WHERE IS THE PATIENT HEADING? MULTI-PATH FORECAST-AWARE MORTALITY PREDICTION IN THE ICU

Anonymous authors
Paper under double-blind review

ABSTRACT

Mortality prediction in the intensive care unit (ICU) requires estimating risk from the patient history available at a given decision time. Although existing predictors can model longitudinal observations and learn associations with future outcomes, they need not explicitly represent anticipated physiological changes or incorporate the outcomes of similar patients. We investigate whether explicit physiological forecasts and retrieved patient outcomes improve prediction under matched observation windows. We propose TRACE (Trajectory-Retrieval Augmented Clinical Estimation), which forecasts quantiles of near-future vital signs and uses the predicted median and interval width as features for mortality classification. A complementary retrieval pathway estimates risk from the observed outcomes of similar training patients, and the two estimates are combined through entropy-conditioned weighting. Across seven forecasting backbones on MIMIC-IV and eICU-CRD, TRACE improves discrimination relative to matched observed-history and point-forecast controls. Component comparisons associate both quantile-derived features and outcome retrieval with further gains, and the observed benefit over matched controls extends across the evaluated decision times. On a matched MIMIC-IV task, TRACE also achieves the highest AUROC and AUPRC among the mortality predictors evaluated. These findings support explicit physiological forecasting and outcome retrieval as complementary components for ICU mortality prediction in the studied settings.

1 INTRODUCTION

Mortality-risk estimation supports risk stratification in the intensive care unit (ICU), where patients' conditions can change rapidly (Churpek et al., 2016; Jones et al., 2009). The usefulness of a risk estimate depends on both its predictive quality and the decision time at which it is available. At a given decision time, prediction must rely on measurements available up to that time, rather than values recorded later (Subbe & Welch, 2013; Franklin & Mathew, 1994; Smith et al., 2014).

Traditional scores such as APACHE II, SOFA, and SAPS II summarize clinical severity (Knaus et al., 1985; Vincent et al., 1996; Gall et al., 1993). Learned predictors can also use longitudinal electronic health record (EHR) data, including recurrent and attention-based models that account for temporal patterns in the observed history (Harutyunyan et al., 2019; Che et al., 2018; Shukla & Marlin, 2021; Rajkomar et al., 2018; Tipirneni & Reddy, 2022; Wang et al., 2024b). Such models may implicitly account for different future courses through their associations with observed outcomes. However, predicting mortality directly from a history does not require the model to explicitly estimate near-future physiology or to form a query-specific risk estimate from the observed outcomes of similar patients. This motivates a broader empirical question: *At the same decision time, can a forecast-aware, retrieval-augmented approach improve ICU mortality prediction beyond direct classification of the observed history?*

Forecasting offers one way to represent anticipated physiology. A point forecast estimates a single value at each future time, while quantile forecasts describe a predicted median and the spread of future values (Duggal et al., 2024; Chang et al., 2024; Koenker & Bassett, 1978; Gneiting & Raftery, 2007). Predicted clinical events have also been used in future-aware EHR models (Renc et al., 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

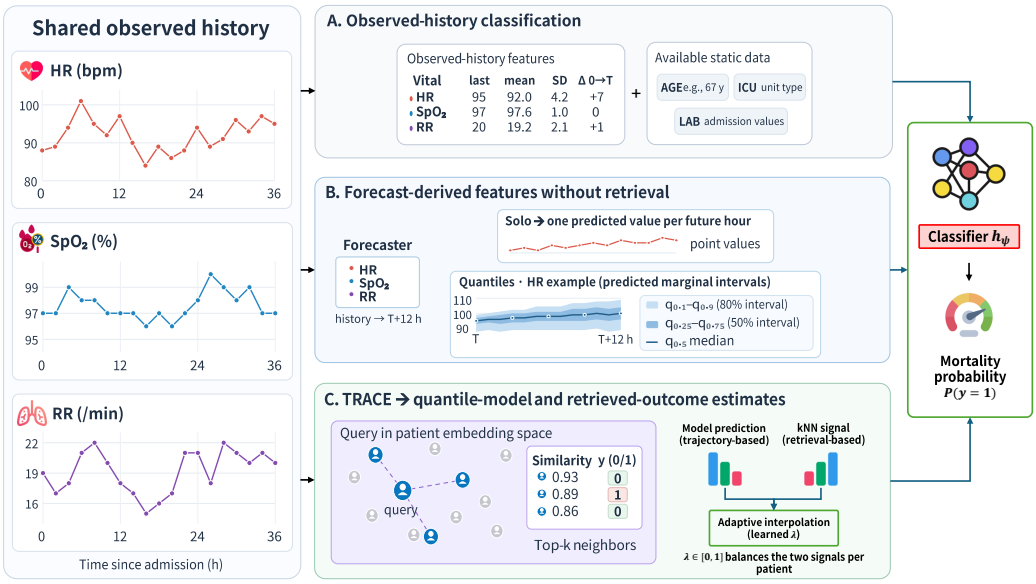


Figure 1: Conceptual comparisons at the same ICU decision time T. **(A)** An observed-history classifier estimates mortality risk from available patient data; it can use longitudinal information. **(B)** Solo uses a point forecast, whereas Quantile-only supplies predicted median and interval-width features, without retrieval. **(C)** TRACE combines the Quantile-only model estimate with a separate estimate based on observed outcomes of retrieved training patients.

Waxler et al., 2025; Pang et al., 2025). Patient-similarity methods provide a separate motivation for estimating risk from related patients' observed outcomes (Zhang et al., 2014; Seymour et al., 2019). Our study examines a specific combination: fixed-horizon forecasts of continuous ICU vital signs as features for mortality classification, together with an outcome estimate from patients retrieved using the observed history, as shown in Figure 1.

We propose TRACE (**T**rajectory-**R**etrieval **A**ugmented **C**linical **E**stimation), a forecaster predicts quantiles of near-future heart rate, respiratory rate, and oxygen saturation from observations available at the decision time. A mortality classifier uses the predicted median and interval width alongside observed-history features. A patient embedding trained with a severity proxy and auxiliary mortality supervision retrieves similar patients from the training cohort; their recorded outcomes form a second risk estimate. TRACE combines the two estimates using entropy-conditioned interpolation. We evaluate matched internal controls at the same observation cutoff and mortality endpoint on MIMIC-IV and eICU-CRD, examine the component configurations across seven forecasting backbones, and study additional decision times. We also compare with independently implemented mortality classifiers on a matched MIMIC-IV task.

Our contributions are as follows:

- We introduce TRACE, a forecast-aware framework that makes predicted near-future physiology available to an ICU mortality classifier. Vital-sign medians and quantile-interval widths provide the classifier with explicit information about anticipated values and forecast spread alongside the observed patient history.
- We augment this forecast-aware estimate with observed mortality outcomes from similar patient outcomes. An entropy-conditioned interpolation rule combines the model and retrieval estimates, allowing retrieved outcomes to contribute a separate predictive signal.
- Across seven forecasting backbones and two ICU cohorts, TRACE achieves higher AUROC and AUPRC than matched observed-history and point-forecast controls at the primary decision time. Comparisons with quantile-only models and at additional decision times further support the value of combining forecast-derived features with outcome retrieval in the evaluated settings.

2 Related Work

Risk prediction from observed clinical histories. ICU risk models can use temporally ordered measurements rather than only static snapshots. Recurrent and attention-based methods model clinical time series, including irregular observations and missing values (Harutyunyan et al., 2019; Che et al., 2018; Shukla & Marlin, 2021; Tipirneni & Reddy, 2022; Wang et al., 2024b). Tabular models can also summarize longitudinal records over defined look-back windows (Oufattole et al., 2024), while long-context EHR studies examine how the amount of available history affects clinical prediction (Wornow et al., 2024). These approaches may learn associations between the observed history and heterogeneous subsequent patient courses without explicitly forecasting future physiological measurements. We investigate whether such explicit forecasts add predictive value when risk is estimated at the same decision time.

Future-aware EHR modeling. Prior work represents future clinical events through simulated timelines (ETHOS-ARES, Curiosity) (Renc et al., 2025; Waxler et al., 2025), generated records (CEHR-XGPT) (Pang et al., 2025), event-stream and mixed-measurement models (McDermott et al., 2023; Loza et al., 2025), and time-to-event or query-specific prediction (Steinberg et al., 2024; Jing et al., 2026; Chandak et al., 2026). TRACE uses fixed-horizon quantiles of continuous hourly ICU physiology as mortality-prediction features and evaluates their added value at matched observation cutoffs.

Physiological Forecasting and Outcome Retrieval. Clinical forecasting studies have modeled subsequent physiological measurements and their uncertainty (Duggal et al., 2024; Chang et al., 2024). General-purpose architectures provide forecasting backbones including PatchTST, iTransformer, Crossformer, TimeMixer, and DLinear (Nie et al., 2023; Liu et al., 2024; Zhang & Yan, 2023; Wang et al., 2024a; Zeng et al., 2023); Chronos and Moirai provide pretrained backbones (Ansari et al., 2024; Woo et al., 2024). Patient-similarity methods use related clinical records to inform personalized decisions (Zhang et al., 2014), while the kNN-LM illustrates probability-level interpolation with an estimate derived from retrieved examples (Khandelwal et al., 2020). TRACE brings these two ideas together by combining a forecast-aware mortality estimate with an outcome estimate from similar training patients.

3 Problem Formulation

3.1 Patient Observation and Task Definition

Let $\mathcal{D} = \{(X_i, Y_i, y_i)\}_{i=1}^{N}$ denote the evaluated cohort of ICU stays lasting at least 48 hours. At a decision time T, each stay provides an observed history X_i, subsequent physiological measurements Y_i used as a forecasting target, and an in-hospital mortality label y_i. The observed data $X_i = (V_i, S_i)$ contain an hourly clinical time series $V_i \in \mathbb{R}^{T \times D}$ with $D = 61$ features and an encoded covariate vector $S_i \in \mathbb{R}^{33}$ constructed only from information available by hour T. The subsequent H hours of $K = 3$ target vital signs (heart rate, respiratory rate, and SpO_2) form $Y_i \in \mathbb{R}^{H \times K}$, and $y_i \in \{0, 1\}$ is the in-hospital mortality label. For retrieval, we augment V_i with observation counts and elapsed time since each feature was measured to form V_i^{rich} (Section 4.2).

Direct classifiers map available clinical histories to mortality risk and can capture temporal patterns without explicitly forecasting subsequent physiology (Harutyunyan et al., 2019; Che et al., 2018; Wang et al., 2024b). Our **forecast-aware** formulation provides predicted near-future physiology to the risk model and separately estimates risk from retrieved training-patient outcomes. Let $\mathcal{I}_{\text{train}}$ be the indices of training stays. The three model stages are:

$$\text{Stage 1 (Forecasting):} \quad g_\theta : V_i \mapsto \hat{Y}_i \in \mathbb{R}^{Q \times H \times K}, \tag{1}$$

$$\text{Stage 2 (Retrieval):} \quad r_\phi : (X_i, \{z_j\}_{j \in \mathcal{I}_{\text{train}}}) \mapsto \mathcal{R}(i) \subseteq \mathcal{I}_{\text{train}}, \tag{2}$$

$$\text{Stage 3 (Model risk):} \quad h_\psi : \mathbf{f}_i \mapsto p_{\text{model}}(i) \in [0, 1], \tag{3}$$

where $Q = 5$ is the number of forecast quantile levels and $\mathbf{f}_i$ (Eq. 6) contains observed-history summaries, available covariates, and selected forecast-derived features. The mortality labels of patients in $\mathcal{R}(i)$ form a separate estimate $p_{\text{kNN}}(i)$, which is interpolated with $p_{\text{model}}(i)$ in Eq. 8.

The forecaster and retrieval embedding are trained separately and fixed when training the mortality classifier.

3.2 Quantile Forecasting

The forecaster g_θ predicts future vital-sign values at Q quantile levels for each forecast hour and target variable:

$$g_\theta(V_i) = \{\hat{Y}_i^{(q)} \in \mathbb{R}^{H\times K} \mid q \in \mathcal{Q}\}, \quad \mathcal{Q} = \{0.1,\, 0.25,\, 0.5,\, 0.75,\, 0.9\}. \tag{4}$$

Each $\hat{Y}_i^{(q)}$ collects predictions at quantile level q over the forecast horizon and target variables. They specify per-hour marginal quantiles, not independently sampled patient futures. The model is trained with a composite loss $\mathcal{L}_{\text{forecast}} = \mathcal{L}_Q + \lambda_T \mathcal{L}_T + \lambda_D \mathcal{L}_D$, where $\mathcal{L}_Q$ is the masked pinball loss for observed future targets (Koenker & Bassett, 1978), $\mathcal{L}_T$ penalizes errors in temporal differences of the predicted median, and $\mathcal{L}_D$ is an auxiliary loss for future vital-sign abnormality. Their definitions and weights are given in Appendix D.3.

