From Forecasting to Features: Zero-Shot Forecasting for Feature Extraction

Nassim Oufattole NASSIM@MIT.EDU
EECS
MIT
Cambridge, MA, USA

Matthew McDermott MATTMCDERMOTT8@GMAIL.COM
DBMI
Columbia University
New York City, New York, USA

Collin Stultz CMSTULTZ@MIT.EDU
EECS
MIT
Cambridge, MA, USA

Abstract

Generative EHR models can estimate clinical risk in a zero-shot fashion by sampling future trajectories and computing outcome probabilities from generated rollouts. Yet in practice, direct use of these zero-shot risk estimates often underperforms supervised tabular baselines built from summaries of the observed past. In this work, we ask a narrower question: can future trajectories sampled from a generative foundation model be used to construct features that enhance the performance of downstream predictors? To test this, we convert model rollouts into horizon-specific *generated future features* (GFFs) that summarize, for each patient, the probability of all future clinical events under the model, and train a simple XGBoost classifier that uses these features for downstream classification tasks. Across post-discharge prediction tasks in MIMIC-IV and a large private heart failure cohort, GFFs consistently achieve higher AUROC than direct zero-shot prediction and supervised learning on features derived from patient history. In our experiments we observe that stronger zero-shot forecasting performance tends to yield more performant GFFs. These results suggest that the value of current zero-shot generative EHR models is not only in their ability to generate future clinical trajectories, but also in their role as probabilistic feature generators that can be used for simple supervised risk models.

1. Introduction

Clinical risk prediction using electronic health records (EHRs) asks a basic but difficult question: given a patient's longitudinal history, what is likely to happen next? Of particular interest is the likelihood of specific adverse future events; e.g., death, heart attack, stroke, etc. This question sits at the center of many healthcare applications, including post-discharge risk stratification, early detection of deterioration, and anticipating future resource utilization. Although modern deep learning models can operate directly on irregular longitudinal records to predict specific clinical outcomes, they often underperform

when compared to simple supervised models using carefully designed summaries of the observed past. This is particularly evident in recent EHR studies using MEDS-Tab-style representations, which achieve competitive and in some cases foundation-model-matching or foundation-model-exceeding performance across multiple tasks and datasets (Oufattole et al., 2024; Pang et al., 2025; Kolo et al., 2024). More broadly, this fits a wider pattern in machine learning: for tabular prediction problems, tree-based methods such as XGBoost (Chen and Guestrin, 2016) remain remarkably strong baselines (Grinsztajn et al.; Shwartz-Ziv and Armon, 2021).

At the same time, recent generative EHR models have introduced a qualitatively different capability. Rather than producing a prediction for one pre-specified task, EHR generative models simulate plausible future patient trajectories and support *zero-shot* forecasting across many endpoints by estimating event probabilities from sampled rollouts (McDermott; Renc et al., 2024, 2025; Yang et al., 2023; Redekop et al., 2025; McDermott et al.; Waxler et al., 2025). This makes them attractive as flexible models of patient state: once pretrained, a single generative model can in principle be queried for many downstream questions without retraining a separate predictor for each one.

In sum, strong tabular models are highly predictive, but they depend on manually designed summaries of the observed past. Generative models are more flexible and expressive, but their direct zero-shot risk estimates are often weaker than supervised tabular baselines. This creates a natural question: can the relative benefits of both approaches be combined; i.e., supervised learning with tabular features for a specific task and zero-shot forecasting across many different endpoints.

In this paper, we study exactly that question. Rather than treating a generative forecaster as a standalone predictor, we treat it as a *probabilistic feature generator*. For each patient and prediction horizon, we sample future trajectories from a frozen generative model and estimate, for every event in a fixed vocabulary, the probability that the event occurs by the horizon. This produces a dense, horizon-specific generated-future feature (GFF) vector[1] that summarizes the model-implied future state of the patient (see Figure 1). We then train a simple XGBoost classifier on top of these features.

Across post-discharge prediction tasks in MIMIC-IV and a large private heart failure cohort, generated-future features consistently achieve higher AUROC than direct zero-shot prediction and a strong tabular baseline fitted on patient history in all experiments.

In our experimental setting, we find higher zero-shot performance tends to yield stronger downstream generated-future features, supporting the interpretation that improvements in generative forecasting quality can translate into better downstream feature quality.

Summary of contributions.

- We frame zero-shot forecasting with a generative EHR model as *forecast-based feature generators* for downstream supervised prediction, rather than only as direct predictors.
- We introduce a simple generated-future feature pipeline that converts Monte Carlo rollout forecasts into horizon-specific tabular feature vectors and trains XGBoost on top.

1. Code and ACES task definitions are available at https://github.com/Oufattole/gff.

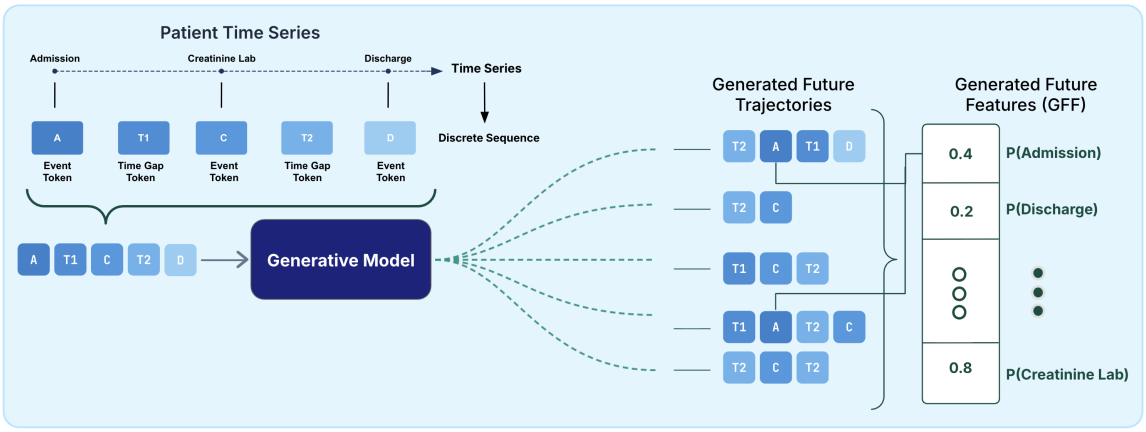


Figure 1: **Overview of generated-future features (GFF).** A patient's discretized pre-discharge EHR sequence is provided to a frozen generative model, which samples multiple possible post-discharge trajectories. For each code in the evaluation vocabulary and prediction horizon, we compute the fraction of sampled futures in which that code occurs by the horizon, yielding a horizon-specific probability vector. This generated-future feature vector is then used as input to a downstream XGBoost classifier for supervised risk prediction.

- We show across two EHR cohorts, multiple clinical tasks, and short- and long-horizon settings that GFF consistently achieves higher AUROC than both direct zero-shot prediction and a strong past-window tabular baseline.
- We provide evidence that under experimental settings similar to ours, stronger zero-shot performance tends to produce stronger downstream generated-future features.

Generalizable Insights about Machine Learning in the Context of Healthcare

These results suggest that the practical value of foundation-style generative EHR models may often lie less in direct zero-shot deployment and more in the representations they provide to simpler supervised learners. More broadly, when new generative models do not outperform strong tabular baselines in direct comparisons, they may still be valuable if recast as probabilistic feature generators for downstream predictors.

2. Related Work

Tabular models remain strong baselines for EHR prediction. Across a broad range of prediction problems in the healthcare space, supervised tabular models built on carefully engineered features remain highly competitive (Shwartz-Ziv and Armon, 2021; Grinsztajn et al.). In EHR modeling, MEDS-Tab (Oufattole et al., 2024) has been shown to achieve competitive discriminative performance relative to several EHR foundation models while remaining simple to train and robust across tasks (Pang et al., 2025; Kolo et al., 2024). MEDS-Tab converts patient histories into tabular feature vectors that can be used

to train supervised models such as XGBoost (Chen and Guestrin, 2016). In this paper, we compare GFF against history-based tabularization using MEDS-Tab. For clarity, we refer to the MEDS-Tab-based baseline as XGB-Past, since it trains XGBoost on features derived directly from the observed patient history, such as counts of prior admissions within a fixed look-back window before prediction time.

Generative EHR models enable zero-shot forecasting across diverse tasks. Recent autoregressive and sequence-based generative models for EHR data have made it possible to estimate clinical risk in a zero-shot manner by sampling future trajectories conditioned on a patient's history and computing endpoint probabilities from those rollouts (McDermott; Renc et al., 2024, 2025; Redekop et al., 2025; Waxler et al., 2025). These results demonstrate that generative EHR models can provide clinically meaningful discriminative performance across diverse forecasting tasks without task-specific supervised training. However, while these methods generate rich future trajectories, they are typically evaluated only through zero-shot endpoint probabilities, collapsing each rollout set to the probability that a particular event occurs. This leaves potentially useful trajectory-level structure underutilized for downstream prediction, and the resulting zero-shot risks are not always competitive with strong supervised discriminative baselines.

This work combines generative forecasting with tabular prediction. The strengths of these two lines of work are complementary: tabular models provide strong supervised discrimination from structured feature vectors, while generative EHR models support zero-shot forecasting by simulating possible patient futures. We combine these approaches by using sampled future trajectories not as endpoints in themselves, but as a source of generated-future features for a downstream tabular learner. This allows us to test whether predictive structure in zero-shot generative forecasts can be more effectively exploited for supervised prediction.

3. Methods

Our main experimental goal is to test whether zero-shot forecasting using a generative EHR model is most useful as a standalone predictor or as a generator of downstream features for supervised tabular learning. We therefore compare five approaches that use patient history at discharge: (i) direct zero-shot risk estimates from generated future trajectories, (ii) supervised learning with a transformer baseline trained on patient history; (iii) supervised learning with tabular summaries of the observed past; (iv) our proposed generated-future features, which summarize the model-implied future and are consumed by XGBoost for supervised learning; (v) linear probing, where we fit a supervised logistic regression model on generative model embeddings. Figure 11 provides graphical overviews of the four supervised and representation-based approaches (ii-v). Figure 1 illustrates the shared rollout procedure: XGB-GFF uses the resulting probabilities for all vocabulary codes as a feature vector, whereas Zero-Shot predictions (i) are computed as the fraction of trajectories containing the target endpoint.

3.1. Prediction setting

Let h_j denote the observed history of patient j up to an index time, defined here as the time at hospital discharge. For task k and prediction horizon H, let

$$Y_{j,H}^{(k)} \in \{0, 1\}$$

indicate whether the outcome occurs at least once in $(0, H]$ after discharge. Our objective is to estimate the patient-specific horizon risk

$$\pi_{j,H}^{(k)} = \mathrm{P}\big(Y_{j,H}^{(k)} = 1 \mid h_j\big).$$

We evaluate post-discharge prediction at horizons $H \in \{30, 730\}$ days for readmission, death, and abnormal laboratory outcomes for creatinine, hematocrit, hemoglobin, leukocyte, and platelet.

All methods begin from the same discretized MEDS-style event sequence representation. Non-numeric events are represented directly as discrete tokens, numeric laboratory values are discretized into value-bin tokens, and time gaps are represented using explicit delta-time tokens. Using a shared sequence representation ensures that differences between methods arise from how the same patient history is summarized and used downstream, rather than from differences in raw-data preprocessing.

3.2. Baselines and comparisons

We compare five different approaches:

3.2.1. Zero-shot prediction with frozen ETHOS generative model

Each cohort uses one frozen, institution-specific ETHOS architecture autoregressive transformer decoder (Renc et al., 2024) with 6 layers, 12 attention heads, hidden size 768, and context length 512. Training was configured for 200 epochs at learning rate 6×10^{-4}, with checkpoint selection by validation performance. The generative model defines a predictive distribution over future trajectories conditioned on the observed history h_j. Given a task k, the model implies a horizon probability

$$\hat{\pi}_{j,H}^{(k)}(\theta) = \mathrm{P}_\theta\big(Y_{j,H}^{(k)} = 1 \mid h_j\big),$$

where θ denotes the pretrained generative model parameters.

Because this quantity is not available in closed form, we estimate it via Monte Carlo rollout. For each patient, we sample S future trajectories and compute the fraction in which the endpoint occurs by horizon H:

$$\hat{\pi}_{j,H}^{(k)} = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\{\text{outcome } k \text{ occurs within } H \text{ in rollout } s\}.$$

Throughout the experiments we use $S = 50$ sampled futures per patient. When this scalar estimate is used directly as the predictor, we refer to the method as **Zero-Shot**.

3.2.2. Supervised Learning with a task-specific transformer model

Our first baseline represents the standard approach that directly uses the patient history h_j:

$$h_j \longrightarrow \text{Transformer} \longrightarrow \hat{Y}^{(k)}_{j,H}.$$

3.2.3. Supervised Learning with tabular patient features

Our second baseline represents the standard strong tabular approach. We summarize the observed pre-discharge history using rolling lookback windows and construct a fixed-dimensional feature vector from counts of observed codes within each window. Concretely, we use MEDS-Tab count features over 1-day, 7-day, 30-day, 365-day, and full-history windows. These features are passed to a classifier:

$$x^{\text{past}}_j \longrightarrow \text{XGBoost} \longrightarrow \hat{Y}^{(k)}_{j,H}.$$

These baselines capture an important practical fact in EHR prediction: simple supervised models built on carefully designed summaries of the observed past are often very strong.

