is therefore preferable: a training set for representation learning, a calibration set for fitting post-hoc maps, and a test set for final reporting. When data are limited, cross-validation can be used, but it must respect patient-level grouping. In addition, calibration should be evaluated under simulated prevalence shift. One can create subsets with different base rates, for example by stratifying by ICU type or by time since admission, then evaluating whether the calibrated probabilities remain reliable. This provides an estimate of robustness to deployment drift [25].

Error analysis should be structured around failure modes that are specific to multimodal ICU forecasting. One failure mode is documentation-induced shortcut learning, where risk spikes when notes are present regardless of structured signals. This can be diagnosed by comparing risk distributions conditional on note availability and by evaluating text-only and structured-only ablations. Another failure mode is missingness-driven behavior, where high missingness is treated as high risk. This can be assessed through missingness stratification and by perturbation tests that artificially mask additional features to observe the model's response. Perturbation tests must be interpreted carefully because masking changes the input distribution, but they can still reveal brittle dependence on observation patterns.

Another important diagnostic is calibration by subgroup. Even if global calibration is good, a model may be miscalibrated for certain populations, such as specific age groups or ICU types, due to representation biases or prevalence differences. Subgroup calibration evaluation requires sufficient data in each subgroup and must therefore be treated as a statistically noisy estimate, but it can reveal systematic reliability issues that are masked in aggregate metrics. When subgroup miscalibration is detected, one can consider subgroup-specific calibration maps, but this introduces complexity and can be unstable if subgroup sizes are small [26]. A more scalable approach is to incorporate subgroup indicators into a calibration model that learns a modestly more expressive map without overfitting.

Finally, calibration-aware experiments should include a clear separation between score quality and probability quality. A model can be improved in discrimination by architectural changes, but those changes can worsen calibration. Therefore, experimental comparisons should report both sets of metrics and should present calibrated and uncalibrated results side by side. When calibration is applied, the method and calibration data must be reported explicitly, because calibration performance is otherwise not reproducible.

9. Conclusion

Calibration-aware multimodal forecasting for ICU deterioration requires treating probability quality as a maintained property rather than a static benchmark metric. The ICU environment combines irregular sampling, informative missingness, sparse asynchronous text, and severe class imbalance, all of which can distort probability semantics even when deep models achieve strong ranking. A technically coherent solution separates representation learning from calibration recovery: robust encoders and modality-stable fusion mechanisms are used to learn discriminative scores, and monotone post-hoc calibration maps are used to repair probability scale without sacrificing discrimination.

This paper presented a formal framework for temporally valid conditioning under partial observability, de-

scribed multimodal architectures that degrade gracefully when text is absent, analyzed how imbalance-aware objectives affect probabilistic interpretation, and developed post-hoc calibration and uncertainty strategies aligned with operational decision policies. It further framed monitoring and recalibration as systems problems driven by prevalence and missingness shifts, emphasizing that reliable deployment demands continuous evaluation, drift detection, and policy-aware threshold management. The resulting blueprint is intended to support the construction of ICU risk models whose outputs can be meaningfully interpreted as probabilities, enabling stable alerting behavior and more transparent integration into clinical workflows under evolving real-world conditions [27].

References

[1] A. M. Cole, K. A. Stephens, G. A. Keppel, H. Estiri, and L.-M. Baldwin, "Extracting electronic health record data in a practice-based research network: Lessons learned from collaborations with translational researchers," *EGEMS (Washington, DC)*, vol. 4, pp. 1206–1206, 3 2016.

[2] A. Amirahmadi, M. Ohlsson, and K. Etminani, "Deep learning prediction models based on ehr trajectories: A systematic review.," *Journal of biomedical informatics*, vol. 144, pp. 104430–104430, 6 2023.

[3] R. E. Johnson, Y. Ding, V. Venkateswaran, A. Bhattacharya, A. M. Chiu, T. Schwarz, M. K. Freund, L. Zhan, K. S. Burch, C. Caggiano, B. L. Hill, N. Rakocz, B. Balliu, J. H. Sul, N. Zaitlen, V. A. Arboleda, E. Halperin, S. Sankararaman, M. J. Butte, C. Lajonchere, D. H. Geschwind, and B. Pasaniuc, "Leveraging genomic diversity for discovery in an ehr-linked biobank: the ucla atlas community health initiative," 9 2021.

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

[5] A. Dubovitskaya, F. Baig, Z. Xu, R. Shukla, P. S. Zambani, A. Swaminathan, M. M. Jahangir, K. Chowdhry, R. Lachhani, N. Idnani, M. Schumacher, K. Aberer, S. D. Stoller, S. Ryu, and F. Wang, "Action-ehr: Patient-centric blockchain-based electronic health record data management for cancer care (preprint)," 2 2019.

[6] Z. Davis and H. Davis, "Usability assessment in the adaptation phase of the electronic health record," *Journal of Information Technology Case and Application Research*, vol. 24, pp. 119–134, 2 2022.