3.3 Retrieval-Augmented Risk Classification

We learn an L2-normalized, 256-dimensional patient embedding $\phi : (V_i^{\text{rich}}, S_i) \mapsto z_i$ through self-supervised pretraining and SOFA-proxy-based fine-tuning with auxiliary mortality supervision (Section 4.2). Given a query stay i, we retrieve the k training stays with the highest cosine similarity:

$$\mathcal{R}(i) = \underset{j\in\mathcal{I}_{\text{train}}\setminus\{i\}}{\arg\,\text{top-k}}\ \cos(z_i, z_j). \tag{5}$$

For a validation or test query, $i \notin \mathcal{I}_{\text{train}}$; the exclusion applies to a training query retrieving from its own pool. The **model pathway** uses the feature vector:

$$\mathbf{f}_i = \big[\mathbf{h}_i;\ \text{vec}(\hat{Y}_i^{(0.5)});\ \text{vec}(\hat{Y}_i^{(0.9)} - \hat{Y}_i^{(0.1)});\ S_i\big] \tag{6}$$

where $\mathbf{h}_i \in \mathbb{R}^{12}$ contains the last value, mean, standard deviation, and recent trend for each target vital sign. The remaining terms are the predicted median, the 0.9–0.1 quantile width at each future hour, and S_i. The **retrieval pathway** aggregates observed mortality labels:

$$p_{\text{kNN}}(i) = \sum_{j\in\mathcal{R}(i)} w_{ij} y_j, \qquad w_{ij} = \frac{\exp(-d_{ij}/\tau)}{\sum_{j'\in\mathcal{R}(i)} \exp(-d_{ij'}/\tau)}, \tag{7}$$

where $d_{ij} = 1 - \cos(z_i, z_j)$ and $\tau > 0$ is a temperature parameter. The final mortality estimate interpolates the two pathway probabilities:

$$\hat{p}_i = (1-\lambda_i)\, p_{\text{model}}(\mathbf{f}_i) + \lambda_i\, p_{\text{kNN}}(i), \quad \lambda_i = \lambda_{\max} \cdot \big(1 - \mathcal{H}_{\text{b}}(p_{\text{kNN}}(i))\big), \tag{8}$$

where $\lambda_{\max} \in [0,1]$ bounds the retrieval weight and $\mathcal{H}_{\text{b}}(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is normalized binary entropy. Hence $\mathcal{H}_{\text{b}}(p) \in [0,1]$ and $\hat{p}_i$ is the final estimated in-hospital mortality probability.

4 Method

4.1 Overview

TRACE implements the three stages in Section 3 (Eqs. 1–3): quantile forecasting with g_θ, patient retrieval with an embedding ϕ, and mortality prediction with h_ψ. The prediction from h_ψ is subsequently combined with a separate estimate derived from retrieved patients' mortality labels (Eq. 8). Figure 2 summarizes these components. The embedding/classification and forecasting pipelines use the same patient cohort, normalization statistics, and dataset splits; see Appendix A.

4.2 Patient Embedding for Retrieval

Two-stage training. The embedding ϕ maps (V_i^{rich}, S_i) to an L2-normalized, 256-dimensional vector z_i. First, we pretrain the temporal encoder on V_i^{rich} with self-supervised contrastive learning (Yue et al., 2022; Chen et al., 2020). We then fine-tune the embedding using a SOFA-derived

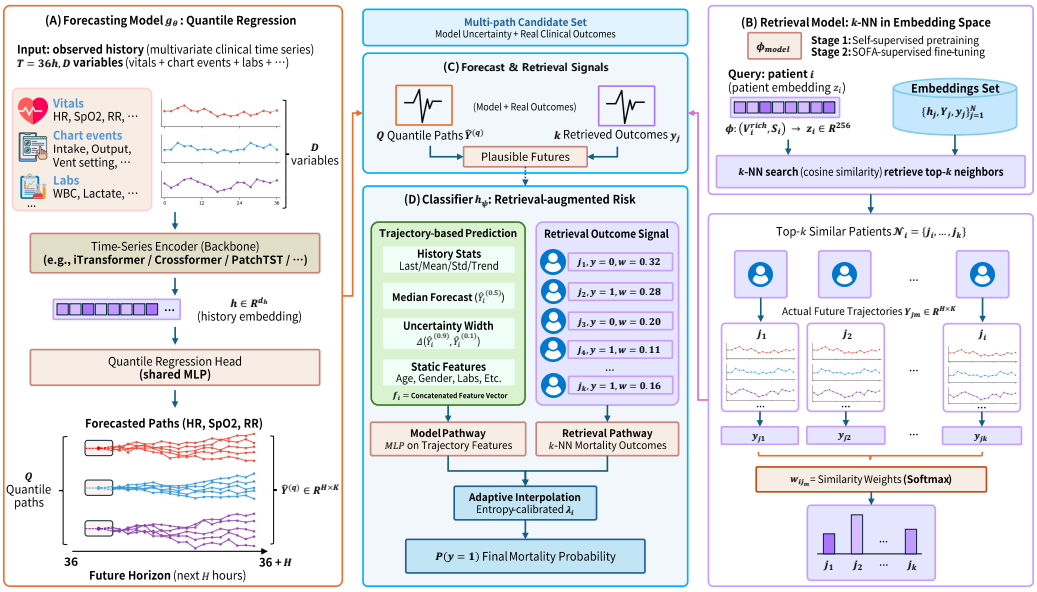


Figure 2: Overview of TRACE. **(A)** A quantile forecaster g_θ predicts future vital signs at Q=5 quantile levels from the observed history. **(B)** An embedding model ϕ, trained with self-supervised pretraining and SOFA-proxy-based fine-tuning, retrieves similar patients from the training cohort. **(C)** Predicted quantiles provide the median and interval-width features for the model pathway; the retrieved patients' observed mortality labels provide a separate neighbor-based risk estimate. **(D)** The final mortality estimate interpolates the model and neighbor-based estimates using a weight determined by the binary entropy of the neighbor-based estimate.

severity proxy as its primary similarity signal, together with auxiliary mortality-supervised objectives. Thus, the learned embedding can reflect outcome information even though the retrieved outcomes are combined with the model estimate only in the final prediction step (Appendix C.2).

Fine-tuning objective. The objective contains four terms. A soft supervised contrastive loss assigns pairwise attraction weights according to the standardized SOFA-derived proxy $\hat{s}_i$:

$$w_{ij} = \exp\left(-\frac{(\hat{s}_i - \hat{s}_j)^2}{2\sigma^2}\right), \tag{9}$$

where σ sets the scale of proxy-score similarity. An intra-SOFA outcome loss uses mortality labels to distinguish patients within proxy-score bins (Robinson et al., 2021). A uniformity term discourages embedding collapse (Wang & Isola, 2020), and an auxiliary binary cross-entropy loss supervises a mortality head during training; this head is discarded for retrieval at inference. We initialize the static encoder with a warm-up phase before fine-tuning the full embedding model. Loss weights, the proxy definition, and training details are given in Appendix C.3.

4.3 Quantile Forecasting

We train the forecaster on sliding windows from each ICU stay and use a single admission-aligned history window for each patient at evaluation (Iwana & Uchida, 2021). Each forecasting backbone outputs $H \times Q$ values per target variable, indexed by forecast hour and quantile level. These values define marginal predictive quantiles at each hour and variable (Eq. 4). In addition to the quantile loss $\mathcal{L}_Q$ (Koenker & Bassett, 1978), we use a temporal trend loss $\mathcal{L}_T$ and an auxiliary loss $\mathcal{L}_D$ for clinically abnormal vital-sign ranges (Singer et al., 2016; Chourpiliadis & Bhardwaj, 2022; Jubran, 1999). Gaussian input noise and random temporal shifts augment training windows (Iwana & Uchida, 2021). Full implementation details are in Appendix D. At prediction time, the classifier receives the predicted median $\hat{Y}_i^{(0.5)}$ and the predicted quantile-interval width $\hat{Y}_i^{(0.9)} - \hat{Y}_i^{(0.1)}$, together with observed historical and static features (Eq. 6). Interval width describes the spread

between the model's predicted quantiles. For retrospective analysis, we also compute a Coverage Score: the fraction of actual future measurements Y_i inside the corresponding predicted $[0.1, 0.9]$ quantile interval.

4.4 Retrieval-Augmented Risk Classification

We freeze ϕ and g_θ while training the mortality classifier and cache the resulting embeddings and forecast-derived features. The model pathway estimates p_{model} with an MLP h_ψ operating on $\mathbf{f}_i$ (Eq. 6). Its input excludes the query embedding z_i and retrieved patient information, keeping the model and retrieval inputs separate (Appendix E.2). In parallel, the retrieval pathway selects the k most similar training patients by cosine similarity and combines their observed mortality labels into a similarity-weighted estimate p_{kNN} (Eq. 7). We exclude a training query's own stay from its retrieval pool to prevent self-retrieval. The final prediction combines p_{model} and p_{kNN} at the probability level (Eq. 8). Its entropy-conditioned weight gives greater influence to retrieval when the neighbor-based estimate is near 0 or 1 and reduces that influence when it is near 0.5. Appendix E.4 details the retrieval pool, k, temperature τ, and maximum retrieval weight $\lambda_{\max}$.

5 Experiments

5.1 Experimental Setup

Evaluation protocol. We predict in-hospital mortality after observing the first $T{=}36$ hours of an ICU stay, with $T{=}36$, $H{=}12$ as our primary configuration. TRACE predicts the next H hours of vital signs from the available history; neither TRACE nor any comparator observes those future measurements at the decision time. For each dataset and decision time, all variants use the same eligible cohort, patient-level data split, endpoint, and first T hours of available observations. We additionally study $(T, H) \in \{(24, 24), (30, 18), (42, 6)\}$ hours, so that $T{+}H{=}48$ in every configuration. The cohort comprises ICU stays lasting at least 48 hours; thus the results at $T{<}48$ characterize this selected cohort, rather than all patients present in an ICU at hour T.

Datasets. We evaluate TRACE on two datasets. For primary evaluation, we use **MIMIC-IV**[1] (Johnson et al., 2023), a large single-center ICU database from Beth Israel Deaconess Medical Center. Following the cohort construction procedure described in Appendix A.1, we split patient stays at the subject level to prevent the same patient's stays from occurring in both training and evaluation splits. For a second cohort evaluation, we use **eICU-CRD**[2] (Pollard et al., 2019), a multi-center ICU database covering over 200 hospitals across the United States. Models are trained and evaluated using a separate patient-level split for each dataset.

Baselines. We compare TRACE with three internal controls at matched decision times. **Direct Classification** is our observed-history MLP control, which uses historical and static features without forecasts or retrieval; we train a separate control for each reported backbone configuration. **Solo** adds a point forecast from the corresponding forecasting backbone to the mortality-model input but has no quantile-width feature or retrieval ($\lambda{=}0$). **Quantile-only** uses the forecast median and predicted quantile-interval widths while disabling retrieval ($\lambda{=}0$). Full TRACE adds the neighbor-based outcome estimate through entropy-conditioned probability interpolation. The direct model can learn associations between observed histories and future outcomes; the comparison tests whether providing explicit forecast-derived features and retrieved outcomes changes predictive performance in this implementation.

We evaluate[3] five forecasting backbones trained from scratch—PatchTST (Nie et al., 2023), iTransformer (Liu et al., 2024), Crossformer (Zhang & Yan, 2023), TimeMixer (Wang et al., 2024a), and DLinear (Zeng et al., 2023)—and two pretrained forecasting backbones, Chronos (Ansari et al., 2024) and Moirai (Woo et al., 2024). For Chronos and Moirai, the pretrained forecasting backbone is frozen; their predictive samples supply quantile estimates, while the downstream mortality model is trained for this task. Implementation details are given in Appendix D and E.

[1] Available at `https://physionet.org/content/mimiciv/3.1/`.

[2] Available at `https://physionet.org/content/eicu-crd/2.0/`.

[3] All models were trained and evaluated on NVIDIA A100 GPU.

Table 1: Matched variants at T/H=36/12 hours on MIMIC-IV and eICU-CRD test sets. TRACE uses k=30. † Frozen pretrained forecasters with supervised downstream mortality classifiers.