3.2.4. Supervised learning with XGBoost and GFFs

Our method asks whether the generative model's usefulness extends beyond its direct zero-shot scalar prediction. Instead of querying the model only for the probability of one target endpoint, we summarize the model-implied future more broadly.

For each patient j, horizon H, and code c in a vocabulary $\mathcal{V}$, we estimate the probability that c occurs at least once by horizon H across generated futures:

$$\hat{p}_{j,H}(c) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{\text{code } c \text{ occurs within } H \text{ in rollout } s\}.$$

We then stack these probabilities into a generated-future feature vector

$$x^{\text{GFF}}_{j,H} = \big(\hat{p}_{j,H}(c)\big)_{c\in\mathcal{V}} \in [0,1]^{|\mathcal{V}|}.$$

This vector summarizes the model-implied future state of the patient: not only the risk of the target endpoint, but also the model's beliefs about other downstream events, abnormalities, and utilization patterns that may be predictive of that endpoint. We then train a simple XGBoost classifier on top of these features:

$$x^{\text{GFF}}_{j,H} \longrightarrow \text{XGBoost} \longrightarrow \hat{Y}^{(k)}_{j,H}.$$

We refer to the features as **Generated-Future Features (GFF)** and to the method of fitting XGBoost on these features as **XGB-GFF**.

3.2.5. Linear Probing on token embeddings from an EHR generative model

To help isolate what drives any downstream gains, we additionally compare GFF against linear probing on learned token embeddings from the frozen generative EHR model. This comparison tests whether the advantage comes specifically from summarizing the *forecasted future*, rather than simply from reusing a pretrained neural representation in some generic way. More details on the setup are in Appendix A.

3.3. Training and evaluation

All tabular classifiers use XGBoost with the same tuning protocol across feature sets. Hyperparameters are tuned separately for each dataset, task, horizon, and feature group by randomized search over 250 configurations on the train/tuning split, selecting the best tuning AUROC and then refitting with five random seeds. The supervised transformer is also trained separately for each dataset–task–horizon configuration. See Appendix A for more details.

We report AUROC as mean ± standard deviation. Boldface in tables denotes statistically significant improvements at $\alpha = 0.05$ via a one-sided t-test. Additional details on uncertainty estimation are provided in Appendix A.

3.4. Questions tested by the experiments

The experiments are designed to answer six focused questions:

- Q1. Does converting generated futures into feature vectors outperform direct zero-shot prediction?
- Q2. Are direct zero-shot risks from generative models competitive with strong tabular baselines?
- Q3. Do generated-future features rival or exceed strong supervised baselines in this setting?
- Q4. Are the gains explained by generic reuse of the pretrained representation, without forecasting?
- Q5. Are these features simply proxies for the target endpoint, or do they have value beyond that?
- Q6. Do stronger zero-shot forecasters tend to produce stronger downstream generated-future features?

Together, these comparisons test whether the value of current generative EHR forecasters lies more in direct deployment as predictors or in their use as probabilistic feature generators for downstream supervised learning. Note, we categorize the results subsections by which of these research questions they address.

Table 1: **Cohort summary.** Summary statistics after cohort construction and discharge anchoring.

Characteristic	Hospital A heart failure cohort	MIMIC-IV discharge cohort
Patients	179,227	123,059
Discharges	1,531,893	397,737
Median Age at discharge (Years)	69.8	63.0

4. Cohort

We evaluate post-discharge prediction in two EHR cohorts: a private heart failure cohort from a large tertiary health system assembled into the MEDS format (Hospital A) and the public MIMIC-IV (Johnson et al., 2024) discharge cohort. The Hospital A cohort focuses on patients with heart failure and provides a long-horizon evaluation setting with repeated hospital discharges per patient. The MIMIC-IV cohort is a general discharge cohort constructed from the public MEDS-MIMIC-IV ETL (McDermott and Xu, 2025). For both cohorts, each prediction time is anchored at hospital discharge, and outcomes are evaluated over fixed post-discharge horizons. Both datasets are split by patient into 95% train, 2.5% tuning, and 2.5% held-out splits. See additional summary statistics for the datasets in Table 1.

4.1. Cohort Selection

Each prediction example corresponds to a hospital discharge, with the discharge time serving as the prediction index. We include only discharges from patients with at least 10 events remaining after vocabulary filtering. Additional cohort characteristics are reported in Appendix C, and split sizes and comparisons between full-cohort and held-out outcome prevalences are provided in Appendix F.

4.2. Data Extraction

MIMIC-IV (Johnson et al., 2024) is processed using the MEDS-MIMIC-IV ETL (McDermott and Xu, 2025). The Hospital A heart failure cohort is converted into the MEDS event format using a reproducible private-data extraction pipeline. In both cohorts, raw longitudinal records are represented as timestamped MEDS events and then processed using a common MEDS-Transforms pipeline. The retained event vocabulary includes death, birth or registration markers when available, hospital admission and discharge events, time-derived tokens, and the laboratory measurements needed for the evaluated abnormal-lab endpoints. Numeric-valued observations are discretized before modeling, as described below and in Appendix B.

4.3. Outcomes

We evaluate post-discharge binary prediction tasks at horizons $H \in \{30, 730\}$ days. The tasks are readmission, death, and abnormal laboratory value detection for leukocyte, platelet, hemoglobin, hematocrit, and creatinine. Creatinine corresponds to elevated creatinine > 2.3

mg/dL. Hematocrit, hemoglobin, leukocyte, and platelet correspond to decreased values $< 24.6\%$, < 8.0 g/dL, < 5.1 K/uL, and < 86.0 K/uL, respectively.

For each discharge-indexed example and task, the label is positive if the corresponding event occurs at least once within $(0, H]$ after discharge. For abnormal laboratory tasks, the label is defined on the discretized laboratory representation: an example is positive if an observed laboratory value, represented by its discretized laboratory token, crosses the task-specific threshold within the prediction horizon.

4.4. Feature Choices

All models operate on the same discretized sequence representation. Non-numeric events are represented directly as discrete tokens. Numeric-valued measurements are discretized into within-code value bins. Time gaps are represented using time-derived tokens. This yields the event vocabulary used by both the cohort-specific ETHOS models and the generated-future feature construction. MEDS data is processed into this format using the MEDS-EIC-AR library (McDermott).

Past features use MEDS-Tab count summaries over 1-day, 7-day, 30-day, 365-day, and full-history lookback windows. Generated-future features convert frozen generative-model rollouts into horizon-specific event-probability features using 50 sampled futures per patient. Full implementation details and additional preprocessing choices are provided in Appendix A.

5. Results on Real Data

Our experiments test a simple hypothesis: current generative EHR forecasters may be more useful as probabilistic feature generators than as standalone predictors. Across datasets, tasks, and horizons, we find consistent evidence for this view. Direct zero-shot risks often underperform strong past-window tabular baselines, but the same generative rollouts become substantially more useful when converted into generated-future feature vectors and consumed by XGBoost. These forecast-derived features not only outperform direct zero-shot use, but also outperform both strong past-window tabular baselines and a supervised transformer baseline in all experiments, suggesting that the main practical value of these models lies in the quality of the future-state representations they induce.

5.1. XGB-GFF outperforms both zero-shot and past-window baselines–(Q1, Q3)

In Table 2, GFF-based predictions outperform zero-shot performance for all 28 settings. The gains were visible in both hospital systems and at both horizons, indicating that the effect is not confined to a single dataset or prediction regime. Because XGB-GFF uses labeled downstream supervision while zero-shot does not, this improvement in performance is not unexpected. Nevertheless, this performance still supports our central claim: the task-relevant information contained in sampled future trajectories is not fully captured by a single zero-shot endpoint probability.

The more consequential comparison is against XGB-Past, since both methods use the same supervised XGBoost learner and differ primarily in the features they receive. XGB-

Table 2: **Comparison of methods:** Zero-Shot, XGBoost fitted directly on patient history features (XGB-Past), Supervised Transformer, and XGBoost fitted on generated-future features (XGB-GFF). AUROC (%) reported as mean ± standard deviation. Bold indicates a statistically significant improvement at $\alpha = 0.05$ under a one-sided t-test.

				Supervised		
Outcome	Horizon	Dataset	Zero-Shot	Transformer	XGB-Past	XGB-GFF (Ours)
Creatinine	30	Hospital A	95.01 ± 0.13	95.33 ± 0.10	95.02 ± 0.19	**96.07 ± 0.02**
Death	30	Hospital A	80.70 ± 0.47	78.94 ± 0.51	82.19 ± 0.05	**83.50 ± 0.04**
Hematocrit	30	Hospital A	91.87 ± 0.06	91.38 ± 0.34	91.28 ± 0.32	**92.43 ± 0.00**
Hemoglobin	30	Hospital A	92.05 ± 0.06	91.68 ± 0.14	91.82 ± 0.01	**92.63 ± 0.01**
Leukocyte	30	Hospital A	88.16 ± 0.10	88.28 ± 0.49	87.51 ± 0.01	88.67 ± 0.01
Platelet	30	Hospital A	93.74 ± 0.12	93.66 ± 0.12	94.22 ± 0.02	**94.91 ± 0.00**
Readmission	30	Hospital A	86.53 ± 0.06	85.02 ± 0.32	85.77 ± 0.01	**86.81 ± 0.01**
Creatinine	730	Hospital A	90.37 ± 0.11	90.60 ± 0.20	90.73 ± 0.41	**91.60 ± 0.02**
Death	730	Hospital A	74.19 ± 0.35	74.25 ± 1.07	77.64 ± 0.08	**78.37 ± 0.02**
Hematocrit	730	Hospital A	83.71 ± 0.06	82.69 ± 0.52	83.87 ± 0.07	**84.36 ± 0.01**
Hemoglobin	730	Hospital A	84.45 ± 0.07	83.57 ± 0.57	84.34 ± 0.44	**85.12 ± 0.01**
Leukocyte	730	Hospital A	82.55 ± 0.14	81.86 ± 0.29	82.48 ± 0.04	**83.47 ± 0.02**
Platelet	730	Hospital A	86.17 ± 0.12	84.99 ± 0.23	85.80 ± 0.07	**87.25 ± 0.01**
Readmission	730	Hospital A	78.28 ± 0.10	74.21 ± 1.17	78.30 ± 0.02	**78.79 ± 0.01**
Creatinine	30	MIMIC	91.18 ± 0.44	93.22 ± 0.68	93.51 ± 0.28	**94.14 ± 0.07**
Death	30	MIMIC	83.08 ± 1.63	87.20 ± 0.16	84.04 ± 1.03	**88.65 ± 0.32**
Hematocrit	30	MIMIC	87.73 ± 0.43	88.55 ± 0.51	88.76 ± 0.12	**89.86 ± 0.00**
Hemoglobin	30	MIMIC	87.06 ± 0.58	88.46 ± 0.06	88.71 ± 0.10	**89.52 ± 0.03**
Leukocyte	30	MIMIC	87.20 ± 0.29	85.64 ± 0.38	85.77 ± 0.12	**87.58 ± 0.07**
Platelet	30	MIMIC	93.63 ± 0.38	93.98 ± 0.50	94.62 ± 0.00	**95.68 ± 0.12**
Readmission	30	MIMIC	71.58 ± 0.25	67.78 ± 0.87	69.69 ± 0.15	**72.26 ± 0.00**
Creatinine	730	MIMIC	89.58 ± 0.25	90.74 ± 0.25	90.92 ± 0.00	**91.68 ± 0.04**
Death	730	MIMIC	78.02 ± 0.44	79.86 ± 0.72	79.59 ± 0.06	**82.54 ± 0.05**
Hematocrit	730	MIMIC	84.37 ± 0.18	83.31 ± 0.13	84.42 ± 0.13	**85.63 ± 0.06**
Hemoglobin	730	MIMIC	84.22 ± 0.21	83.82 ± 0.42	84.74 ± 0.00	**85.83 ± 0.04**
Leukocyte	730	MIMIC	85.29 ± 0.13	84.50 ± 0.58	83.53 ± 0.48	**85.72 ± 0.05**
Platelet	730	MIMIC	85.42 ± 0.57	85.70 ± 0.30	85.89 ± 0.09	**86.91 ± 0.00**
Readmission	730	MIMIC	73.12 ± 0.21	70.17 ± 1.24	70.27 ± 0.08	**74.20 ± 0.09**

Past is trained on tabular summaries of the observed patient history, whereas XGB-GFF is trained on tabular summaries of the model-implied future. The fact that XGB-GFF performs better indicates that sampled future trajectories capture predictive structure not fully represented by tabular featurizations of the observed patient history.