[7] A. A. Mallah, P. Guelpa, S. Marsh, and T. van Rooij, "Integrating genomic-based clinical decision support into electronic health records.," *Personalized medicine*, vol. 7, pp. 163–170, 3 2010.

[8] J. Hou, R. Zhao, J. Gronsbell, B. K. Beaulieu-Jones, G. Webber, T. Jemielita, S. Wan, C. Hong, Y. Lin, T. Cai, J. Wen, V. A. Panickan, C.-L. Bonzel, K.-L. Liaw, K. P. Liao, and T. Cai, "Harnessing electronic health records for real-world evidence," 11 2022.

[9] J. Berg, C. O. Aasa, B. A. Thorell, and S. Aits, "Openchart-se: A corpus of artificial swedish electronic health records for imagined emergency care patients written by physicians in a crowd-sourcing project," 12 2022.

[10] M. VanBeek, R. A. Swerlick, B. Mathes, M. J. Reeder, T. Kaye, A. Aninos, M. Fitzgerald, C. D. Etkin, and J. P. Jacobs, "The completeness and accuracy of dataderm™: The database of the american academy of dermatology (aad)," *Journal of the American Academy of Dermatology*, vol. 86, pp. 394–398, 6 2021.

[11] O. Asan, E. K. Chiou, and E. Montague, "Quantitative ethnographic study of physician workflow and interactions with electronic health record systems.," *International journal of industrial ergonomics*, vol. 49, pp. 124–130, 9 2015.

[12] K. E. Furlong, "Ehr learning – it's about nursing, leadership and long-term commitments," *Nursing leadership (Toronto, Ont.)*, vol. 28, pp. 38–47, 1 2016.

[13] R. D. Wijaya and L. O. A. Rahman, "Implementasi electronic health records (ehrs) pada pelayanan kesehatan di komunitas: Literature review," *Jurnal Kesehatan*, vol. 8, pp. 28–38, 2 2020.

[14] 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.

[15] J. Li, V. Pandey, and R. Hendawi, "A blockchain-based personal health knowledge graph for secure integrated health data management," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–7, IEEE, 7 2023.

[16] A. G. Hirsch and A. S. McAlearney, "Measuring diabetes care performance using electronic health record data: The impact of diabetes definitions on performance measure outcomes.," *American journal of medical quality : the official journal of the American College of Medical Quality*, vol. 29, pp. 292–299, 9 2013.

[17] M. F. Funk, "A survey of chiropractic intern experiences learning and using an electronic health record system.," *The Journal of chiropractic education*, vol. 32, pp. 145–151, 6 2018.

[18] T. Tajirian, V. Stergiopoulos, G. Strudwick, L. Sequeira, M. Sanches, J. Kemp, K. Ramamoorthi, T. Zhang, and D. Jankowicz, "The influence of electronic health record use on physician burnout: Cross-sectional survey (preprint)," 4 2020.

[19] S. Palojoki, K. Saranto, E. Reponen, N. Skants, A. Vakkuri, and R. Vuokko, "Classification of electronic health record–related patient safety incidents: Development and validation study (preprint)," 5 2021.

[20] Y. Li, J. Chok, G. Cui, D. Roosan, and K. Shultz, "Electronic health record adoption among adult day services: Findings from the national study of long-term care providers.," *Journal of the American Geriatrics Society*, vol. 71, pp. 3941–3943, 8 2023.

[21] B. Cyganek, M. Graña, B. Krawczyk, A. Kasprzak, P. Porwik, K. Walkowiak, and M. Woźniak, "A survey of big data issues in electronic health record analysis," *Applied Artificial Intelligence*, vol. 30, pp. 497–520, 7 2016.

[22] E. V. Bernstam, J. L. Warner, J. C. Krauss, E. Ambinder, W. S. Rubinstein, G. Komatsoulis, R. S. Miller, and J. L. Chen, "Quantitating and assessing interoperability between electronic health records.," *Journal of the American Medical Informatics Association : JAMIA*, vol. 29, pp. 753–760, 1 2022.

[23] "A standard model for evaluating return on investment from electronic health record implementation," 1 2014.

[24] A. Keil, R. Brück, and K. Hahn, "Transmission of vital data into the german electronic health record," *Current Directions in Biomedical Engineering*, vol. 9, pp. 583–586, 9 2023.

[25] S. Ballal, S. Awasthi, and M. Panda, "Integrating electronic health records to address clinician burnout and improve workplace well-being in healthcare settings," *Seminars in Medical Writing and Education*, vol. 2, pp. 140–140, 12 2023.

[26] S. T. Rosenbloom, W. W. Stead, J. C. Denny, D. A. Giuse, N. M. Lorenzi, S. H. Brown, and K. B. Johnson, "Generating clinical notes for electronic health record systems.," *Applied clinical informatics*, vol. 1, pp. 232–243, 1 2010.

[27] F. Lau, M. Antonio, K. Davison, R. Queen, and K. Bryski, "Correction: An environmental scan of sex and gender in electronic health records: Analysis of public information sources (preprint)," 5 2021.