Backbone	Direct Classification			Solo			TRACE (k=30)		
	AUROC	AUPRC	F1	AUROC	AUPRC	F1	AUROC	AUPRC	F1
MIMIC-IV									
iTransformer	0.7421	0.3650	0.4080	0.7681	0.3794	0.4436	**0.8462**	**0.5167**	**0.5223**
PatchTST	0.7104	0.2938	0.3589	0.7377	0.3160	0.3927	0.8113	0.4382	0.4590
Crossformer	0.7174	0.3058	0.3716	0.7346	0.3195	0.3944	0.8087	0.4405	0.4629
TimeMixer	0.7229	0.3049	0.3712	0.7382	0.3250	0.3924	0.8100	0.4390	0.4599
DLinear	0.7216	0.3017	0.3698	0.7342	0.3129	0.3933	0.8105	0.4389	0.4666
Chronos†	0.7332	0.3180	0.3820	0.7407	0.3246	0.3929	0.8075	0.4332	0.4511
Moirai†	0.7123	0.2964	0.3618	0.7355	0.3235	0.3910	0.8107	0.4439	0.4568
eICU-CRD									
iTransformer	0.7133	0.1112	0.2365	0.7318	0.1281	0.2508	0.7810	0.2331	0.2832
PatchTST	0.7125	0.1098	0.2341	0.7257	0.1229	0.2454	**0.7844**	**0.2346**	0.2911
Crossformer	0.7035	0.1049	0.2296	0.7117	0.1176	0.2408	0.7802	0.2344	0.2781
TimeMixer	0.7011	0.1028	0.2271	0.7130	0.1188	0.2429	0.7816	0.2345	0.2832
DLinear	0.7046	0.1061	0.2318	0.7111	0.1180	0.2428	0.7790	0.2313	0.2816
Chronos†	0.7101	0.1087	0.2384	0.7126	0.1164	0.2495	0.7825	0.2312	**0.2976**
Moirai†	0.7001	0.1015	0.2254	0.7098	0.1117	0.2425	0.7794	0.2277	0.2736

Table 2: Independently implemented mortality classifiers on the matched 36-hour MIMIC-IV task. All models use the same cohort, patient-level split, observation cutoff, and in-hospital mortality endpoint; inputs are preprocessed for each model. TRACE uses iTransformer as its forecasting backbone and k=30.

Model	AUROC	AUPRC	F1
Logistic Regression	0.8191	0.4575	0.4618
XGBoost	0.8405	0.5115	0.4958
GRU-D	0.8242	0.4760	0.4683
Medformer	0.6943	0.2723	0.3588
TRACE–iTransformer	**0.8462**	**0.5167**	**0.5223**

For a separate task-level comparison, we evaluate Logistic Regression, XGBoost (Chen & Guestrin, 2016), GRU-D (Che et al., 2018), and Medformer (Wang et al., 2024b) on MIMIC-IV. These independently implemented mortality classifiers use the same eligible cohort, patient split, first 36 hours of allowable observations, and in-hospital mortality endpoint, with model-specific input preprocessing. This comparison places TRACE in the context of alternative predictors on the matched task; the internal variants examine the pipeline components.

5.2 Main Results

Table 1 compares TRACE with the observed-history and point forecast controls at T/H=36/12 across seven forecasting backbones on both datasets. At k=30, TRACE has higher AUROC and AUPRC than both controls in every reported backbone configuration. The pattern is also observed with frozen pretrained forecasting backbones, although the downstream mortality predictors are trained on the respective datasets. These comparisons demonstrate consistency for the evaluated forecasters. We assess that comparison in Section 5.3 and report the remaining per-backbone Quantile-only results in Appendix E.8. We also compare TRACE–iTransformer with independently implemented classifiers on the matched 36-hour MIMIC-IV task in Table 2.

Table 3: Stepwise comparison on MIMIC-IV at T/H=36/12 hours. Forecasting variants use iTransformer; TRACE uses k=30. The Direct Classification model is an observed-history MLP.

Variant	AUROC	AUPRC	F1
A1: Direct Classification (observed history)	0.7421	0.3650	0.4080
A2: Solo (point forecast)	0.7681	0.3794	0.4436
A3: Quantile-only (median and interval width)	0.7882	0.3941	0.4336
TRACE (A3 + retrieved outcomes)	**0.8462**	**0.5167**	**0.5223**

5.3 Ablation Study

To examine the changes associated with each stage, we compare four configurations on the MIMIC-IV test set, using iTransformer for the forecasting variants and k=30 for TRACE (Table 3). The variants add a point forecast, quantile-derived inputs, and finally neighbor-outcome interpolation.

Adding point forecasting (A1→A2) yields ΔAUROC= $+0.0260$, showing the predictive value of explicitly forecasting near-future physiology in this setting. Moving from point forecasts to quantile-derived features (A2→A3) yields a further ΔAUROC= $+0.0201$, indicating that the quantile-based configuration provides additional information for risk discrimination. Finally, incorporating retrieved patient outcomes (A3→TRACE) produces the largest AUROC gain (ΔAUROC= $+0.0580$) and increases AUPRC by 0.1226. Together, these matched ablations show incremental AUROC improvements as forecasting, quantile-derived features, and outcome retrieval are introduced into the TRACE pipeline. Quantile-only results for all seven backbones on both datasets are provided in Appendix E.8.

5.4 Decision-Time and Forecast-Horizon Sensitivity

We evaluate four decision times, $T \in \{24, 30, 36, 42\}$ hours, with $H = 48 - T$, using iTransformer for the forecasting variants and $k = 30$ for TRACE. At each T, all variants use the same first T hours of observations. The primary $36/12$ results appear in Tables 1 and 3; complete results are in Appendix E.6. At the additional decision times $T = 24, 30, 42$ hours, TRACE improves AUROC over the matched Direct Classification controls by $+0.049, +0.050, +0.059$ on MIMIC-IV and $+0.040, +0.050, +0.054$ on eICU-CRD, respectively. The corresponding AUPRC gains are $+0.040, +0.039, +0.043$ and $+0.060, +0.079, +0.100$. These comparisons show consistent gains at multiple observation cutoffs within the evaluated cohort of ICU stays lasting at least 48 hours.

5.5 Quality of Mortality-Risk Probabilities

To assess the quality of the final mortality-risk estimates alongside AUROC and AUPRC, we evaluate Brier score, log-loss, and expected calibration error (ECE) at T/H=36/12 (Table 20). The forecast-based variants are summarized across seven forecasting backbones, while Direct Classification represents one fitted observed-history control. TRACE achieves the lowest mean value on all three metrics among the evaluated variants on both datasets. Its mean Brier score, log-loss, and ECE are $(0.1483, 0.4473, 0.1885)$ on MIMIC-IV and $(0.1201, 0.3920, 0.2284)$ on eICU-CRD. These results show that the improvements in risk discrimination are accompanied by improvements in the reported probability-quality metrics.

6 Discussion

6.1 What Forecasts and Retrieved Outcomes Add

The matched-decision-time comparisons show how explicit forecasts and retrieved patient outcomes contribute to mortality-risk prediction. In the iTransformer ablation, point forecasting improves discrimination over the observed-history classifier, quantile-derived features provide a further gain, and adding retrieved outcomes yields the largest incremental improvement (Table 3). These results support using predicted physiology as an input to the risk model and combining its estimate with observed outcomes from similar training patients. The point-forecast and quantile configurations

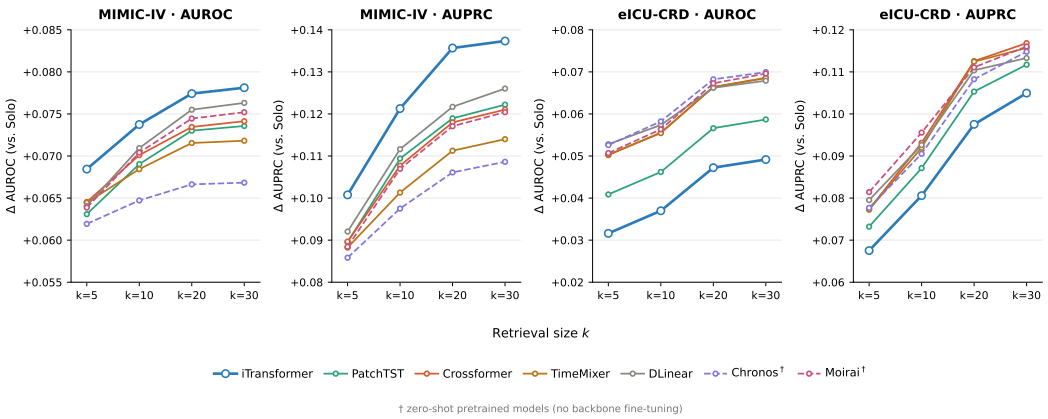


Figure 3: Sensitivity to retrieval size k at T/H=36/12 hours. The four panels show differences in AUROC and AUPRC between full TRACE and the corresponding Solo model for seven forecasting backbones on MIMIC-IV and eICU-CRD at $k \in \{5, 10, 20, 30\}$. Solo uses a point forecast; full TRACE uses quantile-derived features and retrieved outcomes. Chronos and Moirai are evaluated as frozen pretrained forecasting backbones.

represent different forecasting approaches, so the latter comparison reflects the quantile-based configuration as a whole.

Figure 3 shows gains from the full pipeline over Solo across all seven backbones and both cohorts. These differences increase from $k = 5$ to $k = 30$. For both metrics and cohorts, the improvement from $k = 20$ to $k = 30$ is smaller than that from $k = 10$ to $k = 20$ across the seven backbones. The iTransformer Quantile-only comparison separately shows that retrieved outcomes improve discrimination on top of quantile-derived features (Table 3 and Appendix E.8). Together, the comparisons show that forecast-derived inputs and outcome retrieval play complementary roles in the evaluated pipeline. The forecast diagnostics in Appendix G.2 show pooled coverage near the nominal 80% level for the five trained backbones. This characterizes the predicted intervals as summaries of marginal future vital-sign values at each hour and variable, which the classifier uses through their median and interval width.

6.2 Probability Quality and Post-hoc Forecast Behavior

The gains in discrimination are accompanied by lower mean Brier score, log-loss, and ECE for TRACE than for the evaluated internal variants on both datasets (Table 20). These metrics assess the final mortality-risk probabilities and complement the forecast-interval diagnostics for physiological values. Post-hoc interval coverage offers a second view of forecast behavior. Across the five trained backbones on both datasets, observed future vital signs of surviving patients fall inside the predicted intervals more often than those of patients who died (Figure 4). This outcome-stratified pattern indicates that forecast-interval behavior differs between the two groups. Coverage Score is computed from measurements recorded after the decision time for this retrospective analysis, while the prediction-time classifier uses only forecast-derived features.

7 Conclusion

We introduced TRACE, which combines quantile-derived forecasts of near-future vital signs with a risk estimate based on the observed outcomes of similar training patients. On MIMIC-IV and eICU-CRD, TRACE achieved higher AUROC than our observed-history MLP controls across seven forecasting backbones at the same decision time. On the matched MIMIC-IV task, TRACE obtained the highest observed scores among the independently implemented mortality classifiers. These findings support explicit physiological forecasts and outcome retrieval as useful additions under the evaluated cohorts and decision times.

AI USE STATEMENT

Generative AI tools were used during the preparation of this paper for drafting and editorial refinement of text, including paraphrasing and grammar correction; feedback on research framing, experimental comparisons, and interpretation of results; and drafts of illustrative figures and associated plotting code. The authors reviewed and revised AI-assisted material. The TRACE framework and its model components were developed and implemented by the authors, who ran the reported experiments and verified the results. The authors take responsibility for the final text, figures, code, and scientific claims.

ETHICS STATEMENT

This retrospective study uses de-identified MIMIC-IV and eICU-CRD records under the applicable credentialed-access agreements. No patient-level records are redistributed with this paper. Mortality-risk estimates learned from historical care may reflect variation in measurement practices, case mix, and treatment across patients and institutions. The evaluated cohorts include ICU stays of at least 48 hours, and the reported results have not been prospectively evaluated for clinical use. Subgroup and prospective evaluation would be needed before applying the model to patient care.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe the task, model, and evaluation protocol. Appendix A documents cohort construction, input variables, and preprocessing; Appendices C–E give the embedding, forecasting, retrieval, and classifier details. Additional decision-time, ablation, probability-quality, and forecast diagnostics are reported in Appendices E–G. The source ICU datasets are available through their providers subject to credentialed access.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda-Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Trans. Mach. Learn. Res.*, 2024. ISSN 2835-8856. doi: 10.48550/arXiv.2403.07815. URL https://openreview.net/forum?id=gerNCVqqtR.