5.2. Zero-shot generative prediction is weaker and often underperforms strong tabular baselines – (Q2)

Comparing the zero-shot and XGB-Past columns in Table 2, we observe XGB-Past outperformed Zero-Shot in 60.7% of experiments. This pattern shows that zero-shot rollout-derived risks are clinically informative, but not consistently competitive with strong supervised tabular baselines. The gap was especially noticeable for several mortality and

Table 3: **Linear Probing (LP) vs. XGBoost on generated-future features (XGB-GFF).** AUROC (%) reported as mean ± standard deviation. Bold indicates a statistically significant improvement at $\alpha = 0.05$ under a one-sided t-test.

	MIMIC				Hospital A			
	30-Day Horizon		2-Year Horizon		30-Day Horizon		2-Year Horizon	
Outcome	LP	XGB-GFF	LP	XGB-GFF	LP	XGB-GFF	LP	XGB-GFF
Creatinine	93.41 ± 0.03	**94.14 ± 0.07**	90.51 ± 0.13	**91.68 ± 0.04**	94.52 ± 0.23	**96.07 ± 0.02**	90.94 ± 0.05	**91.60 ± 0.02**
Death	**89.10 ± 0.19**	88.65 ± 0.32	82.59 ± 0.08	82.54 ± 0.05	80.69 ± 0.39	**83.50 ± 0.04**	**78.49 ± 0.03**	78.37 ± 0.02
Hematocrit	88.44 ± 0.22	**89.86 ± 0.00**	84.57 ± 0.44	**85.63 ± 0.06**	91.35 ± 0.38	**92.43 ± 0.00**	83.94 ± 0.01	**84.36 ± 0.01**
Hemoglobin	88.47 ± 0.20	**89.52 ± 0.03**	84.98 ± 0.63	**85.83 ± 0.04**	91.59 ± 0.49	**92.63 ± 0.01**	84.73 ± 0.00	**85.12 ± 0.01**
Leukocyte	85.11 ± 0.34	**87.58 ± 0.07**	85.00 ± 0.01	**85.72 ± 0.05**	88.16 ± 0.04	**88.67 ± 0.01**	82.90 ± 0.06	**83.47 ± 0.02**
Platelet	94.45 ± 0.13	**95.68 ± 0.12**	85.03 ± 0.27	**86.91 ± 0.00**	92.06 ± 0.65	**94.91 ± 0.00**	86.50 ± 0.13	**87.25 ± 0.01**
Readmission	**72.50 ± 0.07**	72.26 ± 0.00	73.99 ± 0.05	**74.20 ± 0.09**	86.56 ± 0.01	**86.81 ± 0.01**	77.46 ± 0.76	**78.79 ± 0.01**

longer-horizon settings, where a single scalar zero-shot endpoint estimate appears to discard predictive structure that a supervised learner can use when given a broader representation. Thus, while zero-shot forecasting demonstrates that these generative models capture clinically meaningful patient state, direct deployment of their scalar risk estimates does not appear to be the setting in which they deliver the greatest downstream predictive value.

5.3. Supervised learning with GFFs achieves higher AUROC than the supervised transformer baseline across all experiments – (Q3)

Additionally, Table 2 shows XGB-GFF compares favorably against the supervised transformer baseline. Across all 28 dataset–task–horizon settings, XGB-GFF achieved higher AUROC than the transformer. This result is notable because the downstream model used with XGB-GFF is a simple XGBoost classifier rather than a fully end-to-end sequence model. The comparison suggests that much of the task-relevant predictive information can already be captured in the generative model's forecasted future state and then efficiently exploited by a simple supervised learner.

5.4. Comparison to Linear Probing on Token Embeddings – (Q4)

In Table 3 we observe XGB-GFF generally outperforms Linear Probing (LP) on learned token embeddings, winning in 24 of 28 dataset–task–horizon comparisons. The average improvement was 0.93 AUROC points. This comparison is important because it helps motivate the source of the improvement: the benefit is not simply due to reusing a pretrained neural representation. Instead, the gains appear to come from summarizing the model-implied *future* as horizon-specific event probabilities. For the majority of our experiments, forecast-derived features are more useful for downstream prediction than static embedding summaries alone, supporting the view that the generative model's value lies in the structure of its sampled futures rather than merely in its internal hidden states.

Table 4: **Joint multi-horizon prediction.** Each row defines a joint outcome that is positive only if event 1 occurs within 30 days and event 2 occurs within 730 days for the same patient. This setting therefore requires reasoning jointly over multiple clinical signals across different prediction horizons, rather than predicting either event independently. AUROC (%) reported as mean ± standard deviation. Bold indicates the better method within each row.

Dataset	Outcome Pair	Past	GFF
Hospital A	Creatinine → Death	91.47 ± 0.11	**92.70 ± 0.03**
Hospital A	Creatinine → Platelet	91.58 ± 0.02	**93.36 ± 0.03**
Hospital A	Hematocrit → Readmission	91.16 ± 0.03	**92.00 ± 0.02**
Hospital A	Leukocyte → Death	87.98 ± 0.00	**88.73 ± 0.03**
Hospital A	Leukocyte → Hemoglobin	88.24 ± 0.05	**89.23 ± 0.02**
MIMIC	Creatinine → Death	93.00 ± 0.20	**94.05 ± 0.07**
MIMIC	Creatinine → Platelet	91.75 ± 1.55	**94.60 ± 0.48**
MIMIC	Hematocrit → Readmission	88.71 ± 0.00	**89.61 ± 0.11**
MIMIC	Leukocyte → Death	87.66 ± 0.28	**89.77 ± 0.07**
MIMIC	Leukocyte → Hemoglobin	91.14 ± 0.02	**91.86 ± 0.10**

5.5. Joint Multi-Horizon Tasks Comparison – (Q5)

The results in Table 4 demonstrate the advantage of GFF was not limited to marginal single-endpoint prediction. On the joint multi-horizon tasks, GFF outperformed the past-window baseline in all experiments, with an average gain of 1.32 AUROC points. Because these labels require both event 1 and event 2 to occur within different horizons for the same patient, they demand reasoning over multiple event types and timescales simultaneously. The fact that GFF remains strong in this setting suggests that the generated-future feature vectors encode meaningful cross-event structure rather than merely serving as surrogates for individual marginal risks. This strengthens the interpretation that generative forecasts provide a useful higher-order representation of patient state for downstream prediction.

We further probe this question in Appendix H.1 by removing all generated features belonging to the endpoint's code family before tuning and fitting the downstream model (XGB-GFF*). XGB-GFF* still outperforms XGB-Past in 16 of 28 settings, indicating that the gains are not solely driven by generated features directly related to the target endpoint. However, full XGB-GFF outperforms XGB-GFF* in 27 of 28 settings, showing that endpoint-related generated features nevertheless contribute substantial predictive signal.

5.6. Zero-Shot Performance Strongly Correlates with Performance of GFF – (Q6)

In Figure 2 we plot Zero-Shot AUROC against XGB-GFF AUROC for each task. Across dataset–task–horizon settings, zero-shot forecasting AUROC was strongly correlated with downstream GFF performance, with Pearson correlation $r = 0.976$. This relationship suggests that task–horizon settings with better ETHOS zero-shot performance tend to produce

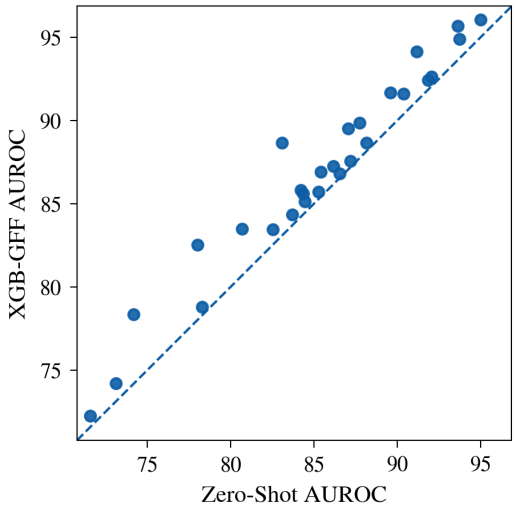


Figure 2: **Relationship between zero-shot forecasting AUROC and downstream (XGB-GFF) AUROC.** Each point corresponds to one dataset–task–horizon setting and plots the AUROC of ETHOS in zero-shot mode against the corresponding AUROC when ETHOS generated-future features (GFF) are fed to XGBoost. A Pearson correlation coefficient of 0.976 was achieved.

better generated-future features for downstream supervised learning. This suggests progress on zero-shot generative EHR forecasting may have a second practical payoff beyond direct prediction: it may improve the quality of GFFs that can be consumed by simple downstream models to achieve even higher performance.

6. Discussion

Our results support a simple interpretation: in this setting, the main practical value of current zero-shot generative EHR models lies less in direct use of rollout-derived scalar risks and more in the richer representations of future patient state that those models make available. When those representations are exposed to a downstream supervised tabular learner, they become substantially more useful for discrimination.

This perspective helps reconcile two observations that might otherwise seem in tension. First, strong tabular baselines remain challenging to consistently beat in EHR prediction. Second, generative EHR models can still encode meaningful information about future patient state even when their direct zero-shot predictions are not the strongest downstream predictors. Our results suggest that these observations are not contradictory: a generative model can be weak as a standalone predictor yet valuable as a probabilistic feature generator.

The comparison to linear probing is important for this interpretation. The downstream gains are not explained merely by reusing a pretrained neural representation. Rather, they appear to depend specifically on summarizing the *forecasted future*. This suggests that the model-implied distribution over possible patient trajectories is carrying predictive

structure that is particularly useful once converted into horizon-specific event probabilities and consumed by a supervised learner.

The correlation between zero-shot forecasting quality and downstream GFF quality points in the same direction. Improvements in the underlying generative forecaster appear to translate into improvements in the downstream features it produces. This is encouraging because it suggests that progress in zero-shot EHR forecasting may additionally drive up performance of models fitted on GFF.

Overall, these results suggest a practical division of labor for current generative EHR models. Rather than asking them to serve only as final zero-shot predictors, it may be more effective to use them to generate forecast-derived representations of likely future patient state and then train simple supervised models on top for the endpoint of interest. Under this view, progress in generative forecasting matters not only because it may improve zero-shot risk estimation, but also because it can improve the quality of the downstream representations available for clinical prediction. Additional analyses of discrimination and calibration, clinically relevant operating points, subgroup performance, and computational costs are provided in Appendices E.2, E.3, I, and J.

Additionally, concatenating past-window and generated-future features further improved AUROC over GFF alone in 24 of 28 settings (Appendix K), suggesting that observed-history and forecast-derived features provide complementary predictive information.

Limitations Generating future trajectories is computationally expensive because it requires autoregressive rollout sampling. In our setting, however, these trajectories are already produced for zero-shot evaluation, so fitting XGB-GFF on top adds little additional cost. More broadly, the practicality of this approach will depend on whether rollout generation is acceptable in the intended use case.

We evaluate only post-discharge binary prediction tasks in two cohorts, using AUROC as the primary metric. Generalization to other task families, clinical settings, and evaluation criteria remains to be established.

All experiments use the ETHOS architecture (Renc et al., 2024), with a separate model trained for each cohort; consequently, we do not establish generality across all generative architectures. We additionally do not evaluate transport of one trained model across hospitals or robustness to distribution shift. We use the same patient-level splits for ETHOS pretraining and downstream evaluation, and ETHOS is never trained on patients in the held-out split, preventing patient-level data leakage.

We additionally only compare to a single supervised transformer baseline, which is not intended to be a comprehensive survey of all possible supervised sequence model approaches.

In addition, generated-future features depend on rollout quality and on the chosen future summarization scheme. Different sampling budgets, vocabularies, or richer summaries of the forecast distribution may change downstream performance, so our method should be viewed as one simple instantiation of forecast-derived feature extraction.

7. Conclusion

We showed that, in our evaluated setting, zero-shot generative EHR forecasting can be used not only for direct risk estimation, but also as a practical feature extraction procedure for downstream clinical prediction. By converting sampled future trajectories from a

frozen generative model into horizon-specific generated-future features and training a simple XGBoost classifier on top, we obtained consistent improvements over direct zero-shot prediction across post-discharge tasks in MIMIC-IV and a large private heart failure cohort. Generated-future features also outperformed strong supervised baselines in most settings, suggesting that an important practical use of current generative EHR models is to provide forecast-derived representations of likely future patient state that can be exploited by simple downstream learners.

References

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data?

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58.

Aleksia Kolo, Chao Pang, Edward Choi, Ethan Steinberg, Hyewon Jeong, Jack Gallifant, Jason A Fries, Jeffrey N Chiang, Jungwoo Oh, Justin Xu, and others. Meds decentralized, extensible validation (meds-dev) benchmark: Establishing reproducibility and comparability in ml for health. 2024.

Matthew McDermott. MEDS "everything-is-code" autoregressive model. URL https://github.com/mmcdermott/MEDS_EIC_AR.

Matthew McDermott and Justin Xu. MIMIC-IV MEDS: An ETL pipeline to extract MIMIC-IV data into the MEDS format, May 2025. URL https://github.com/Medical-Event-Data-Standard/MIMIC_IV_MEDS.

Matthew B A McDermott, Bret Nestor, Peniel Argaw, Ye Jin, and Isaac Kohane. Event Stream GPT: A Data Pre-processing and Modeling Library for Generative, Pre-trained Transformers over Continuous-time Sequences of Complex Events.

Matthew B. A. McDermott, Justin Xu, Teya S. Bergamaschi, Hyewon Jeong, Simon A. Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, Robin P. van de Water, Michael Wornow, John Wu, and Zhenbang Wu. MEDS: Building Models and Tools in a Reproducible Health AI Ecosystem. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 6243–6244, New York, NY, USA, August 2025. Association for Computing Machinery. ISBN 979-8-4007-1454-2. doi: 10.1145/3711896.3737608. URL https://dl.acm.org/doi/10.1145/3711896.3737608.

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, October 2024. URL http://arxiv.org/abs/2411.00200. arXiv:2411.00200 [cs].

Chao Pang, Vincent Jeanselme, Young Sang Choi, Xinzhuo Jiang, Zilin Jing, Aparajita Kashyap, Yuta Kobayashi, Yanwei Li, Florent Pollet, Karthik Natarajan, and Shalmali Joshi. FoMoH: A clinically meaningful foundation model evaluation for structured electronic health records, June 2025. URL http://arxiv.org/abs/2505.16941. arXiv:2505.16941 [cs].

Ekaterina Redekop, Zichen Wang, Rushikesh Kulkarni, Mara Pleasure, Aaron Chin, Hamid Reza Hassanzadeh, Brian L Hill, Melika Emami, William F Speier, and Corey W Arnold. Zero-shot medical event prediction using a generative pretrained transformer on electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 32(12):1833–1842, October 2025. ISSN 1067-5027. doi: 10.1093/jamia/ocaf160. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12646381/.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1):256, September 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01235-0. URL https://www.nature.com/articles/s41746-024-01235-0.

Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew B A McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, David W Bates, and Arkadiusz Sitek. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14:giaf107, January 2025. ISSN 2047-217X. doi: 10.1093/gigascience/giaf107. URL https://doi.org/10.1093/gigascience/giaf107.

Ravid Shwartz-Ziv and Amitai Armon. Tabular Data: Deep Learning is Not All You Need. July 2021. URL https://openreview.net/forum?id=vdgtepS1pV.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative Medical Event Models Improve with Scale, November 2025. URL http://arxiv.org/abs/2508.12104. arXiv:2508.12104 [cs].

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHR Data. *International Conference on Learning Representations*, 2025:16518–16555, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2a246ec4437808525a088abae744c2d3-Abstract-Conference.html.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857, November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43715-z. URL https://www.nature.com/articles/s41467-023-43715-z.

Appendix A. Architecture and Experiment Details

Generative backbone. We pretrain ETHOS from scratch using the same architecture and training recipe in (Renc et al., 2024, 2025). We note that the dataset is split by patient id into a train, validation, and test set. The same splits are used for pretraining the generative model as are used for downstream tasks, and the generative model is never trained on patients in the test set. The epoch checkpoint with the best validation performance is selected. Generative rollouts are anchored at hospital discharge (i.e. the dataset-specific discharge index event). For each patient and split, we sample 50 future trajectories and convert each trajectory into time-to-first-event values for all codes in the evaluation vocabulary. ETHOS rollouts are treated identically downstream; in our runs these were read from the 512-context-length ETHOS trajectory cache and converted into the same time-to-event/probability representation. The generative model is not fine-tuned for downstream tasks.

Generated-future features. For a patient j, horizon H, code c, and $S = 50$ generated futures, we define

$$x_{j,c}^{\text{future}}(H) = \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\{T_{j,c}^{(s)} \leq H\},$$

where $T_{j,c}^{(s)}$ is the generated time to first occurrence of code c in rollout s. Direct zero-shot prediction uses the same estimator for the target endpoint only. XGB-GFF uses the full vector over generated endpoint/code probabilities. Mean and standard deviations for zero-shot AUROC are estimated using a leave one out jackknife estimator over the 50 generated trajectories.

Past-window tabular features. XGB-Past uses MEDS-Tab rolling-window features derived only from observed history before the index time. We use code-count aggregations over five lookback windows: 1 day, 7 days, 30 days, 365 days, and the full available history. No additional imputation or normalization is applied in the XGBoost pipeline.

XGBoost training and tuning. For each dataset, endpoint, horizon, feature group, and generative backbone, XGBoost is tuned independently. We sample 250 hyperparameter configurations using search seed 1337 and select the configuration with the highest tuning AUROC. The search space is:

$$\eta \in \{0.003, 0.01, 0.03, 0.1, 0.2, 0.3\}, \quad \lambda \in \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0\},$$

$$\alpha \in \{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0\}, \quad \text{subsample} \in \{0.5, 0.7, 0.85, 1.0\},$$

$$\text{min\_child\_weight} \in \{0.01, 0.1, 1, 5, 10, 50, 100\}, \quad \text{max\_depth} \in \{2, 3, 4, 6, 8, 10, 12, 14, 16\},$$

$$\text{n\_estimators} \in \{50, 100, 200, 400, 800, 1200\}.$$

All models use binary logistic objective, AUC evaluation metric, histogram tree construction, column subsampling fixed at 1.0, $\gamma = 0$, and 10 early-stopping rounds. After selecting hyperparameters, we train five models with random seeds 0–4 and report mean $\pm$ standard deviation on the held-out split.

Supervised transformer baseline. The supervised sequence baseline is trained separately for each dataset–task–horizon setting. It uses transformer decoder architecture with: 4 layers, 4 heads, head dimension 64, hidden size 256, and feed-forward dimension 1024. The classifier is a linear head applied to the final non-padding hidden state. Models are trained with binary cross-entropy, AdamW, learning rate 10^{-3}, weight decay 10^{-3}, batch size 512, maximum sequence length 128, 10 epochs, mixed precision, and gradient clipping at 1.0. Results are aggregated over seeds 0–4. Mean and standard deviations for AUROC are estimated over five seeds.

Linear Probing. For each task we trained a logistic regression model fitted on the last embedding token from ETHOS. Past work Wornow et al. (2025) has shown simply using the last token in this setup is most performant. Embedding features were z-score normalized (statistics computed on the training split), and logistic regression was fitted with ℓ_2 regularization. We swept the regularization strength α over

$$\alpha \in \{10^{-6}, 3\times10^{-6}, 10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}.$$

Model selection was performed using only the validation split. After hyperparameter selection, we refit the probe using the selected α across five seeds, and evaluated these models on the test split.

Appendix B. Discretized MEDS Sequence Representation

Raw MEDS (McDermott et al., 2025) data must be discretized before being passed to ETHOS. We use the same reproducible MEDS discretization pipeline introduced in the MEDS-EIC-AR (McDermott) repository. The preprocessing pipeline first filters the raw MEDS event stream to a retained vocabulary containing timeline markers, hospital utilization events, death indicators, and the laboratory concepts required for the evaluated endpoints. Patients are retained only if they have at least 10 observations after this vocabulary filtering step.

Non-numeric events are represented directly as discrete codes. Numeric-valued measurements are converted into discrete tokens by fitting per-code empirical quantile bins on the training data and replacing raw numeric values with bin identifiers. We use decile binning, with cut points at the 0.1, 0.2, ..., 0.9 quantiles. The raw numeric value is dropped after binning, so the sequence models observe only discrete event/value-bin tokens. Time intervals are also discretized into explicit timeline delta tokens, using bins ranging from minutes to multi-year gaps.

For Hospital A, the retained laboratory codes are LOINC 2160-0 for creatinine, LOINC 777-3 for platelet count, LOINC 6690-2 for leukocyte count, LOINC 718-7 for hemoglobin, and LOINC 4544-3 for hematocrit. For MIMIC, the retained laboratory item codes map to the same clinical concepts: creatinine, platelet, leukocyte, hemoglobin, and hematocrit. In both datasets, abnormal laboratory outcomes are computed by resolving each observed discretized lab token back to its associated value interval and checking whether the interval crosses the task threshold by horizon H.

This discretization has two roles. First, it provides a common input representation for autoregressive sequence modeling over EHR events. Second, it ensures that direct zero-shot

risks, generated-future features, past-window features, and supervised sequence baselines are evaluated on a shared event vocabulary rather than on task-specific raw-data feature engineering.

Table 5: **Full-index cohort age and follow-up.** Age is years at indexed discharge; follow-up is observed post-discharge follow-up in days. Leading values are medians and values in brackets are the interquartile range. Mean indexed discharges use all preprocessed patients as the denominator.

Metric	Hospital A	MIMIC-IV
Median age at discharge, y [IQR]	69.8 [58.6, 79.3]	63.0 [49.4, 75.2]
Median observed follow-up, d [IQR]	1019.1 [207.0, 2584.0]	620.6 [189.2, 1720.6]
Mean indexed discharges per preprocessed patient	8.0	2.6
Patients	179,227	123,059

Appendix C. Cohort Context

Hospital A is a heart-failure cohort from a large tertiary health system, whereas MIMIC-IV is a general critical-care discharge cohort. Table 5 provides additional context on age, observed follow-up, and repeated indexed discharges. Statistics are computed over the union of train, tuning, and held-out prediction indices.

Appendix D. Additional Evaluation Curves

This appendix section provides additional held-out evaluation curves comparing XGB-GFF and XGB-Past. Each panel aggregates the corresponding task-level curves within a dataset and horizon.

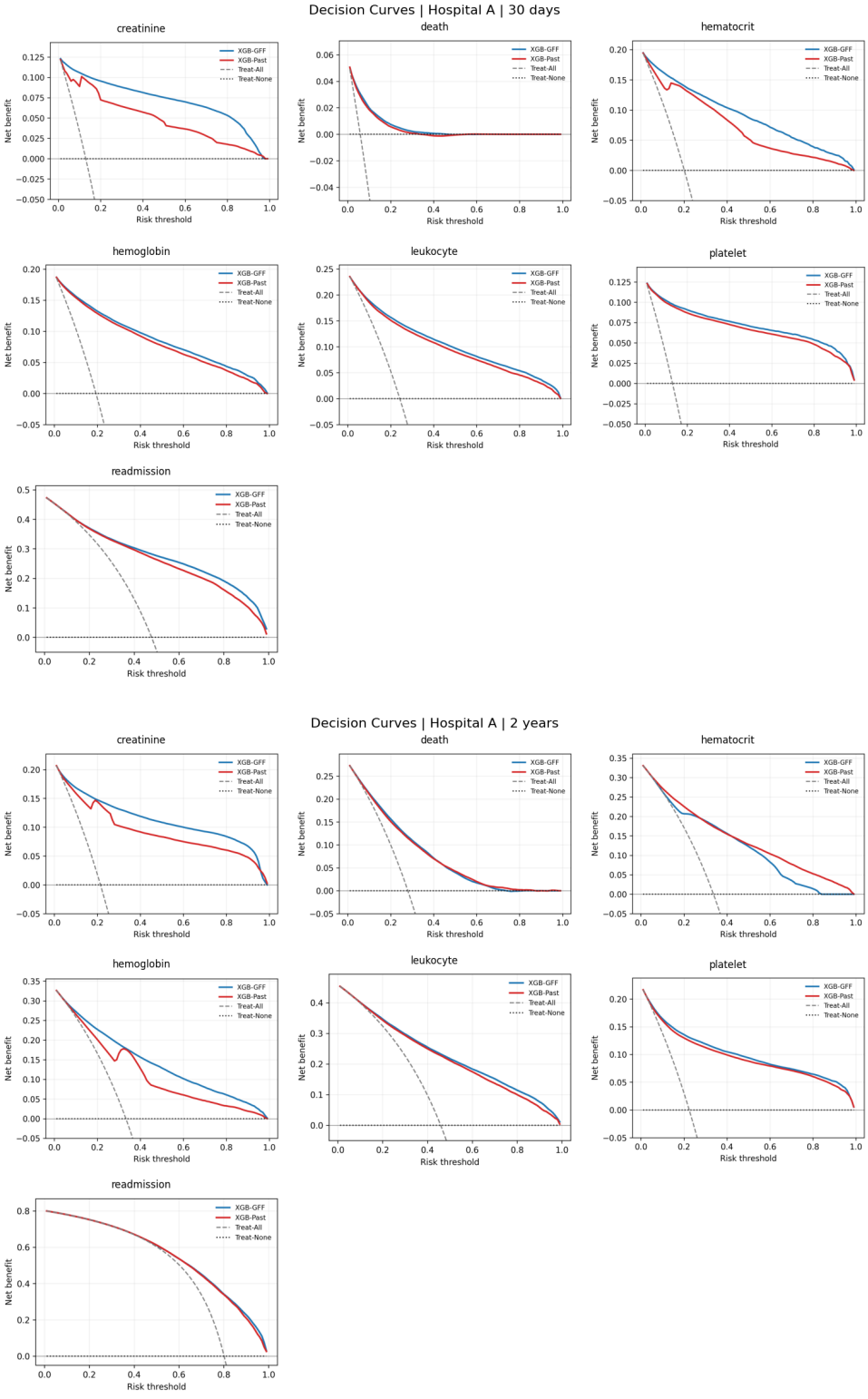


Figure 3: **Decision-curve analysis.** Net benefit for XGB-GFF, XGB-Past, treat-all, and treat-none strategies across risk thresholds for Hospital A.