Payal Chandak, Gregory Kondas, Isaac Kohane, and Matthew B. A. McDermott. EveryQuery: Zero-shot clinical prediction via task-conditioned pretraining over electronic health records, 2026. URL https://arxiv.org/abs/2603.07900.

Ping Chang, Huayu Li, Stuart F. Quan, Shuyang Lu, Shu-Fen Wung, Janet Roveda, and Ao Li. A transformer-based diffusion probabilistic model for heart rate and blood pressure forecasting in intensive care unit. *Computer Methods and Programs in Biomedicine*, 246:108060, 2024. doi: 10.1016/j.cmpb.2024.108060.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Charilaos Chourpiliadis and Abhishek Bhardwaj. Physiology, respiratory rate. In *StatPearls*. StatPearls Publishing, Treasure Island, FL, 2022. URL https://www.ncbi.nlm.nih.gov/books/NBK537306/. Updated 2022 Sep 12. PMID: 30725991.

Matthew M. Churpek, Trevor C. Yuen, Christopher Winslow, David O. Meltzer, Michael W. Kattan, and Dana P. Edelson. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical Care Medicine*, 44(2):368–374, 2016. doi: 10.1097/CCM.0000000000001571.

A. Duggal, Rachel G. Scheraga, G. Sacha, Xiao-Feng F. Wang, Shuaqui Huang, S. Krishnan, M. Siuba, Heather Torbic, S. Dugar, Simon Mucha, et al. Forecasting disease trajectories in critical illness: Comparison of probabilistic dynamic systems to static models to predict patient status in the intensive care unit. *BMJ Open*, 14(2):e079243, 2024. doi: 10.1136/bmjopen-2023-079243.

Cory Franklin and Joseph Mathew. Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22(2):244–247, 1994. doi: 10.1097/00003246-199402000-00014.

J. le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *Journal of the American Medical Association (JAMA)*, 270(24):2957–2963, 1993. doi: 10.1001/jama.1993.03510240069035.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. doi: 10.1038/s41597-019-0103-9.

Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE*, 16(7):e0254841, 2021. doi: 10.1371/journal.pone.0254841.

Zilin Jing, V. Jeanselme, Y. Kobayashi, Simon A. Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xin-Zhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured EHR foundation models, 2026. URL https://arxiv.org/abs/2602.00541.

A. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, S. Horng, T. Pollard, Sicheng Hao, Benjamin Moody, B. Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Alan E. Jones, Stephen Trzeciak, and Jeffrey A. Kline. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical Care Medicine*, 37(5):1649–1654, 2009. doi: 10.1097/CCM.0b013e31819def97.

Amal Jubran. Pulse oximetry. *Critical Care*, 3(2):R11–R17, 1999. doi: 10.1186/cc341.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.00172.

William A. Knaus, Elizabeth A. Draper, Douglas P. Wagner, and Jack E. Zimmerman. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985. doi: 10.1097/00003246-198510000-00009.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Andrew J. Loza, Jun Yup Kim, Shangzheng Song, Yihang Liu, Joseph J. Y. Sung, R. Andrew Taylor, and Dennis L. Shung. multivariateGPT: A decoder-only transformer for multivariate categorical and numeric data, 2025. URL https://arxiv.org/abs/2505.21680.

Matthew B. A. McDermott, Bret Nestor, Peniel Argaw, and Isaac S. Kohane. Event stream GPT: A data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. In *Advances in Neural Information Processing Systems 36*, pp. 24322–24334. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1057. URL https://arxiv.org/abs/2306.11547.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2211.14730. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Nassim Oufattole, Teya Bergamaschi, Aleksia Kolo, Hyewon Jeong, Hanna Gaggin, Collin M. Stultz, and Matthew B. A. McDermott. MEDS-Tab: Automated tabularization and baseline methods for MEDS datasets, 2024. URL https://arxiv.org/abs/2411.00200.

Chao Pang, Jiheum Park, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Shalmali Joshi, Noémie Elhadad, and Karthik Natarajan. CEHR-XGPT: A scalable multi-task foundation model for electronic health records, 2025. URL https://arxiv.org/abs/2509.03643.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database. *PhysioNet*, 2019. doi: 10.13026/C2WM1R. URL https://physionet.org/content/eicu-crd/2.0/. Version 2.0.

A. Rajkomar, E. Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018. doi: 10.1038/s41746-018-0029-1.

Pawel Renc, Michal K. Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, S. Biegański, Matthew B. A. McDermott, J. Was, Anthony E. Samir, J. Cunningham, et al. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14, 2025. doi: 10.1093/gigascience/giaf107. URL https://doi.org/10.1093/gigascience/giaf107.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36 (3):1181–1191, 2020. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2019.07.001. URL https://www.sciencedirect.com/science/article/pii/S0169207019301888.

C. Seymour, J. Kennedy, Shu Wang, Chung-Chou H. Chang, Corrine F. Elliott, Zhongying Xu, S. Berry, G. Clermont, Gregory F. Cooper, H. Gómez, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *Journal of the American Medical Association (JAMA)*, 321(20):2003–2017, 2019. doi: 10.1001/jama.2019.5791.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Journal of the American Medical Association (JAMA)*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

M. E. Beth Smith, J. Chiovaro, M. O'Neil, D. Kansagara, Ana R. Quiñones, M. Freeman, M. Motu'apuaka, and C. Slatore. Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014. doi: 10.1513/AnnalsATS.201403-102OC.

Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b36554b97da741b1c48c9de05c73993e-Abstract-Conference.html.

Christian Peter Subbe and John Robert Welch. Failure to rescue: Using rapid response systems to improve care of the deteriorating patient in hospital. *Clinical Risk*, 19(1):6–11, 2013. doi: 10.1177/1356262213486451.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6):Article 105, 2022. doi: 10.1145/3516367.

J. Vincent, R. Moreno, J. Takala, S. Willatts, A. de Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996. doi: 10.1007/BF01709751.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/a7ac8a21e5a27e7ab31a5f42a0117bdb-Abstract-Conference.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh (eds.), *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems 37*, volume 37, pp. 36314–36341. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024b. doi: 10.52202/079017-1145. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Shane Waxler, Paul J. Blazek, D. White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, P. Williams, H. Voeller, Karen K. Wong, Matthew Swanhorst, et al. Generative medical event models improve with scale, 2025. URL https://arxiv.org/abs/2508.12104.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Michael Wornow, Suhana Bedi, Miguel Hernandez, Ethan H. Steinberg, Jason Fries, Christopher Ré, Sanmi Koyejo, and Nigam Haresh Shah. Context clues: Evaluating long context models for clinical prediction tasks on EHRs, 2024. URL https://arxiv.org/abs/2412.16178.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. *AAAI Conference on Artificial Intelligence*, 36(8):8980–8987, Jun. 2022. doi: 10.1609/aaai.v36i8.20881. URL https://ojs.aaai.org/index.php/AAAI/article/view/20881.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Knowledge Discovery and Data Mining*, pp. 3273–3285. ACM, ACM, 2023. doi: 10.1145/3580305.3599543.

Ping Zhang, Fei Wang, Jianying Hu, and Robert Sorrentino. Towards personalized medicine: Leveraging patient similarity and drug similarity analytics. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2014:132–6, 2014.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

A Dataset and Preprocessing

A.1 Cohort Construction and Data Splits

MIMIC-IV. We extract ICU stays from MIMIC-IV v3.1 (Johnson et al., 2023) using the following inclusion criteria: (i) minimum ICU length of stay of 48 hours (`MIN_LOS_DAYS`$= 2.0$), ensuring a sufficient observation window plus forecast horizon; and (ii) availability of at least one measurement for each of the three forecasting target variables (HR, RespRate, SpO2) within the observation window. Stays are split at the *subject* level (not stay level) to prevent data leakage across train, validation, and test sets, using an 80/10/10 ratio (`RANDOM_STATE`$= 42$). The resulting cohort comprises 46,236 ICU stays (train: 36,942, val: 4,651, test: 4,643), with an overall in-hospital mortality rate of 14.72% (train: 14.76%, val: 14.99%, test: 14.11%).

eICU-CRD. We extract ICU stays from eICU-CRD v2.0 (Pollard et al., 2019) applying the same 48-hour minimum LOS criterion. Stays are split at the *uniquepid* level (patient level) to prevent leakage across the 200+ hospital cohort. The resulting cohort comprises 73,576 ICU stays (train: 58,899, val: 7,338, test: 7,339), with an overall in-hospital mortality rate of 6.51% (train: 6.51%, val: 6.64%, test: 6.35%).

Normalization statistics. For MIMIC-IV, normalization statistics (per-feature min/max) are computed from the training split only and applied to validation and test splits. The eICU-CRD models are trained and evaluated on the eICU-CRD cohort separately; these experiments are not a zero-shot transfer evaluation of a MIMIC-IV-trained model. Normalization is fitted on the respective training split and applied to its validation and test splits.

Table 4 summarizes cohort statistics for both datasets.

Table 4: Cohort statistics for MIMIC-IV and eICU-CRD. Splits are performed at the subject/patient level to prevent data leakage.

Dataset	Split	# Stays	# Deaths	Mortality (%)
MIMIC-IV	Train	36,942	5,454	14.76
	Val	4,651	697	14.99
	Test	4,643	655	14.11
	Total	46,236	6,806	14.72
eICU-CRD	Train	58,899	3,835	6.51
	Val	7,338	487	6.64
	Test	7,339	466	6.35
	Total	73,576	4,788	6.51

A.2 Missing Value Imputation

Temporal resolution. All clinical measurements are binned into 1-hour intervals aligned to ICU admission time. Vital signs and vasopressors are represented by the *last observed value* within each 1-hour bin (hourly resolution). Laboratory values are represented by the *last observed value* within each calendar day (daily resolution); if no measurement exists for the current day, the most recent prior measurement is carried forward (longitudinal forward-fill within the observation window).

Missing value handling. To encode missingness structure explicitly, the enriched representation $V_i^{\text{rich}} \in \mathbb{R}^{T \times 3D}$ concatenates three channels per feature:

- **Value** (D dims): imputed value (forward-filled within day for labs; last-observed for vitals).
- **Observation count** (D dims): number of actual measurements in the current 1-hour bin (0 = missing).
- **Hours since last observation** (D dims): elapsed time since the most recent actual measurement for each feature (`HOURS_SINCE_MAX`$= 99.0$ if never observed).

The count and recency channels distinguish observed measurements from values carried forward within the available history. No measurement recorded after the prediction cutoff enters these channels.

A.3 NORMALIZATION

Outlier clipping. Before normalization, each feature is clipped to physiologically plausible hard bounds. For most variables, bounds are derived from established ICU reference ranges; for eight variables with heavy-tailed distributions in extreme critical illness (Creatinine, BUN, WBC, Platelets, TotalBili, AST, ALT, ALP), we adopt the broader bounds from Warpformer (Zhang et al., 2023) to avoid discarding rare-but-real extreme values (e.g., high-grade AKI, leukemia, severe hepatitis). Table 5 lists the specified clipping ranges for the selected features; the complete 61-variable input set appears in Table 6.

Feature normalization. Following outlier clipping, each feature is normalized via per-feature min-max scaling computed from the training split:

$$\tilde{x} = \frac{x - \min_{\text{train}}}{\max_{\text{train}} - \min_{\text{train}} + \epsilon}, \tag{10}$$

where $\epsilon = 10^{-6}$ prevents division by zero.

Static feature normalization. Continuous static features (age, height, weight) are normalized using training-split statistics: missing values are imputed with the training median before normalization.

A.4 VITAL SIGN AND LABORATORY FEATURES

Table 6 lists all 61 input features grouped by clinical category. The three forecasting targets (HR, RespRate, SpO2) are selected based on their high hourly observation rates in the MIMIC-IV training split (HR: 95.0%, RespRate: 94.0%, SpO2: 93.5%), ensuring sufficient signal density for quantile forecasting. In eICU-CRD, the same targets retain the highest observation rates among vital signs (HR: 77.2%, RespRate: 71.6%, SpO2: 65.6%), though overall rates are lower due to the multi-center heterogeneity of the cohort. All remaining features serve as context only and are included in the observation window V_i but excluded from the forecasting target Y_i.

A.5 STATIC FEATURES

The processed static vector $S_i \in \mathbb{R}^{33}$ used by the classifier contains available demographic, admission, and ICU-unit information. Inputs are constructed only from information available by the evaluated decision time. The same encoded dimensionality is used in the embedding and classifier pathways (Appendices C.1 and E.1).