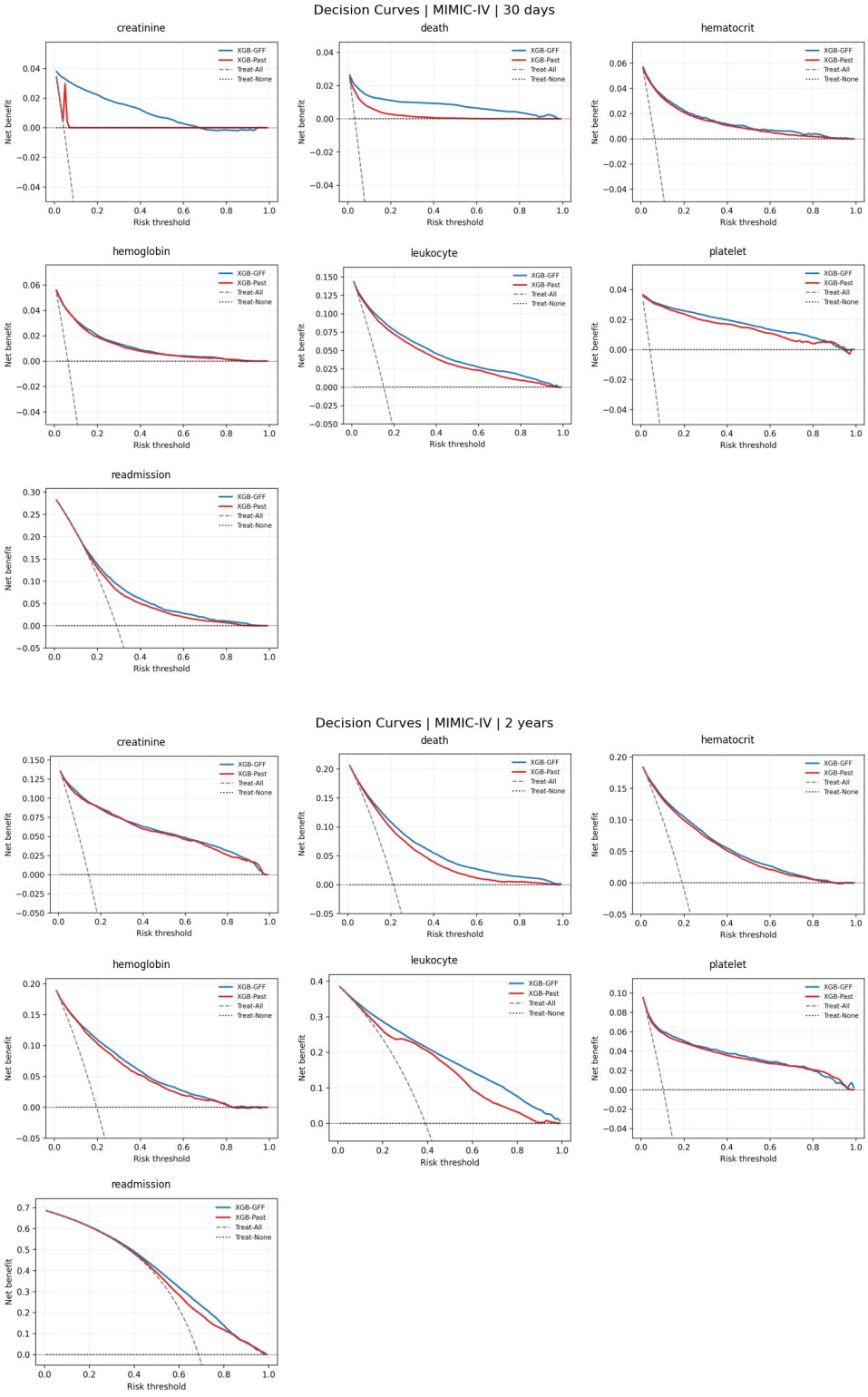


Figure 4: **Decision-curve analysis.** Net benefit for XGB-GFF, XGB-Past, treat-all, and treat-none strategies across risk thresholds for MIMIC-IV.

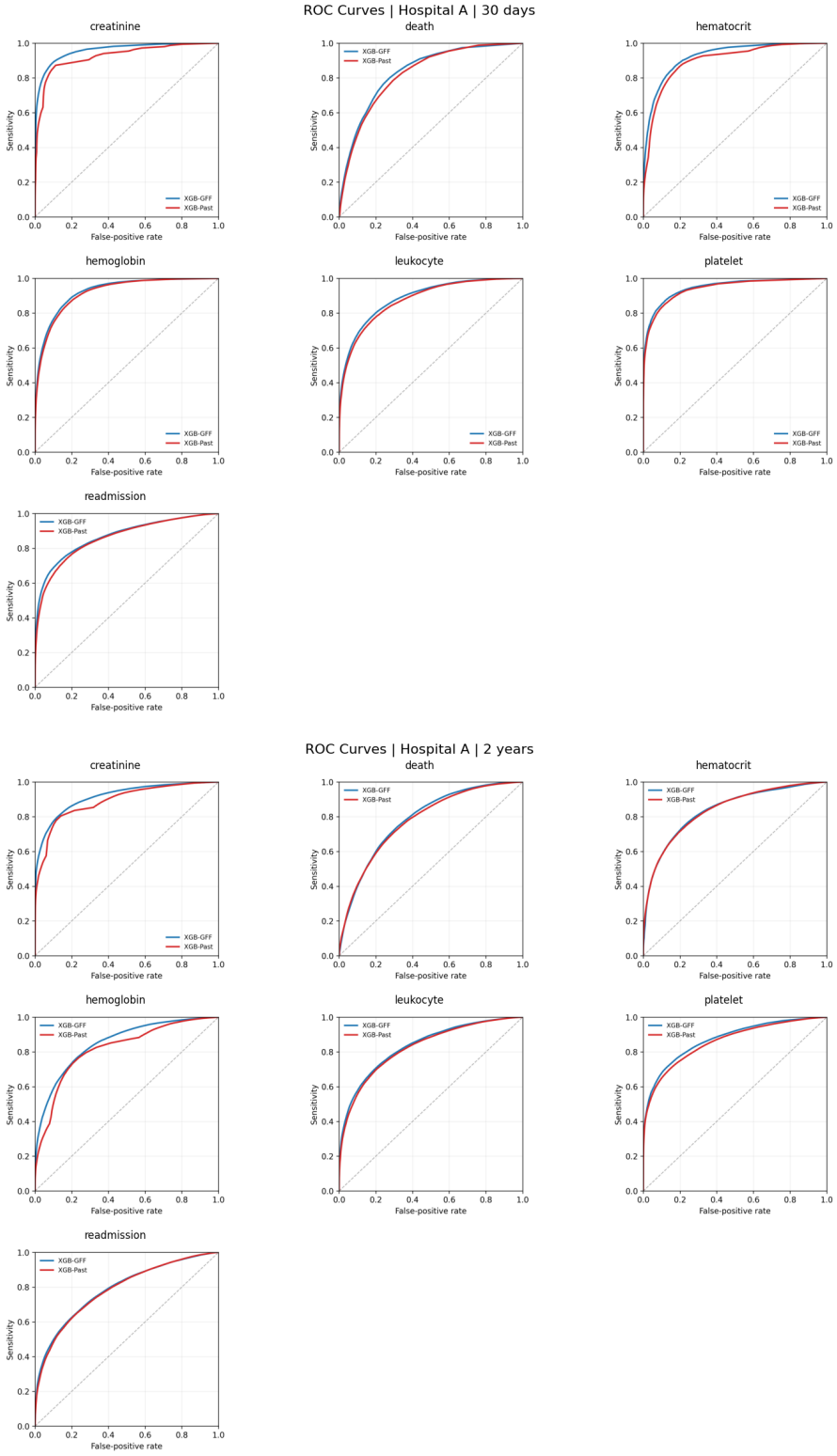


Figure 5: **ROC curves.** Sensitivity versus false-positive rate (1−specificity) for XGB-GFF and XGB-Past for Hospital A.

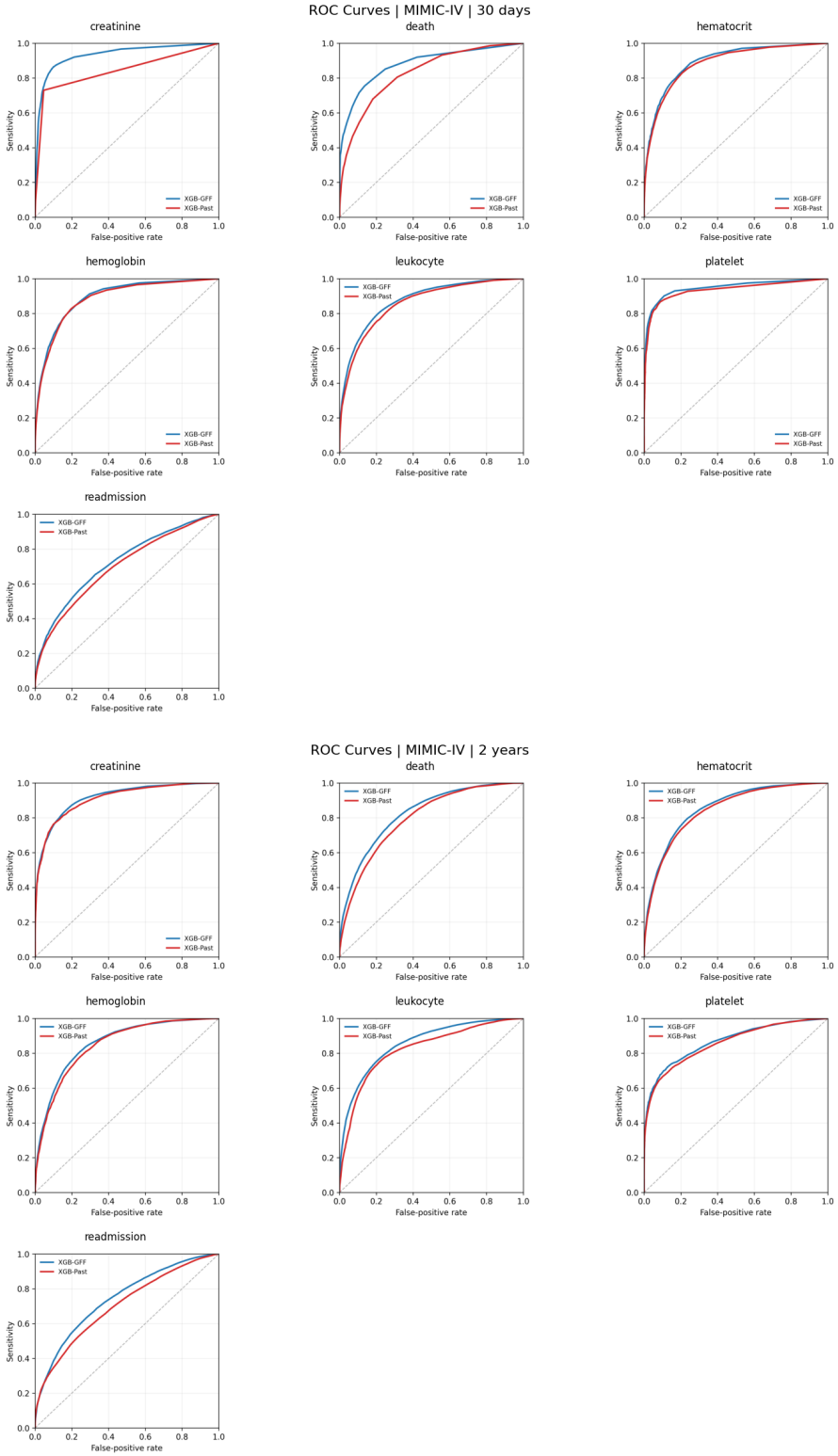


Figure 6: **ROC curves.** Sensitivity versus false-positive rate (1−specificity) for XGB-GFF and XGB-Past for MIMIC-IV.

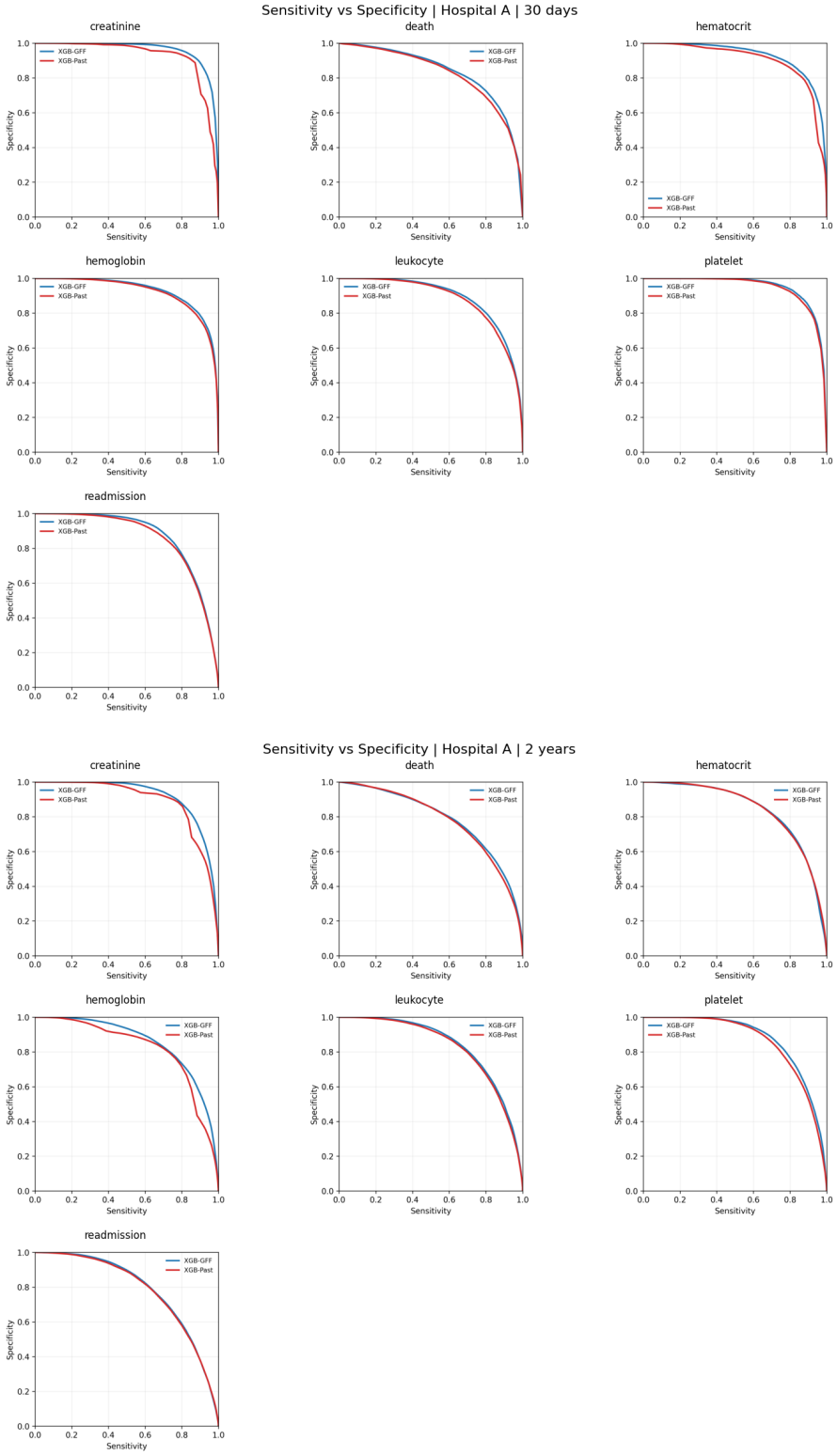


Figure 7: **Sensitivity–specificity curves.** Tradeoff between sensitivity and specificity for XGB-GFF and XGB-Past across prediction thresholds for Hospital A.

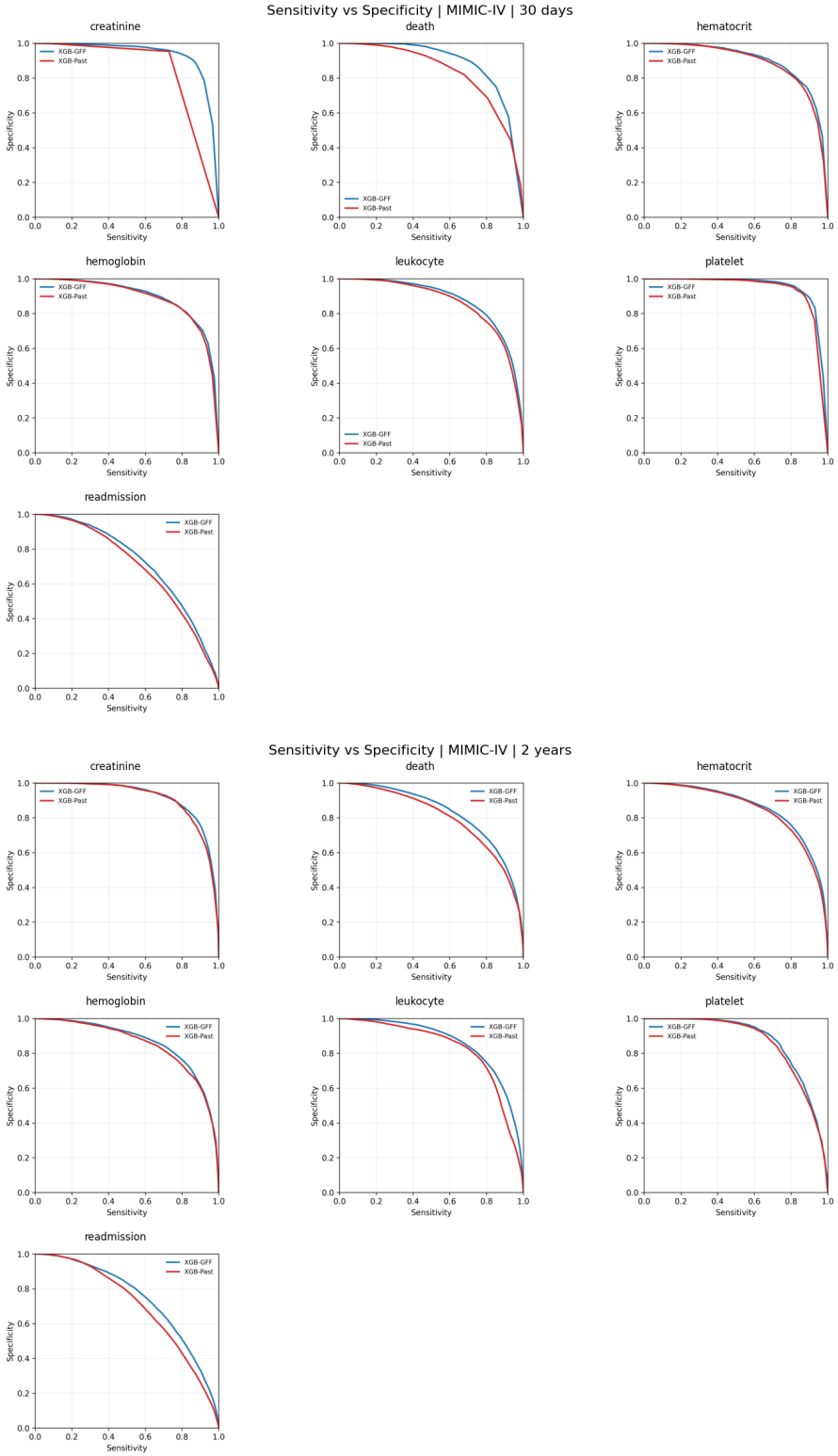


Figure 8: **Sensitivity–specificity curves.** Tradeoff between sensitivity and specificity for XGB-GFF and XGB-Past across prediction thresholds for MIMIC-IV.

Appendix E. Additional Evaluation Results

We computed additional held-out metrics comparing XGB-GFF against XGB-Past using the same task, horizon, dataset, and seed settings as the main supervised XGBoost comparisons. All values below are computed on the held-out split and averaged over five XGBoost random seeds. We refer to the private cohort as Hospital A throughout.

E.1. Aggregate Paired Significance Check

As a complementary aggregate check to the per-setting tests in Table 2, we compared the mean XGB-GFF and XGB-Past AUROCs across the 28 paired dataset–endpoint–horizon settings using a one-sided exact Wilcoxon signed-rank test. XGB-GFF was higher in all 28 pairs; the mean gain was 1.38 AUROC points and the exact one-sided p-value was 3.73×10^{-9}.

E.2. AUPRC, Brier Score, and Calibration

Across the 14 task–horizon settings in each dataset, XGB-GFF improved AUPRC in 13/14 settings in Hospital A and 14/14 settings in MIMIC-IV. XGB-GFF also achieved lower Brier score in 13/14 Hospital A settings and 14/14 MIMIC-IV settings. Positive AUPRC deltas favor XGB-GFF.

Table 6: **Additional discrimination and calibration metrics by task.** AP denotes AUPRC; Br denotes Brier score. Subscripts P and G denote XGB-Past and XGB-GFF, respectively.

Dataset	Task	H	AP_P	AP_G	ΔAP	Br_P	Br_G
Hospital A	creatinine	30	0.820	0.857	+0.036	0.063	0.042
Hospital A	death	30	0.219	0.254	+0.035	0.050	0.049
Hospital A	hematocrit	30	0.760	0.789	+0.029	0.099	0.082
Hospital A	hemoglobin	30	0.768	0.790	+0.022	0.083	0.079
Hospital A	leukocyte	30	0.746	0.769	+0.022	0.110	0.105
Hospital A	platelet	30	0.814	0.834	+0.019	0.049	0.046
Hospital A	readmission	30	0.867	0.882	+0.015	0.150	0.142
Hospital A	creatinine	730	0.811	0.827	+0.017	0.097	0.079
Hospital A	death	730	0.575	0.567	-0.009	0.162	0.161
Hospital A	hematocrit	730	0.759	0.767	+0.007	0.147	0.158
Hospital A	hemoglobin	730	0.759	0.770	+0.011	0.166	0.141
Hospital A	leukocyte	730	0.818	0.831	+0.013	0.168	0.163
Hospital A	platelet	730	0.742	0.762	+0.020	0.104	0.099
Hospital A	readmission	730	0.935	0.937	+0.003	0.133	0.132
MIMIC-IV	creatinine	30	0.491	0.527	+0.036	0.041	0.027
MIMIC-IV	death	30	0.261	0.487	+0.226	0.029	0.022
MIMIC-IV	hematocrit	30	0.466	0.494	+0.027	0.045	0.043
MIMIC-IV	hemoglobin	30	0.433	0.449	+0.016	0.045	0.044
MIMIC-IV	leukocyte	30	0.590	0.635	+0.046	0.090	0.085
MIMIC-IV	platelet	30	0.663	0.713	+0.050	0.022	0.020
MIMIC-IV	readmission	30	0.527	0.560	+0.034	0.181	0.175
MIMIC-IV	creatinine	730	0.730	0.742	+0.012	0.067	0.066
MIMIC-IV	death	730	0.526	0.599	+0.073	0.134	0.124
MIMIC-IV	hematocrit	730	0.588	0.610	+0.021	0.112	0.108
MIMIC-IV	hemoglobin	730	0.588	0.613	+0.025	0.114	0.110
MIMIC-IV	leukocyte	730	0.773	0.808	+0.035	0.167	0.147
MIMIC-IV	platelet	730	0.627	0.644	+0.017	0.059	0.057
MIMIC-IV	readmission	730	0.844	0.861	+0.017	0.191	0.181

Table 7: **Win rates and average improvements for AUPRC, Brier score.** Improvements are XGB-GFF minus XGB-Past for AUPRC, and XGB-Past minus XGB-GFF for Brier score.

Dataset	Horizon	Metric	GFF wins	Past mean	GFF mean	Avg. improvement
Hospital A	30d	AUPRC	7/7 (1.000)	0.714	0.739	+0.026
Hospital A	730d	AUPRC	6/7 (0.857)	0.771	0.780	+0.009
MIMIC-IV	30d	AUPRC	7/7 (1.000)	0.490	0.552	+0.062
MIMIC-IV	730d	AUPRC	7/7 (1.000)	0.668	0.697	+0.029
Hospital A	30d	Brier	7/7 (1.000)	0.086	0.078	+0.009
Hospital A	730d	Brier	6/7 (0.857)	0.139	0.133	+0.006
MIMIC-IV	30d	Brier	7/7 (1.000)	0.065	0.059	+0.005
MIMIC-IV	730d	Brier	7/7 (1.000)	0.121	0.113	+0.007

E.3. Fixed-Sensitivity Operating Points

At matched sensitivity targets of 50%, 80%, and 90%, XGB-GFF generally reached the target sensitivity with higher specificity and PPV while flagging fewer patients. At the 80% sensitivity setting, averaged over tasks and horizons, XGB-GFF increased specificity by 0.019 in Hospital A and 0.035 in MIMIC-IV, increased PPV by 0.025 and 0.027, respectively, and reduced false positives by 14.0 and 26.8 per 1,000 discharges. Values in Table 8 are averaged over tasks within each dataset and horizon. Negative changes in fraction flagged and false positives per 1,000 discharges favor XGB-GFF. Across all 3 sensitivity targets and operating point metrics in Table 8, XGB-GFF achieves higher performance than XGB-Past.

Table 8: **Fixed-sensitivity operating points.** Values are averaged over tasks within each dataset and horizon.