B DETERIORATION-LABEL CONSTRUCTION

The auxiliary deterioration head in Appendix D.3 is trained from the observed future values of the three forecast targets: heart rate (HR), respiratory rate (RespRate), and oxygen saturation (SpO_2). These future measurements supply a *training label*.

For each variable, abnormal burden is the fraction of its *observed* future hourly values outside the corresponding reference range. The binary abnormality label for that variable is one when this fraction exceeds 0.30. Missing future hours are excluded from the denominator; variables without an observed future value receive no supervised loss. This definition applies to the auxiliary forecaster head only.

Table 5: Selected outlier clipping ranges for clinical features. Ranges marked $*$ are updated from original clinical references to Warpformer bounds (Zhang et al., 2023) to accommodate extreme critical illness values.

Category	Feature	Min	Max
Vitals	HR	30	220
	SysABP	40	250
	DiasABP	20	180
	RespRate	5	80
	SpO2	50	100
	Temp_C	30	42
	Temp_F	86	108
Blood gas	PaO2	20	500
	PaCO2	10	150
	ArterialPH	6.8	7.8
	VenousPH	6.8	7.8
	SaO2_art	50	100
	HCO3	5	50
	AnionGap	2	40
	LacticAcid	0	20
BMP (selected)	Creatinine*	0	66
	BUN*	0	275
	Sodium_serum	100	180
	Potassium_serum	1.5	8.0
	Glucose_serum	20	1000
	Magnesium	0.5	6.0
CBC	WBC*	0	1100
	Platelets*	0	2200
	Hemoglobin	2	25
	HCT_serum	5	80
	HCT_calc	5	80
LFT	AST*	0	22000
	ALT*	0	11000
	ALP*	0	4000
	TotalBili*	0	66
Vasopressors	Epinephrine	0	200
	Norepinephrine	0	200
	Dopamine	0	200
	Phenylephrine	0	200
	Vasopressin	0	10

C Patient Retrieval Embedding

C.1 Architecture and Pretraining

Temporal encoder. The temporal encoder f_{ts} maps the enriched representation $V_i^{\text{rich}} \in \mathbb{R}^{T\times 183}$ to a 256-dimensional sequence embedding. The architecture follows TS2Vec (Yue et al., 2022): a linear input projection ($183 \to 256$) followed by $N{=}6$ dilated temporal convolution blocks with exponentially increasing dilation factors $(1, 2, 4, 8, 16, 32)$, and temporal mean pooling to obtain a patient-level representation. Each dilated block applies a two-layer residual structure:

$$\mathbf{h} \leftarrow \mathrm{LayerNorm}\big(\mathbf{h} + \mathrm{Conv}_1(\mathrm{GELU}(\mathrm{Dropout}_{0.1}(\mathrm{Conv}_d(\mathbf{h}))))\big), \tag{11}$$

where Conv_d denotes a dilated causal convolution with kernel size 3 and dilation d, and Conv_1 is a pointwise projection back to the residual dimension.

Static encoder. The static encoder f_{static} processes the same processed static vector $S_i \in \mathbb{R}^{33}$ described in Appendix A.5, producing a 64-dimensional representation.

Patient embedding. The full patient embedding $z_i \in \mathbb{R}^{256}$ is produced by concatenating the temporal and static representations and passing through a two-layer fusion MLP ($320 \to 256 \to 256$), followed by L2 normalization:

$$z_i = \frac{\mathrm{MLP}_{\mathrm{fuse}}\big([f_{\mathrm{ts}}(V_i^{\mathrm{rich}});\ f_{\mathrm{static}}(S_i)]\big)}{\|\cdot\|_2}. \tag{12}$$

In Stage 1, the static encoder and fusion MLP are absent; the temporal encoder output is passed directly through a two-layer projection head ($256 \to 256 \to 256$) for contrastive pretraining.

Stage 1 pretraining. The temporal encoder is pretrained via TS2Vec hierarchical contrastive learning on the augmented ICU time series. For each patient, two augmented views are generated (see Appendix D.4) and the encoder is trained to produce similar per-timestep embeddings for the same patient (temporal CL) and similar pooled embeddings (instance CL):

$$\mathcal{L}_{\mathrm{Stage1}} = \alpha\mathcal{L}_{\mathrm{temporal}} + (1-\alpha)\mathcal{L}_{\mathrm{instance}}, \quad \alpha = 0.5. \tag{13}$$

Training hyperparameters are summarized in Table 8.

C.2 SOFA Proxy Computation

Severity proxy and outcome supervision. An approximate, SOFA-inspired severity score computed from the available observation window supplies one source of similarity supervision. The embedding also receives mortality-label supervision through the within-proxy-bin outcome loss and an auxiliary BCE head below. Retrieved training-patient outcomes are subsequently used to form a separate neighbor-based risk estimate.

Five-domain SOFA-inspired proxy. We compute a continuous 5-domain SOFA proxy $\hat{s}_i \in [0,5]$ from the last available observation at or before the decision time for each relevant feature in $V_i \in \mathbb{R}^{T\times 61}$:

$$\hat{s}_i = \underbrace{(1-\tilde{x}_{\mathrm{PaO2}})}_{\text{respiratory}} + \underbrace{\tilde{x}_{\mathrm{Creatinine}}}_{\text{renal}} + \underbrace{\tilde{x}_{\mathrm{TotalBili}}}_{\text{hepatic}} + \underbrace{(1-\tilde{x}_{\mathrm{Platelets}})}_{\text{coagulation}} + \underbrace{\max_{j\in\mathcal{V}} \tilde{x}_j}_{\text{cardiovascular}}, \tag{14}$$

where $\tilde{x} \in [0,1]$ denotes the min-max normalized feature value, $\mathcal{V}$ indexes the five vasopressor features (Epinephrine, Norepinephrine, Dopamine, Phenylephrine, Vasopressin), and values are clipped to $[0,5]$. Respiratory and coagulation domains are inverted ($1-\tilde{x}$) so that higher values uniformly indicate worse physiological status. This is a heuristic proxy rather than a clinical SOFA score: the neurological domain (GCS) is omitted, and its components are based on normalized laboratory and vasopressor features.

SOFA normalization. The proxy $\hat{s}_i$ is standardized using training-split mean and standard deviation before being used as supervision signal (μ_{SOFA}, σ_{SOFA} computed on the training split and applied to validation and test splits).

C.3 Composite Contrastive Objective

Stage 2 loss formulation. The Stage 2 fine-tuning loss combines four terms across two training phases:

$$\mathcal{L}_{\mathrm{total}} = (1-\alpha_{\mathrm{bce}})\big[\alpha_{\mathrm{sofa}}\mathcal{L}_{\mathrm{sofa}} + \alpha_{\mathrm{outcome}}\mathcal{L}_{\mathrm{outcome}} + \alpha_{\mathrm{uniform}}\mathcal{L}_{\mathrm{uniform}}\big] + \alpha_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}}. \tag{15}$$

Soft SupCon loss $\mathcal{L}_{\mathrm{sofa}}$. SOFA-similarity-weighted contrastive loss with temperature τ=0.2 and SOFA kernel width σ=0.3:

$$w_{ij} = \exp\left(-\frac{(\hat{s}_i - \hat{s}_j)^2}{2\sigma^2}\right), \quad \mathcal{L}_{\mathrm{sofa}} = -\frac{1}{B}\sum_i \frac{\sum_j w_{ij}\log p_{ij}}{\sum_j w_{ij}}, \tag{16}$$

where $p_{ij} = \exp(\langle z_i, z_j\rangle/\tau)/\sum_{k\neq i}\exp(\langle z_i, z_k\rangle/\tau)$.

Intra-SOFA outcome loss $\mathcal{L}_{\text{outcome}}$. Applied *only within* SOFA bins to introduce outcome separation without globally collapsing the embedding space. The proxy $\hat{s}$ is quantized into n_{bins}=5 bins using batch-level quantiles. Within each bin, patients with the same outcome label are attracted and those with different labels serve as hard negatives, with temperature τ_{out}=0.1:

$$\mathcal{L}_{\text{outcome}} = -\frac{1}{|V|} \sum_{i \in V} \frac{\sum_{j \in \mathcal{P}(i)} \log p_{ij}^{\text{bin}}}{|\mathcal{P}(i)|}, \tag{17}$$

where V is the set of anchor samples with at least one positive in their bin, $\mathcal{P}(i)$ is the set of same-bin same-outcome positives, and p_{ij}^{bin} is the softmax probability restricted to same-bin pairs. Rows with fewer than `min_pairs`= 2 valid same-bin pairs are excluded.

Uniformity loss $\mathcal{L}_{\text{uniform}}$. Wang & Isola (2020) uniformity loss with Gaussian kernel bandwidth t=4.0:

$$\mathcal{L}_{\text{uniform}} = \log \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \exp\bigl(-t \cdot \|z_i - z_j\|^2\bigr), \tag{18}$$

where $\mathcal{P}$ is the set of all upper-triangular pairs in the batch and $\|z_i - z_j\|^2 = 2(1 - \langle z_i, z_j \rangle)$ for L2-normalized embeddings.

BCE loss $\mathcal{L}_{\text{bce}}$. A linear classifier head maps embeddings to mortality logits, trained with binary cross-entropy and positive class weight w_{pos}=9.0 to address class imbalance in this auxiliary objective. This fixed weight is distinct from the training-prevalence weight used for the final mortality classifier in Appendix E.3. This head is used only during Stage 2 training and discarded at inference time.

Two-phase training schedule. Training proceeds in two phases to prevent the randomly initialized static encoder from corrupting pretrained temporal representations. Hyperparameters for both phases are summarized in Table 9.

C.4 Embedding Quality Evaluation

Quantitative evaluation. Table 10 reports embedding quality metrics on the MIMIC-IV test set, comparing Stage 1 pretraining alone with the full Stage 2 embedding objective. We report kNN classification AUROC using the embedding alone (p_{kNN} with k=30) as a proxy for retrieval quality, alongside embedding-space alignment (mean cosine similarity within SOFA bins) and uniformity.

D Forecasting Model Details

D.1 Sliding Window Statistics

Construction. The forecasting model is trained on a sliding window dataset extracted from the same MIMIC-IV and eICU-CRD cohorts used for the embedding and classifier. For each ICU stay, windows are extracted with a step size of `slide_step`= 6 hours, subject to a maximum stay length cap of `max_los_hours`= 168 hours (7 days); stays longer than 7 days use only the first 168 hours. The primary setting uses a T=36h observation period followed by an H=12h forecast horizon; the additional matched decision times are reported in Appendix E.6.

Retention criterion. A window is retained for training only if all three forecasting target variables (HR, RespRate, SpO2) each have at least 30% observed (non-missing) measurements within the historical window (`MIN_OBS_FRACTION`= 0.30). Windows failing this criterion are discarded to avoid training the forecaster on sequences with insufficient target signal.

Inference. At inference time, each patient is represented by a single window starting at ICU admission (no sliding), ensuring that the observation window is anchored to the clinically relevant admission period.

D.2 Quantile Output Head

All five fine-tuned backbones are extended to produce predictions at $Q{=}5$ quantile levels $\mathcal{Q} = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ via a shared quantile projector appended to the backbone output. Two integration strategies are used depending on whether the backbone exposes a single output projection layer:

Internal quantile head (iTransformer, PatchTST). For backbones whose final layer is a single linear projection, the quantile head replaces the output projection:

$$\text{Linear}(d_{\text{model}},\ H \times Q) \xrightarrow{\text{reshape}} (B,\, Q,\, H,\, K), \tag{19}$$

where $H{=}12$, $Q{=}5$, and $K{=}3$.

Wrapper quantile projector (Crossformer, TimeMixer, DLinear). For backbones that do not expose a single output projection, a lightweight wrapper is attached to the backbone's prediction output without changing its internal architecture:

$$\hat{Y}_{\text{base}} \in \mathbb{R}^{B\times H\times K} \xrightarrow{\text{Linear}(K,\, K\times Q)} \mathbb{R}^{B\times H\times K\times Q} \xrightarrow{\text{permute}} (B,\, Q,\, H,\, K). \tag{20}$$

The wrapper linear layer uses Xavier uniform initialization and zero bias.

Deterioration head. All backbones include a lightweight deterioration head h_d that maps the encoder's final hidden state (or last input timestep for wrapper-based models) to per-variable abnormality logits (B, K), trained with $\mathcal{L}_D$ and discarded at inference time.