Dataset	Horizon	Sens.	Spec_P	Spec_G	ΔSpec	PPV_P	PPV_G	ΔPPV	ΔFlag	ΔFP/1000
Hospital A	30d	0.50	0.963	0.970	+0.007	0.772	0.800	+0.028	-0.005	-5.3
Hospital A	30d	0.80	0.829	0.850	+0.021	0.553	0.588	+0.035	-0.017	-16.7
Hospital A	30d	0.90	0.696	0.725	+0.028	0.434	0.463	+0.029	-0.023	-22.9
Hospital A	730d	0.50	0.927	0.933	+0.005	0.812	0.827	+0.015	-0.003	-3.1
Hospital A	730d	0.80	0.695	0.712	+0.017	0.605	0.620	+0.015	-0.011	-11.4
Hospital A	730d	0.90	0.509	0.530	+0.021	0.516	0.527	+0.011	-0.015	-14.8
MIMIC-IV	30d	0.50	0.929	0.946	+0.018	0.481	0.544	+0.064	-0.015	-15.5
MIMIC-IV	30d	0.80	0.772	0.805	+0.033	0.293	0.321	+0.028	-0.029	-29.2
MIMIC-IV	30d	0.90	0.639	0.677	+0.038	0.202	0.223	+0.020	-0.035	-34.8
MIMIC-IV	730d	0.50	0.909	0.927	+0.018	0.688	0.725	+0.037	-0.010	-10.4
MIMIC-IV	730d	0.80	0.689	0.727	+0.038	0.478	0.504	+0.027	-0.024	-24.5
MIMIC-IV	730d	0.90	0.519	0.563	+0.045	0.398	0.420	+0.023	-0.030	-29.6

Table 9: **Full fixed-sensitivity operating point results.** Each row is one dataset–endpoint–horizon experiment. Within each target sensitivity block, columns report specificity, PPV, fraction flagged, and false positives per 1,000 discharges for XGB-Past (P) and XGB-GFF (G).

			50% sensitivity								80% sensitivity								90% sensitivity							
Dataset	Endpoint	Horizon	Spec_P	Spec_G	PPV_P	PPV_G	Flag_P	Flag_G	FP_P	FP_G	Spec_P	Spec_G	PPV_P	PPV_G	Flag_P	Flag_G	FP_P	FP_G	Spec_P	Spec_G	PPV_P	PPV_G	Flag_P	Flag_G	FP_P	FP_G
Hospital A	creatinine	30d	0.994	0.996	0.922	0.954	0.070	0.068	5.5	3.1	0.940	0.958	0.667	0.738	0.156	0.141	51.9	36.9	0.856	0.889	0.482	0.547	0.242	0.214	125.6	96.9
Hospital A	death	30d	0.890	0.901	0.220	0.238	0.133	0.123	103.4	93.6	0.691	0.725	0.139	0.153	0.337	0.305	290.7	258.7	0.553	0.581	0.111	0.118	0.474	0.447	421.3	394.0
Hospital A	hematocrit	30d	0.966	0.975	0.788	0.833	0.128	0.121	27.2	20.2	0.859	0.882	0.590	0.632	0.274	0.255	112.2	93.8	0.756	0.787	0.483	0.516	0.376	0.352	194.4	170.1
Hospital A	hemoglobin	30d	0.972	0.978	0.812	0.846	0.119	0.114	22.4	17.6	0.864	0.878	0.586	0.612	0.265	0.253	109.6	98.4	0.762	0.787	0.476	0.504	0.367	0.346	192.2	171.4
Hospital A	leukocyte	30d	0.958	0.965	0.793	0.821	0.153	0.148	31.7	26.5	0.774	0.802	0.531	0.564	0.365	0.344	171.4	150.1	0.605	0.644	0.422	0.448	0.518	0.488	299.3	269.3
Hospital A	platelet	30d	0.995	0.997	0.941	0.957	0.069	0.068	4.1	2.9	0.924	0.938	0.613	0.659	0.170	0.158	65.7	53.9	0.822	0.844	0.430	0.463	0.272	0.253	155.0	135.6
Hospital A	readmission	30d	0.963	0.976	0.925	0.951	0.258	0.251	19.4	12.3	0.752	0.766	0.747	0.758	0.512	0.504	129.3	122.0	0.522	0.540	0.633	0.642	0.680	0.670	249.7	240.3
Hospital A	creatinine	730d	0.987	0.992	0.914	0.945	0.117	0.114	10.1	6.3	0.865	0.876	0.619	0.639	0.277	0.269	105.8	97.1	0.699	0.720	0.450	0.467	0.430	0.413	236.4	220.2
Hospital A	death	730d	0.854	0.855	0.571	0.572	0.245	0.245	105.2	104.7	0.595	0.615	0.434	0.447	0.516	0.501	291.7	277.3	0.429	0.462	0.380	0.394	0.663	0.639	411.2	387.0
Hospital A	hematocrit	730d	0.935	0.937	0.797	0.802	0.212	0.210	43.1	41.5	0.705	0.717	0.580	0.590	0.465	0.457	195.6	187.6	0.520	0.533	0.488	0.495	0.622	0.613	318.1	309.5
Hospital A	hemoglobin	730d	0.931	0.936	0.784	0.795	0.212	0.209	45.8	43.0	0.722	0.733	0.589	0.599	0.452	0.444	185.7	178.0	0.540	0.560	0.494	0.505	0.607	0.593	307.1	293.5
Hospital A	leukocyte	730d	0.925	0.938	0.850	0.873	0.270	0.263	40.4	33.5	0.670	0.685	0.673	0.683	0.546	0.538	178.7	170.6	0.466	0.494	0.588	0.601	0.702	0.687	289.0	273.9
Hospital A	platelet	730d	0.969	0.975	0.822	0.851	0.137	0.132	24.4	19.6	0.726	0.766	0.458	0.497	0.392	0.361	212.3	181.6	0.529	0.562	0.356	0.373	0.567	0.542	365.5	339.8
Hospital A	readmission	730d	0.891	0.898	0.949	0.952	0.422	0.421	21.7	20.2	0.580	0.590	0.885	0.887	0.725	0.723	83.5	81.5	0.379	0.380	0.854	0.854	0.845	0.844	123.4	123.1
MIMIC-IV	creatinine	30d	0.983	0.984	0.568	0.588	0.039	0.037	16.7	15.3	0.926	0.937	0.331	0.369	0.106	0.095	71.0	60.2	0.828	0.838	0.194	0.204	0.204	0.194	164.4	154.8
MIMIC-IV	death	30d	0.914	0.973	0.168	0.388	0.100	0.043	83.0	26.1	0.698	0.816	0.084	0.130	0.319	0.204	292.0	177.9	0.549	0.636	0.064	0.078	0.466	0.381	435.8	351.6
MIMIC-IV	hematocrit	30d	0.952	0.958	0.417	0.450	0.077	0.071	44.7	39.0	0.818	0.826	0.231	0.239	0.221	0.214	170.3	162.6	0.688	0.729	0.164	0.184	0.350	0.312	292.2	254.1
MIMIC-IV	hemoglobin	30d	0.947	0.951	0.389	0.407	0.081	0.077	49.4	45.7	0.826	0.827	0.235	0.236	0.213	0.212	163.2	162.2	0.705	0.725	0.170	0.180	0.333	0.314	276.2	257.6
MIMIC-IV	leukocyte	30d	0.937	0.951	0.585	0.647	0.130	0.117	53.8	41.4	0.753	0.791	0.367	0.406	0.331	0.299	209.4	177.6	0.605	0.636	0.289	0.306	0.472	0.446	335.5	309.1
MIMIC-IV	platelet	30d	0.993	0.995	0.763	0.811	0.027	0.026	6.5	4.9	0.956	0.962	0.442	0.482	0.076	0.070	42.3	36.1	0.850	0.892	0.208	0.267	0.181	0.141	143.3	103.8
MIMIC-IV	readmission	30d	0.774	0.812	0.474	0.520	0.305	0.278	160.3	133.6	0.428	0.474	0.363	0.383	0.638	0.605	406.3	373.6	0.247	0.285	0.327	0.339	0.796	0.768	535.3	508.0
MIMIC-IV	creatinine	730d	0.980	0.982	0.806	0.819	0.089	0.087	17.2	15.8	0.858	0.869	0.485	0.503	0.236	0.227	121.4	112.7	0.700	0.758	0.333	0.382	0.385	0.336	256.9	207.8
MIMIC-IV	death	730d	0.868	0.901	0.507	0.577	0.210	0.185	103.7	78.2	0.630	0.686	0.370	0.409	0.462	0.418	290.9	247.0	0.490	0.533	0.324	0.344	0.593	0.559	401.3	367.1
MIMIC-IV	hematocrit	730d	0.918	0.925	0.592	0.611	0.162	0.157	65.9	60.9	0.731	0.759	0.414	0.440	0.370	0.348	217.2	195.2	0.567	0.596	0.330	0.346	0.523	0.499	350.3	326.5
MIMIC-IV	hemoglobin	730d	0.913	0.925	0.583	0.618	0.168	0.159	70.2	60.6	0.734	0.764	0.424	0.453	0.371	0.347	213.3	189.8	0.606	0.614	0.358	0.363	0.494	0.487	316.9	309.8
MIMIC-IV	leukocyte	730d	0.923	0.942	0.806	0.847	0.242	0.231	46.9	35.3	0.724	0.747	0.650	0.670	0.480	0.466	168.0	154.0	0.502	0.577	0.537	0.577	0.655	0.609	303.3	257.7
MIMIC-IV	platelet	730d	0.973	0.979	0.681	0.734	0.076	0.071	24.3	18.8	0.713	0.751	0.244	0.271	0.340	0.306	257.0	222.8	0.504	0.532	0.174	0.182	0.538	0.513	444.5	419.2
MIMIC-IV	readmission	730d	0.788	0.834	0.838	0.869	0.410	0.395	66.4	51.9	0.431	0.511	0.756	0.783	0.728	0.703	177.8	152.8	0.262	0.332	0.728	0.748	0.849	0.827	230.8	208.7

Table 10: **Compact summary of additional evaluation metrics.** For AUPRC, improvement is XGB-GFF minus XGB-Past. For Brier, improvement is XGB-Past minus XGB-GFF. For fixed-sensitivity burden metrics, positive improvement denotes reduced alert burden or fewer false positives for XGB-GFF.

Dataset	Metric	GFF wins	Avg. improvement
Hospital A	AUPRC	13/14 (0.929)	+0.017
Hospital A	Brier score	13/14 (0.929)	+0.007
Hospital A	50% sens.: specificity	14/14 (1.000)	+0.006
Hospital A	50% sens.: PPV	14/14 (1.000)	+0.022
Hospital A	50% sens.: fraction flagged	14/14 (1.000)	+0.004
Hospital A	50% sens.: FP/1000	14/14 (1.000)	+4.2
Hospital A	80% sens.: specificity	14/14 (1.000)	+0.019
Hospital A	80% sens.: PPV	14/14 (1.000)	+0.025
Hospital A	80% sens.: fraction flagged	14/14 (1.000)	+0.014
Hospital A	80% sens.: FP/1000	14/14 (1.000)	+14.0
Hospital A	90% sens.: specificity	14/14 (1.000)	+0.025
Hospital A	90% sens.: PPV	14/14 (1.000)	+0.020
Hospital A	90% sens.: fraction flagged	14/14 (1.000)	+0.019
Hospital A	90% sens.: FP/1000	14/14 (1.000)	+18.8
MIMIC-IV	AUPRC	14/14 (1.000)	+0.045
MIMIC-IV	Brier score	14/14 (1.000)	+0.006
MIMIC-IV	50% sens.: specificity	14/14 (1.000)	+0.018
MIMIC-IV	50% sens.: PPV	14/14 (1.000)	+0.051
MIMIC-IV	50% sens.: fraction flagged	14/14 (1.000)	+0.013
MIMIC-IV	50% sens.: FP/1000	14/14 (1.000)	+13.0
MIMIC-IV	80% sens.: specificity	14/14 (1.000)	+0.035
MIMIC-IV	80% sens.: PPV	14/14 (1.000)	+0.027
MIMIC-IV	80% sens.: fraction flagged	14/14 (1.000)	+0.027
MIMIC-IV	80% sens.: FP/1000	14/14 (1.000)	+26.8
MIMIC-IV	90% sens.: specificity	14/14 (1.000)	+0.041
MIMIC-IV	90% sens.: PPV	14/14 (1.000)	+0.021
MIMIC-IV	90% sens.: fraction flagged	14/14 (1.000)	+0.032
MIMIC-IV	90% sens.: FP/1000	14/14 (1.000)	+32.2

Appendix F. Held-Out Cohort Prevalence Checks

We additionally checked that held-out outcome prevalences were close to the corresponding full-cohort prevalences for each dataset, task, and horizon. The full cohort is defined as the union of the train, tuning, and held-out indexed discharges. Outcome prevalence is the proportion of indexed discharges with the task event observed within the stated horizon.

Table 11: **Split sizes.** Indexed discharges are rows in the prediction index, and patients are unique subject identifiers in that index. Full cohort size is train plus tuning plus held-out indexed discharges.

Dataset	Split	Indexed discharges	Patients
Hospital A	train	1,454,785	170,201
Hospital A	tuning	37,973	4,508
Hospital A	held-out	39,135	4,518
Hospital A	full	1,531,893	179,227
MIMIC-IV	train	377,505	116,865
MIMIC-IV	tuning	9,592	3,053
MIMIC-IV	held-out	10,640	3,141
MIMIC-IV	full	397,737	123,059

Table 12: **Full-cohort and held-out task prevalences.** Prevalences and absolute differences are reported.