D.3 Composite Loss Hyperparameters

The forecasting model is trained with the composite objective:

$$\mathcal{L}_{\text{forecast}} = \mathcal{L}_Q + \lambda_T \mathcal{L}_T + \lambda_D \mathcal{L}_D. \tag{21}$$

Pinball loss $\mathcal{L}_Q$. The primary term is the Pinball (quantile) loss (Koenker & Bassett, 1978; Salinas et al., 2020) aggregated over all quantiles $q \in \mathcal{Q}$ and observed time steps:

$$\mathcal{L}_Q = \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \sum_{t,k} \mathbf{m}_{t,k} \cdot \rho_q\Big(Y_i[t,k] - \hat{Y}_i^{(q)}[t,k]\Big)\,, \tag{22}$$

where $\mathbf{m}_{t,k} \in \{0,1\}$ is the observation mask (accounting for missing values) and $\rho_q(u) = u \cdot \mathbf{1}[u \geq 0] \cdot q + u \cdot \mathbf{1}[u < 0] \cdot (q-1)$ is the standard pinball function.

Trend loss $\mathcal{L}_T$. Optimizing $\mathcal{L}_Q$ alone risks degenerate flat median predictions that correctly capture marginal distributions but fail to reproduce temporal dynamics. The trend loss penalizes errors in the first-order temporal difference of the median path:

$$\mathcal{L}_T = \sum_t \mathbf{m}_t \cdot \left\| \Delta\hat{Y}_i^{(0.5)}[t] - \Delta Y_i[t] \right\|^2, \qquad \Delta Z[t] = Z[t] - Z[t-1]. \tag{23}$$

This encourages the model to capture the *direction* of physiological change in the observed training targets. The mask is applied only to time differences for which both underlying measurements are available.

Deterioration loss $\mathcal{L}_D$. The deterioration loss attaches a lightweight head h_d to the backbone encoder and trains it to detect whether the future trajectory falls within clinically critical abnormal ranges:

$$\mathcal{L}_D = \text{BCE}(\sigma(h_d(\text{enc}(V_i)))\,,\ \mathbf{1}[Y_i \in \mathcal{A}])\,, \tag{24}$$

where $\mathcal{A}$ is the set of abnormality conditions defined in Appendix B, and h_d is discarded at inference time. This auxiliary signal encourages the encoder to develop representations sensitive to critical state transitions, which are directly relevant to mortality risk.

Loss weights. We set $\lambda_T = 1.0$ and $\lambda_D = 0.1$, applied uniformly across all five fine-tuned backbones to ensure fair comparison. λ_T=1.0 was selected to balance the scale of the pinball loss, and λ_D=0.1 was chosen to treat deterioration detection as an auxiliary signal without dominating the primary forecasting objective.

D.4 Training Augmentation

Overview. The forecasting backbone and Stage 1 embedding pretraining both apply data augmentation to the input time series to improve generalization. We use two augmentation strategies: for Stage 1 (self-supervised pretraining), each patient generates two independently augmented views that serve as a positive pair; for forecasting model training, a single augmented view is used per window.

Stage 1 two-view augmentation. Each view is constructed by applying the following four operations sequentially to the input $V_i^{\text{rich}} \in \mathbb{R}^{T\times 183}$:

1. A sub-sequence of length $l \sim \text{Uniform}(\lceil T/2 \rceil, T)$ is sampled from a random start position within the window. The selected sub-sequence is left-aligned and zero-padded to restore the original length T, preserving temporal ordering while introducing positional diversity.
2. Each timestep is independently masked with probability p_{mask}=0.15; masked timesteps have all feature values set to zero. This simulates missing-data patterns common in ICU recordings.
3. Each feature dimension is independently scaled by a factor drawn from $\text{Uniform}(0.8, 1.2)$, applied uniformly across all timesteps of that feature. This encourages scale-invariant temporal representations.
4. Additive Gaussian noise $\mathcal{N}(0, \sigma_j^2)$ with σ_j=0.05 is applied *only to observed (non-zero) timesteps*, to avoid corrupting the missingness structure encoded in the zero-padding from masking and cropping.

Two views are generated independently by applying all four operations separately to the same input, yielding a positive pair (v_1, v_2) for contrastive learning.

Forecasting model augmentation. For forecasting backbone training, only Gaussian jitter (σ_j=0.05) and random temporal shift are applied to the sliding window input, without cropping or masking, to avoid distorting the temporal structure needed for quantile prediction.

Augmentation hyperparameters. Table 11 summarizes all augmentation hyperparameters.

D.5 Per-Model Hyperparameters

Shared training settings. All five fine-tuned backbones use identical training settings to ensure fair comparison, as summarized in Table 12.

Backbone-specific hyperparameters. Table 13 reports architecture hyperparameters for each fine-tuned backbone.

Zero-shot pretrained models. Two foundation models are evaluated in zero-shot mode without any backbone fine-tuning, using only the quantile head (which is trained on top of frozen backbone outputs).

- **Chronos** (Ansari et al., 2024): we use `amazon/chronos-t5-large`. Each of the three target vitals is predicted independently as a univariate channel, then concatenated to form the multivariate output. Quantile predictions are derived from `num_samples`= 20 predictive samples per channel.
- **Moirai** (Woo et al., 2024): we use `Salesforce/moirai-1.1-R-large` with `target_dim`= 3 (joint multivariate prediction), `patch_size`= 4, and `num_samples`= 100. Quantile predictions are derived from the sample distribution via empirical quantile estimation.

E CLASSIFIER IMPLEMENTATION DETAILS

E.1 FEATURE DIMENSION BREAKDOWN

At the primary T/H=36/12 setting, the 117-dimensional feature vector $\mathbf{f}_i$ fed to the quantile-based model pathway MLP is constructed as:

$$\mathbf{f}_i = \left[\mathbf{h}_i;\ \hat{Y}_i^{(0.5)};\ \left(\hat{Y}_i^{(0.9)} - \hat{Y}_i^{(0.1)}\right);\ S_i\right] \in \mathbb{R}^{117}. \tag{25}$$

Table 14 details each component.

The components break down as follows:

- $\mathbf{h}_i \in \mathbb{R}^{12}$: historical vital sign statistics for the 3 target vitals (HR, RespRate, SpO2), comprising last observed value, mean, standard deviation, and recent trend per variable ($4 \times 3 = 12$).
- $\hat{Y}_i^{(0.5)} \in \mathbb{R}^{H \times K}$ (flattened to $\mathbb{R}^{36}$): the predicted median trajectory over H=12 hours for K=3 target vitals.
- $\hat{Y}_i^{(0.9)} - \hat{Y}_i^{(0.1)} \in \mathbb{R}^{36}$: the predicted interval width between the 0.1 and 0.9 quantiles; this width does not distinguish sources of uncertainty.
- $S_i \in \mathbb{R}^{d_s}$ (d_s=33): static features including age, gender, and admission laboratory values.

Trend computation. The trend statistic in $\mathbf{h}_i$ is computed as the difference between the mean of the last w=6 hours and the mean of the first w=6 hours of the observation window for each target vital:

$$\text{trend}_k = \frac{1}{w}\sum_{t=T-w}^{T-1} V_i[t,k] - \frac{1}{w}\sum_{t=0}^{w-1} V_i[t,k], \quad k \in \{1,\ldots,K\}. \tag{26}$$

Design rationale for pathway separation. The model pathway MLP operates solely on $\mathbf{f}_i$ and has no access to the patient embedding z_i or any retrieved neighbor information. This separates the model-pathway input from the retrieval output. The retrieval embedding itself is trained partly with mortality labels (Appendix C.3), so the pathways are not independent of outcome supervision.

E.2 MLP ARCHITECTURE

The model pathway p_{model} is a four-layer MLP with progressively narrowing hidden dimensions:

$$\mathbf{f}_i \xrightarrow{117\to256} \xrightarrow{256\to128} \xrightarrow{128\to64} \xrightarrow{64\to1} \ \sigma(\cdot) \ \to\ p_{\text{model}} \in (0,1). \tag{27}$$

Each hidden layer applies LayerNorm, GELU activation, and Dropout:

$$\mathbf{h}^{(l+1)} = \text{Dropout}\big(\text{GELU}(\text{LayerNorm}(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}))\big). \tag{28}$$

Dropout is applied after hidden layers at the rates specified in Table 15. All linear layers use Xavier uniform weight initialization and zero bias initialization. The output logit is passed through a sigmoid to produce the final mortality probability p_{model}.

E.3 TRAINING DETAILS

Optimizer and schedule. The MLP classifier is trained with Adam (β_1=0.9, β_2=0.999, weight decay 10^{-4}) with a linear learning rate warm-up over the first 3 epochs from 0 to η=10^{-3}, followed by cosine annealing to $\eta_{\min}$=10^{-5}. Training runs for up to 50 epochs with early stopping based on validation AUROC (patience=10).

Class imbalance handling. To address class imbalance, we apply a positive class weight in the BCE loss computed automatically from the training split:

$$w_{\text{pos}} = \frac{1 - \rho_{\text{train}}}{\rho_{\text{train}}}, \tag{29}$$

where ρ_{train} is the training-split mortality rate. For MIMIC-IV (ρ=14.76%) this yields $w_{\text{pos}}\approx 5.8$, and for eICU-CRD ($\rho$=6.51%) $w_{\text{pos}}\approx 14.4$.

All hyperparameters are summarized in Table 15.

E.4 Retrieval Pool Construction

Pool construction. The retrieval pool is restricted exclusively to the training split $\mathcal{D}_{\text{train}}$, ensuring that ground-truth labels y_j of retrieved neighbors are never observed during inference on validation or test patients. For training patients querying themselves, self-exclusion is applied: the query patient is removed from the pool before computing top-k neighbors.

Similarity computation. Cosine similarities between all query and pool embeddings are computed in GPU-accelerated batches (batch size 2048) using L2-normalized embeddings:

$$\text{sim}(z_i, z_j) = \frac{z_i}{\|z_i\|_2} \cdot \frac{z_j}{\|z_j\|_2} = \cos(z_i, z_j), \quad \text{sim} \in [-1, 1]. \tag{30}$$

kNN weighting and caching. Neighbors are retrieved up to $k_{\max}$=30 and cached to disk; any $k \leq k_{\max}$ can be evaluated by slicing the cached neighbor list without recomputing embeddings. Similarity-weighted mortality estimates use temperature τ=0.05:

$$w_j = \frac{\exp(-d_{ij}/\tau)}{\sum_{j' \in \mathcal{R}(i)} \exp(-d_{ij'}/\tau)}, \quad d_{ij} = 1 - \cos(z_i, z_j). \tag{31}$$

E.5 Additional Results

Full k-sensitivity results. Tables 16 and 17 report AUROC and AUPRC across all evaluated retrieval sizes $k \in \{5, 10, 20, 30\}$ for all seven backbones, providing the numerical values corresponding to Figure 3 in the main text. Direct Classification denotes an observed-history classifier without forecasting. The per-backbone columns report separately fitted controls, which explains why their values are not identical across rows. A2 (Solo) denotes the MLP classifier with single-path forecasting and no retrieval (λ=0).

E.6 Decision-Time and Forecast-Horizon Sensitivity

Tables 18 and 19 give the full matched comparisons underlying Section 5.4. At each decision time $T \in \{24, 30, 36, 42\}$ hours, the forecast horizon is H=48 $- T$. All variants have the same observed history cutoff, endpoint, and eligible cohort within each dataset. The $36/12$ rows repeat the primary-setting results to show the four time points together.

Table 6: Full list of 61 clinical features used as model input, grouped by category and temporal resolution. † Forecasting targets (HR, RespRate, SpO2).

Category	Features	Resolution	Count
Vital signs	HR†, SysABP, DiasABP, RespRate†, SpO2†, Temp_F, Temp_C	Hourly	7
Blood gas	PaO2, SaO2_art, PaCO2, VenousPH, VenousTCO2, ArterialPH, ArterialBaseExcess, ArterialTCO2, VenousPCO2, VenousPO2, AnionGap, HCO3, LacticAcid	Daily	13
BMP	Chloride, Creatinine, Glucose_serum, Sodium_serum, BUN, Calcium, Glucose_finger, Sodium_whole, Glucose_whole, Potassium_serum, Potassium_whole, Magnesium, Phosphate	Daily	13
CBC	Hemoglobin, HCT_serum, HCT_calc, WBC, Platelets	Daily	5
CBC differential	Basophils, Eosinophils, Lymphocytes, Monocytes, Neutrophils	Daily	5
Coagulation	PT, PTT, INR, Fibrinogen	Daily	4
LFT	AST, ALT, ALP, TotalBili	Daily	4
Cardiac enzymes	LDH, CK, CKMB, TroponinT, Albumin	Daily	5
Vasopressors	Epinephrine, Norepinephrine, Dopamine, Phenylephrine, Vasopressin	Hourly	5
Total			**61**

Table 7: Reference ranges for the three forecast variables used to construct the auxiliary abnormality label.

Variable	Reference range	Unit
HR	60–100	bpm
RespRate	12–20	breaths/min
SpO_2	95–100	%

Table 8: Stage 1 (self-supervised pretraining) hyperparameters.

Hyperparameter	Value
Epochs	100
Batch size	256
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
LR schedule	CosineAnnealing ($\eta_{\min}$=10^{-5})
Gradient clipping	max norm = 1.0
Early stopping patience	15 epochs
Temperature τ	0.07
Hidden / embed dim	256
Dilated conv layers	6

Table 9: Stage 2 (SOFA-supervised fine-tuning) hyperparameters. Phase 1 freezes the temporal encoder; Phase 2 unfreezes all parameters.

Hyperparameter	Phase 1 (warm-up)	Phase 2 (full fine-tuning)
Epochs	10	50
Batch size	128	
Optimizer	AdamW, weight decay 10^{-4}	
Learning rate	10^{-3}	10^{-4}
LR schedule	CosineAnnealing ($\eta_{\min}$=10^{-6})	
Gradient clipping	max norm = 1.0	
Frozen parameters	ts_encoder	none
Early stopping patience	—	10 epochs
α_{bce}	0.5	0.1
α_{sofa}	1.0	0.6
α_{outcome}	0.0	0.2
α_{uniform}	0.0	0.2
τ_{sofa} (SoftSupCon temperature)	0.2	
σ (SOFA kernel width)	0.3	
τ_{out} (IntraSOFA temperature)	0.1	
n_{bins} (SOFA quantization bins)	5	
t (Uniformity kernel bandwidth)	4.0	
w_{pos} (BCE positive weight)	9.0	

Table 10: Embedding quality on MIMIC-IV test set. kNN AUROC: retrieval-only classification using p_{kNN} (k=30, train pool $\rightarrow$ test query). Within-SOFA similarity: mean cosine similarity of pairs within the same SOFA bin. Uniformity: log mean pairwise Gaussian kernel (Wang & Isola, 2020) (lower is more uniform). The full Stage 2 objective increases retrieval-only AUROC by 0.108 and changes the embedding uniformity, while within-SOFA similarity decreases slightly in the reported comparison.

Embedding supervision	kNN AUROC ↑	Within-SOFA sim. ↑	Uniformity ↓
Stage 1 only (no fine-tuning)	0.6975	0.8067	−0.770
Stage 2 (full objective)	**0.8058**	0.7464	$\mathbf{-1.866}$

Table 11: Data augmentation hyperparameters for Stage 1 two-view generation and forecasting model training.

Augmentation	Parameter	Stage 1	Forecasting
Temporal crop	min crop ratio	0.5	—
Random masking	mask probability	0.15	—
Random scaling	scale range	[0.8, 1.2]	—
Gaussian jitter	std σ_j	0.05	0.05
Temporal shift	applied	No	Yes

Table 12: Shared training hyperparameters for all five fine-tuned forecasting backbones.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
LR warm-up epochs	5 (linear)
LR schedule	CosineAnnealing
Batch size	2048
Max epochs	100
Early stopping patience	20 (val loss)
Gradient clipping	max norm = 1.0
Quantiles $\mathcal{Q}$	$\{0.1, 0.25, 0.5, 0.75, 0.9\}$
λ_T	1.0
λ_D	0.1

Table 13: Architecture hyperparameters for each fine-tuned forecasting backbone. The primary setting predicts Q=5 marginal quantiles for each of K=3 target vitals over H=12 hours.

Backbone	Hyperparameter	Value
iTransformer (Liu et al., 2024)	d_{model}	512
	n_{heads}	8
	n_{layers}	4
	d_{ffn}	2048
	Dropout	0.1
PatchTST (Nie et al., 2023)	d_{model}	512
	n_{heads}	8
	n_{layers}	3
	d_{ffn}	1024
	Patch length	6h
	Stride	3h
	Dropout	0.2
Crossformer (Zhang & Yan, 2023)	d_{model}	512
	d_{ff}	1024
	n_{heads}	8
	e_{layers}	3
	Segment length	6h
	Window size	2
	Dropout	0.2
TimeMixer (Wang et al., 2024a)	d_{model}	128
	d_{ff}	256
	e_{layers}	3
	Down-sampling layers	2
	Down-sampling window	2
	Moving avg kernel	25
DLinear (Zeng et al., 2023)	Moving avg kernel size	25
	Channel independence	Yes

Table 14: Feature vector $\mathbf{f}_i$ breakdown (117 dimensions total).

Component	Description	Dim	Computation
$\mathbf{h}_i$	Historical vital statistics	12	$[\text{last}, \text{mean}, \text{std}, \text{trend}] \times K$=3 vitals
$\hat{Y}_i^{(0.5)}$	Predicted median trajectory	36	$H \times K = 12 \times 3$, flattened
$\hat{Y}_i^{(0.9)} - \hat{Y}_i^{(0.1)}$	Predicted uncertainty width	36	$(q_{0.9} - q_{0.1})$, flattened
S_i	Static features	33	encoded static information available by decision time
$\mathbf{f}_i$	Total	117	

Table 15: Classifier h_ψ training hyperparameters.

Hyperparameter	Value
MLP hidden dimensions	$256 \to 128 \to 64 \to 1$
Dropout	0.1 and 0.05 (hidden layers)
Weight initialization	Xavier uniform, bias zero
Optimizer	Adam
Learning rate	10^{-3}
LR warm-up epochs	3 (linear)
LR schedule	CosineAnnealing ($\eta_{\min}$=10^{-5})
Weight decay	10^{-4}
Batch size	512
Max epochs	50
Early stopping patience	10 (val AUROC)
Positive class weight	$(1-\rho)/\rho$ (auto-computed)

Table 16: Full k-sensitivity results on MIMIC-IV test set. AUROC (top) and AUPRC (bottom) reported for each backbone across Direct Classification, A2 (Solo), and $k \in \{5, 10, 20, 30\}$. † zero-shot pretrained models.

Backbone	Direct Class.	A2 (Solo)	k=5	k=10	k=20	k=30
AUROC						
iTransformer	0.7421	0.7681	0.8365	0.8418	0.8455	**0.8462**
PatchTST	0.7104	0.7377	0.8008	0.8067	0.8107	**0.8113**
Crossformer	0.7174	0.7346	0.7991	0.8047	0.8080	**0.8087**
TimeMixer	0.7229	0.7382	0.8027	0.8066	0.8097	**0.8100**
DLinear	0.7216	0.7342	0.7983	0.8051	0.8097	**0.8105**
Chronos†	0.7332	0.7407	0.8026	0.8054	0.8073	**0.8075**
Moirai†	0.7123	0.7355	0.7994	0.8059	0.8100	**0.8107**
AUPRC						
iTransformer	0.3650	0.3794	0.4802	0.5007	0.5151	**0.5167**
PatchTST	0.2938	0.3160	0.4055	0.4254	0.4350	**0.4382**
Crossformer	0.3058	0.3195	0.4091	0.4272	0.4375	**0.4405**
TimeMixer	0.3049	0.3250	0.4132	0.4263	0.4363	**0.4390**
DLinear	0.3017	0.3129	0.4049	0.4245	0.4346	**0.4389**
Chronos†	0.3180	0.3246	0.4104	0.4221	0.4307	**0.4332**
Moirai†	0.2964	0.3235	0.4119	0.4304	0.4406	**0.4439**

Table 17: Full k-sensitivity results on eICU-CRD test set. Same notation as Table 16.

Backbone	Direct Class.	A2 (Solo)	k=5	k=10	k=20	k=30
AUROC						
iTransformer	0.7133	0.7318	0.7634	0.7688	0.7790	**0.7810**
PatchTST	0.7125	0.7257	0.7666	0.7719	0.7823	**0.7844**
Crossformer	0.7035	0.7117	0.7619	0.7673	0.7781	**0.7802**
TimeMixer	0.7011	0.7130	0.7633	0.7684	0.7794	**0.7816**
DLinear	0.7046	0.7111	0.7639	0.7685	0.7773	**0.7790**
Chronos†	0.7101	0.7126	0.7652	0.7708	0.7808	**0.7825**
Moirai†	0.7001	0.7098	0.7605	0.7661	0.7771	**0.7794**
AUPRC						
iTransformer	0.1112	0.1281	0.1956	0.2087	0.2256	**0.2331**
PatchTST	0.1098	0.1229	0.1961	0.2100	0.2282	**0.2346**
Crossformer	0.1049	0.1176	0.1949	0.2107	0.2302	**0.2344**
TimeMixer	0.1028	0.1188	0.1961	0.2108	0.2312	**0.2345**
DLinear	0.1061	0.1180	0.1975	0.2106	0.2284	**0.2313**
Chronos†	0.1087	0.1164	0.1941	0.2069	0.2247	**0.2312**
Moirai†	0.1015	0.1117	0.1931	0.2073	0.2228	**0.2277**

Table 18: Decision-time and forecast-horizon sensitivity on MIMIC-IV for the cohort with ICU stays of at least 48 hours. Forecasting variants use iTransformer and TRACE uses k=30. All variants at each time use the same observed first T hours; H=48 − T hours are forecasts, not additional observations.

T/H (hours)	Variant	AUROC	AUPRC	F1
24/24	Direct Classification	0.746	0.367	0.414
	Solo	0.753	0.369	0.417
	Quantile-only	0.783	0.387	0.433
	TRACE	**0.795**	**0.407**	**0.451**
30/18	Direct Classification	0.748	0.368	0.416
	Solo	0.755	0.374	0.422
	Quantile-only	0.786	0.394	0.440
	TRACE	**0.798**	**0.407**	**0.452**
36/12	Direct Classification	0.742	0.365	0.408
	Solo	0.768	0.379	0.444
	Quantile-only	0.788	0.394	0.434
	TRACE	**0.846**	**0.517**	**0.522**
42/6	Direct Classification	0.752	0.389	0.435
	Solo	0.759	0.399	0.444
	Quantile-only	0.797	0.406	0.452
	TRACE	**0.811**	**0.432**	**0.474**

Table 19: Decision-time and forecast-horizon sensitivity on eICU-CRD for the cohort with ICU stays of at least 48 hours. Forecasting variants use iTransformer and TRACE uses k=30. All variants at each time use the same observed first T hours; H=48 − T hours are forecasts, not additional observations.

T/H (hours)	Variant	AUROC	AUPRC	F1
24/24	Direct Classification	0.711	0.109	0.235
	Solo	0.727	0.124	0.247
	Quantile-only	0.739	0.143	0.256
	TRACE	**0.751**	**0.169**	**0.263**
30/18	Direct Classification	0.709	0.107	0.233
	Solo	0.723	0.120	0.244
	Quantile-only	0.741	0.148	0.257
	TRACE	**0.759**	**0.186**	**0.269**
36/12	Direct Classification	0.713	0.111	0.237
	Solo	0.732	0.128	0.251
	Quantile-only	0.753	0.217	0.245
	TRACE	**0.781**	**0.233**	**0.283**
42/6	Direct Classification	0.719	0.116	0.241
	Solo	0.734	0.133	0.252
	Quantile-only	0.760	0.188	0.270
	TRACE	**0.773**	**0.216**	**0.278**

Table 20: Mortality-probability metrics at $T/H{=}36/12$ hours (lower is better). Forecast-based variants: mean $\pm$ standard deviation across seven forecasting-backbone configurations; TRACE uses $k{=}30$. Direct Classification: one fitted observed-history control. ECE is a bin-based calibration diagnostic.

Variant	Brier score ↓	Log-loss ↓	ECE ↓
MIMIC-IV			
Direct Classification	0.1709	0.5061	0.2195
Solo	0.1925 ± 0.015	0.5530 ± 0.036	0.2541 ± 0.030
Quantile-only	0.1730 ± 0.023	0.4956 ± 0.057	0.2361 ± 0.045
TRACE	$\mathbf{0.1483 \pm 0.013}$	$\mathbf{0.4473 \pm 0.034}$	$\mathbf{0.1885 \pm 0.032}$
eICU-CRD			
Direct Classification	0.2218	0.6284	0.3698
Solo	0.1863 ± 0.028	0.5686 ± 0.043	0.3356 ± 0.032
Quantile-only	0.1931 ± 0.009	0.5305 ± 0.022	0.3039 ± 0.015
TRACE	$\mathbf{0.1201 \pm 0.004}$	$\mathbf{0.3920 \pm 0.011}$	$\mathbf{0.2284 \pm 0.010}$

E.7 Mortality-Risk Probability Quality

The table below reports Brier score, log-loss, and ECE for the $T/H{=}36/12$ test sets. Forecast-based variants summarize seven forecasting-backbone configurations (mean $\pm$ standard deviation), whereas Direct Classification is one separately fitted control.

Table 20 reports the final probability-quality metrics.

E.8 A3 Ablation: Quantile Forecasting without Retrieval

Tables 21 and 22 report A3 results (quantile-derived input features without retrieval, $\lambda{=}0$) for all seven backbones on MIMIC-IV and eICU-CRD. A1 (no forecasting), A2 (Solo), and TRACE ($k{=}30$) results are reported in Tables 1 and 3 of the main text. These results support matched comparisons of observed-history, point-forecast, quantile-based, and retrieval-augmented configurations.

Table 21: A3 ablation results on MIMIC-IV test set: quantile-derived inputs without retrieval ($\lambda{=}0$). † zero-shot pretrained models.

Backbone	AUROC	AUPRC	F1
iTransformer	0.7882	0.3941	0.4336
PatchTST	0.7892	0.3929	0.4292
Crossformer	0.7896	0.4017	0.4410
TimeMixer	0.7909	0.3939	0.4343
DLinear	0.7873	0.3918	0.4291
Chronos†	0.7929	0.3959	0.4322
Moirai†	0.7918	0.4042	0.4268

Table 22: A3 ablation results on eICU-CRD test set. Same notation as Table 21.

Backbone	AUROC	AUPRC	F1
iTransformer	0.7527	0.2167	0.2452
PatchTST	0.7544	0.2155	0.2471
Crossformer	0.7505	0.2151	0.2473
TimeMixer	0.7510	0.2184	0.2530
DLinear	0.7532	0.2159	0.2528
Chronos†	0.7514	0.2104	0.2469
Moirai†	0.7468	0.2063	0.2510

E.9 Interpolation Hyperparameter Sensitivity

Table 23 reports AUROC on the MIMIC-IV validation set across a grid of $\lambda_{\max}$ and τ values, using iTransformer as the backbone at k=30. The reported default setting (τ=0.05, $\lambda_{\max}$=0.4; val AUROC= 0.837) was evaluated outside the displayed grid and matches the largest displayed grid value. This table describes sensitivity on the reported validation split.

Table 23: Val AUROC sensitivity to $\lambda_{\max}$ and τ (MIMIC-IV validation set, iTransformer, k=30, adaptive mode). Our default (τ=0.05, $\lambda_{\max}$=0.4) lies outside this grid and achieves val AUROC= 0.837, matching the grid optimum.

	$\lambda_{\max}$			
τ	0.3	0.5	0.7	0.9
0.1	0.836	**0.837**	0.834	0.823
0.3	0.829	**0.835**	0.832	0.816
0.5	0.821	**0.831**	0.829	0.816

E.10 Interpolation Mode Comparison

We evaluate three strategies for setting λ_i in $P(y_i{=}1) = (1-\lambda_i)\, p_{\text{model}} + \lambda_i\, p_{\text{kNN}}$:

- **Fixed**: $\lambda_i = \lambda$ (constant), tuned on the validation set (λ=0.3).
- **Adaptive** (used in TRACE): $\lambda_i = \lambda_{\max} \cdot (1 - \mathcal{H}_{\text{b}}(p_{\text{kNN}}(i)))$, where $\mathcal{H}_{\text{b}}$ is normalized binary entropy and $\lambda_{\max}$=0.4. When retrieved neighbors are label-homogeneous (low entropy), λ_i is large and retrieval dominates; when neighbors are mixed (high entropy), the model pathway dominates.
- **Learned**: $\lambda_i = \sigma(\text{MLP}_\lambda(\mathbf{f}_i)) \cdot \lambda_{\max}$, where MLP_λ is a two-layer network ($117 \to 64 \to 1$, hidden dim 64) trained jointly with the classifier.

We use the entropy-dependent rule as a parameter-free weighting heuristic. In the evaluated MIMIC-IV setting, it yields a higher AUROC and AUPRC than the fixed and learned rules (Table 24).

Table 24: AUROC, AUPRC, and F1 comparison of three interpolation strategies on MIMIC-IV test set (iTransformer, k=30). Adaptive achieves the best performance across all metrics without additional parameters.

Strategy	AUROC	AUPRC	F1
Fixed (λ=0.3)	0.837	0.506	0.511
Adaptive ($\lambda_{\max}$=0.4)	**0.846**	**0.517**	**0.522**
Learned ($\lambda_{\max}$=0.4)	0.833	0.507	0.513

F Post-hoc Forecast-Interval Coverage

F.1 Definition and Comparison

For patient i, the Coverage Score is the fraction of *observed* future measurements of the three target vital signs contained in the predicted 0.1–0.9 quantile interval:

$$\text{CS}_i = \frac{1}{|\mathcal{T}_i|} \sum_{(t,k)\in\mathcal{T}_i} \mathbf{1}[\hat{Y}_i^{(0.1)}[t,k] \le Y_i[t,k] \le \hat{Y}_i^{(0.9)}[t,k]], \tag{32}$$

where $\mathcal{T}_i$ contains only the observed future hours and variables.

For the exploratory deceased-versus-surviving comparison, we report a two-sided Mann–Whitney p-value and the rank-biserial effect size $r = 1 - 2U_{\text{dead}}/(n_{\text{dead}} n_{\text{alive}})$. This sign convention gives positive r when surviving patients tend to have higher coverage. The reported p-values are not corrected for multiple backbone comparisons.

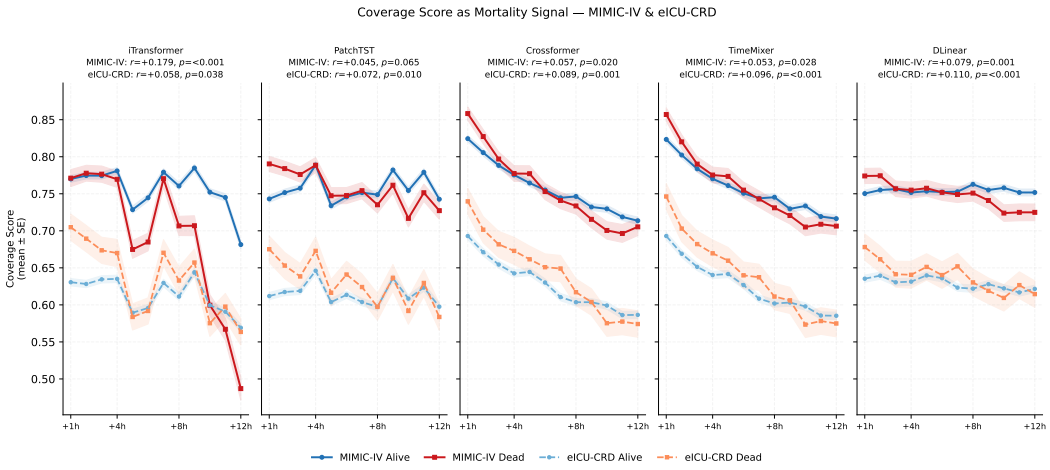


Figure 4: Post-hoc Coverage Score stratified by mortality outcome for five trained forecasting backbones on MIMIC-IV and eICU-CRD. Each panel shows interval coverage of observed future vital signs over the 12-hour forecast horizon. Panel titles give the overall rank-biserial effect size and two-sided Mann–Whitney p-value.

F.2 Outcome-Stratified Coverage on MIMIC-IV

Table 25 provides the results underlying the MIMIC-IV portion of Figure 4. The comparison characterizes forecast intervals conditional on eventual mortality.

Table 25: Coverage Score by mortality outcome for all seven backbones on the MIMIC-IV test set. All p-values from two-sided Mann-Whitney U test. † zero-shot pretrained models.

Backbone	μ_{dead}	μ_{alive}	p-value	Effect size r
iTransformer	0.743	0.797	2.25×10^{-13} ***	+0.179
PatchTST	0.773	0.789	6.48×10^{-2} (ns)	+0.045
Crossformer	0.774	0.791	1.99×10^{-2} *	+0.057
TimeMixer	0.775	0.790	2.79×10^{-2} *	+0.053
DLinear	0.765	0.788	1.13×10^{-3} **	+0.079
Chronos†	0.761	0.769	6.11×10^{-1} (ns)	+0.012
Moirai†	0.456	0.434	2.70×10^{-3} **	−0.073

F.3 Outcome-Stratified Coverage on eICU-CRD

Table 26 reports the corresponding analysis for the separately trained eICU-CRD models.

Table 26: Coverage Score by mortality outcome for all seven backbones on the eICU-CRD test set. † zero-shot pretrained models.

Backbone	μ_{dead}	μ_{alive}	p-value	Effect size r
iTransformer	0.743	0.765	3.76×10^{-2} *	+0.058
PatchTST	0.739	0.764	1.04×10^{-2} *	+0.072
Crossformer	0.768	0.795	1.47×10^{-3} **	+0.089
TimeMixer	0.765	0.793	6.28×10^{-4} ***	+0.096
DLinear	0.760	0.796	8.10×10^{-5} ***	+0.110
Chronos†	0.724	0.726	7.25×10^{-1} (ns)	−0.010
Moirai†	0.436	0.428	1.88×10^{-1} (ns)	−0.037

G Forecasting Objective and Quantile Diagnostics

G.1 Forecasting MAE under Training Objectives

We compare the forecasting MAE of each backbone under two training objectives on the MIMIC-IV validation set:

- **Standard MSE** (non-modified): single-path MSE loss, the default training objective used in the original backbone implementations.
- **Composite loss** (ours): the quantile pinball loss $\mathcal{L}_Q$ combined with trend loss $\mathcal{L}_T$ and deterioration-aware loss $\mathcal{L}_D$ (Appendix D.3, λ_T=1.0, λ_D=0.1).

Table 27 reports the validation MAE for HR, RespRate, and SpO2 under both objectives. The composite loss consistently reduces MAE across all five fine-tuned backbones and all three target variables, with the largest improvement observed for Crossformer on HR (Δ=-1.253 bpm). The observed MAE differences compare the complete training objectives.

Table 27: Validation MAE (HR: bpm, RespRate: breaths/min, SpO2: %) under standard MSE training vs. our composite loss ($\mathcal{L}_Q + \mathcal{L}_T + \mathcal{L}_D$) on the MIMIC-IV validation set. Δ = composite $-$ standard; negative values indicate improvement. RR = RespRate.

	Standard MSE			Composite (ours)			Δ		
Backbone	HR	RR	SpO2	HR	RR	SpO2	HR	RR	SpO2
iTransformer	5.44	3.14	1.56	5.09	3.06	1.52	-0.35	-0.08	-0.03
PatchTST	5.44	3.14	1.55	5.13	3.08	1.54	-0.31	-0.06	-0.01
Crossformer	6.67	3.32	1.59	5.42	3.17	1.58	-1.25	-0.15	-0.01
TimeMixer	5.22	3.13	1.54	5.03	3.07	1.53	-0.19	-0.07	-0.02
DLinear	5.56	3.14	1.56	5.19	3.08	1.53	-0.37	-0.07	-0.02

G.2 Quantile Separation and Pooled Interval Coverage

Table 28: Per-backbone quantile forecast diagnostics on MIMIC-IV and eICU-CRD test sets. *Quantile separation*: mean pairwise MAD across the $\binom{5}{2}$=10 quantile-level prediction pairs (normalized space). *Nominal 80% coverage*: fraction of observed future values within $[q_{0.1}, q_{0.9}]$ at mask=1 timesteps (target: 0.800). Results are reported for the five fine-tuned backbones.

	Quantile Separation (MAD)		Nominal 80% Coverage	
Backbone	MIMIC-IV	eICU-CRD	MIMIC-IV	eICU-CRD
iTransformer	0.239	0.222	0.791	0.775
PatchTST	0.246	0.220	0.788	0.778
Crossformer	0.247	0.235	0.789	0.793
TimeMixer	0.247	0.235	0.788	0.791
DLinear	0.248	0.239	0.784	0.793

Table 28 reports two quantile forecast quality metrics for all five trainable backbones on MIMIC-IV and eICU-CRD.

Quantile separation is the mean pairwise MAD across the $\binom{5}{2}$=10 pairs of quantile-level predictions, averaged over observed future hours, variables, and patients. These differences measure spread between predicted marginal quantiles; they do not represent diversity among independently sampled patient futures.

Nominal 80% coverage is the fraction of observed future vital sign values that fall within the $[q_{0.1}, q_{0.9}]$ interval, computed over all valid (mask=1) timesteps. The nominal target is 0.8.