Dataset	Task	Horizon	Full prev. (%)	Held-out prev. (%)	Abs. diff. (pp)
Hospital A	creatinine	30d	11.4	13.0	1.61
Hospital A	creatinine	730d	19.5	21.5	1.98
Hospital A	death	30d	5.8	5.8	0.10
Hospital A	death	730d	27.9	28.0	0.10
Hospital A	hematocrit	30d	19.9	20.2	0.30
Hospital A	hematocrit	730d	33.8	33.7	0.06
Hospital A	hemoglobin	30d	19.5	19.4	0.13
Hospital A	hemoglobin	730d	33.4	33.3	0.08
Hospital A	leukocyte	30d	24.0	24.3	0.28
Hospital A	leukocyte	730d	45.9	45.9	0.06
Hospital A	platelet	30d	11.5	13.0	1.48
Hospital A	platelet	730d	21.1	22.4	1.30
Hospital A	readmission	30d	47.8	47.8	0.00
Hospital A	readmission	730d	80.1	80.1	0.06
MIMIC-IV	creatinine	30d	4.7	4.4	0.31
MIMIC-IV	creatinine	730d	13.5	14.3	0.81
MIMIC-IV	death	30d	3.2	3.3	0.12
MIMIC-IV	death	730d	20.7	21.3	0.64
MIMIC-IV	hematocrit	30d	6.3	6.4	0.04
MIMIC-IV	hematocrit	730d	19.4	19.2	0.25
MIMIC-IV	hemoglobin	30d	6.1	6.3	0.14
MIMIC-IV	hemoglobin	730d	19.4	19.7	0.28
MIMIC-IV	leukocyte	30d	13.4	15.2	1.78
MIMIC-IV	leukocyte	730d	36.8	39.0	2.28
MIMIC-IV	platelet	30d	4.2	4.2	0.04
MIMIC-IV	platelet	730d	11.0	10.4	0.64
MIMIC-IV	readmission	30d	26.4	28.9	2.50
MIMIC-IV	readmission	730d	66.9	68.7	1.87

Appendix G. Endpoint Code Counts

This section provides Table 13 which specifies the number of codes that indicate each endpoint. I.e. for MIMIC-IV readmission, there are 71 codes that match the readmission task.

Table 13: **Endpoint-defining code counts.** Code counts are expanded from the dataset ACES predicate definitions and ignore prediction horizon. Type is single-code when exactly one concrete code defines the endpoint and multi-code otherwise.

Dataset	Endpoint	Codes	Type
Hospital A	creatinine	1	single-code
Hospital A	death	1	single-code
Hospital A	hematocrit	1	single-code
Hospital A	hemoglobin	1	single-code
Hospital A	leukocyte	2	multi-code
Hospital A	platelet	1	single-code
Hospital A	readmission	1	single-code
MIMIC-IV	creatinine	5	multi-code
MIMIC-IV	death	1	single-code
MIMIC-IV	hematocrit	3	multi-code
MIMIC-IV	hemoglobin	3	multi-code
MIMIC-IV	leukocyte	7	multi-code
MIMIC-IV	platelet	1	single-code
MIMIC-IV	readmission	71	multi-code

Appendix H. XGB-GFF Ablations

We additionally evaluated how sensitive XGB-GFF is to the number of generated rollouts used to construct GFFs and to the size of the retained generated-feature vocabulary. The full GFF vectors contain 100 generated-event probability features for Hospital A and 308 for MIMIC-IV. All rows use the same XGBoost tuning and five-seed refit protocol as the main XGB-GFF experiments. The XGB-Past column is copied from the main past-window baseline. The full-budget XGB-GFF columns, $S = 50$ and $V = 100\%$, are copied from the main XGB-GFF experiment. We note that some identical displayed AUROCs across some vocabulary sizes in Table 16 arise when the additional low-frequency features are not selected by the fitted trees.

H.1. Removal of Target-Adjacent Generated Features

To test whether the GFF advantage is solely explained by direct target proxies, we removed the generated-feature columns matching each endpoint family before tuning and fitting XGBoost; we denote this ablation XGB-GFF*. For laboratory outcomes, removal covers every retained value bin for the corresponding laboratory concept; for death it removes the generated death feature; and for readmission it removes every generated hospital-admission feature. XGB-GFF* remained above XGB-Past in 16/28 settings. Full XGB-GFF was above XGB-GFF* in 27/28 settings; the remaining Hospital A readmission-at-730-days setting differed by 0.07 AUROC points in favor of XGB-GFF*. Thus, target-adjacent features carry substantial signal, but they do not fully account for the advantage over past-window features.

We note that we removed all codes related to each endpoint. For example, for the abnormal creatinine value tasks we removed all creatinine features of all deciles, not just the deciles corresponding to the abnormal creatinine endpoint definition.

Table 14: **Rollout-sample sensitivity.** Held-out AUROC (%) for XGB-Past and XGB-GFF when GFFs are constructed from the first/subsampled S generated trajectories per indexed discharge. The $S = 50$ column is copied from the main full-budget XGB-GFF experiment.

				XGB-GFF rollout samples				
Dataset	Task	Horizon	XGB-Past	$S = 1$	$S = 5$	$S = 10$	$S = 25$	$S = 50$
Hospital A	Creatinine	30d	95.02 ± 0.19	93.08 ± 0.03	95.50 ± 0.02	95.80 ± 0.01	95.93 ± 0.02	96.07 ± 0.02
Hospital A	Creatinine	730d	90.73 ± 0.41	89.52 ± 0.01	91.06 ± 0.01	91.17 ± 0.05	91.48 ± 0.01	91.60 ± 0.02
Hospital A	Death	30d	82.19 ± 0.05	78.77 ± 0.11	81.03 ± 0.03	82.25 ± 0.05	82.92 ± 0.03	83.50 ± 0.04
Hospital A	Death	730d	77.64 ± 0.08	73.71 ± 0.01	76.34 ± 0.02	77.20 ± 0.01	77.90 ± 0.02	78.37 ± 0.02
Hospital A	Hematocrit	30d	91.28 ± 0.32	89.68 ± 0.02	91.70 ± 0.00	92.05 ± 0.01	92.28 ± 0.00	92.43 ± 0.00
Hospital A	Hematocrit	730d	83.87 ± 0.07	80.21 ± 0.03	83.27 ± 0.02	83.86 ± 0.02	84.16 ± 0.02	84.36 ± 0.01
Hospital A	Hemoglobin	30d	91.82 ± 0.01	89.79 ± 0.00	91.93 ± 0.04	92.29 ± 0.01	92.52 ± 0.02	92.63 ± 0.01
Hospital A	Hemoglobin	730d	84.34 ± 0.44	80.93 ± 0.04	84.03 ± 0.02	84.62 ± 0.10	85.01 ± 0.04	85.12 ± 0.01
Hospital A	Leukocyte	30d	87.51 ± 0.01	83.17 ± 0.02	86.96 ± 0.00	87.96 ± 0.00	88.47 ± 0.01	88.67 ± 0.01
Hospital A	Leukocyte	730d	82.48 ± 0.04	77.30 ± 0.02	81.50 ± 0.02	82.50 ± 0.02	83.14 ± 0.01	83.47 ± 0.02
Hospital A	Platelet	30d	94.22 ± 0.02	91.27 ± 0.09	94.16 ± 0.00	94.55 ± 0.02	94.78 ± 0.02	94.91 ± 0.00
Hospital A	Platelet	730d	85.80 ± 0.07	81.94 ± 0.01	85.92 ± 0.02	86.49 ± 0.02	86.95 ± 0.05	87.25 ± 0.01
Hospital A	Readmission	30d	85.77 ± 0.01	82.76 ± 0.05	85.67 ± 0.02	86.25 ± 0.09	86.55 ± 0.05	86.81 ± 0.01
Hospital A	Readmission	730d	78.30 ± 0.02	71.57 ± 0.02	76.98 ± 0.01	77.82 ± 0.01	78.56 ± 0.01	78.79 ± 0.01
MIMIC-IV	Creatinine	30d	93.51 ± 0.28	90.44 ± 0.32	93.44 ± 0.09	93.81 ± 0.19	94.07 ± 0.12	94.14 ± 0.07
MIMIC-IV	Creatinine	730d	90.92 ± 0.00	89.14 ± 0.00	90.56 ± 0.11	91.06 ± 0.19	91.45 ± 0.04	91.68 ± 0.04
MIMIC-IV	Death	30d	84.04 ± 1.03	84.07 ± 0.00	87.25 ± 0.00	87.49 ± 0.35	88.12 ± 0.14	88.65 ± 0.32
MIMIC-IV	Death	730d	79.59 ± 0.06	77.43 ± 0.28	81.31 ± 0.01	81.78 ± 0.07	82.63 ± 0.00	82.54 ± 0.05
MIMIC-IV	Hematocrit	30d	88.76 ± 0.12	85.67 ± 0.00	89.03 ± 0.00	89.09 ± 0.12	89.90 ± 0.03	89.86 ± 0.00
MIMIC-IV	Hematocrit	730d	84.42 ± 0.13	81.35 ± 0.07	84.25 ± 0.00	84.94 ± 0.04	85.46 ± 0.13	85.63 ± 0.06
MIMIC-IV	Hemoglobin	30d	88.71 ± 0.10	85.41 ± 0.06	88.49 ± 0.05	88.79 ± 0.08	89.46 ± 0.07	89.52 ± 0.03
MIMIC-IV	Hemoglobin	730d	84.74 ± 0.00	81.58 ± 0.00	84.70 ± 0.07	85.22 ± 0.16	85.75 ± 0.07	85.83 ± 0.04
MIMIC-IV	Leukocyte	30d	85.77 ± 0.12	81.48 ± 0.10	85.54 ± 0.00	86.62 ± 0.02	87.10 ± 0.06	87.58 ± 0.07
MIMIC-IV	Leukocyte	730d	83.53 ± 0.48	79.54 ± 0.08	83.94 ± 0.00	84.58 ± 0.07	85.42 ± 0.02	85.72 ± 0.05
MIMIC-IV	Platelet	30d	94.62 ± 0.00	92.89 ± 0.24	95.13 ± 0.15	95.01 ± 0.22	95.83 ± 0.00	95.68 ± 0.12
MIMIC-IV	Platelet	730d	85.89 ± 0.09	83.42 ± 0.04	86.61 ± 0.13	87.00 ± 0.00	87.12 ± 0.19	86.91 ± 0.00
MIMIC-IV	Readmission	30d	69.69 ± 0.15	66.45 ± 0.08	69.94 ± 0.09	71.44 ± 0.09	72.09 ± 0.21	72.26 ± 0.00
MIMIC-IV	Readmission	730d	70.27 ± 0.08	68.66 ± 0.00	72.13 ± 0.06	73.12 ± 0.00	73.83 ± 0.12	74.20 ± 0.09

Table 15: **Rollout-sample win rates against XGB-Past.** Each cell reports the number and percentage of task comparisons in which XGB-GFF outperforms XGB-Past for that dataset, horizon, and rollout budget. The combined row aggregates over all dataset–task–horizon comparisons.

Dataset	Horizon	$S = 1$	$S = 5$	$S = 10$	$S = 25$	$S = 50$
Hospital A	30d	0/7 (0.0%)	3/7 (42.9%)	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)
Hospital A	730d	0/7 (0.0%)	2/7 (28.6%)	4/7 (57.1%)	7/7 (100.0%)	7/7 (100.0%)
MIMIC-IV	30d	1/7 (14.3%)	4/7 (57.1%)	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)
MIMIC-IV	730d	0/7 (0.0%)	4/7 (57.1%)	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)
Combined	All	1/28 (3.6%)	13/28 (46.4%)	25/28 (89.3%)	28/28 (100.0%)	28/28 (100.0%)

Table 16: **GFF vocabulary-size sensitivity.** Held-out AUROC (%) for XGB-Past and XGB-GFF when retaining only the top V% of generated-future feature codes by training-set metadata frequency. The $V = 100\%$ column is copied from the main full-vocabulary XGB-GFF experiment.

				XGB-GFF vocabulary size			
Dataset	Task	Horizon	XGB-Past	$V = 25\%$	$V = 50\%$	$V = 75\%$	$V = 100\%$
Hospital A	Creatinine	30d	95.02 ± 0.19	96.02 ± 0.01	96.02 ± 0.02	96.06 ± 0.04	96.07 ± 0.02
Hospital A	Creatinine	730d	90.73 ± 0.41	91.54 ± 0.01	91.62 ± 0.02	91.57 ± 0.02	91.60 ± 0.02
Hospital A	Death	30d	82.19 ± 0.05	81.09 ± 0.05	81.83 ± 0.00	82.22 ± 0.03	83.50 ± 0.04
Hospital A	Death	730d	77.64 ± 0.08	76.67 ± 0.02	77.24 ± 0.02	77.90 ± 0.02	78.37 ± 0.02
Hospital A	Hematocrit	30d	91.28 ± 0.32	92.08 ± 0.02	92.39 ± 0.01	92.40 ± 0.00	92.43 ± 0.00
Hospital A	Hematocrit	730d	83.87 ± 0.07	84.07 ± 0.01	84.30 ± 0.02	84.34 ± 0.01	84.36 ± 0.01
Hospital A	Hemoglobin	30d	91.82 ± 0.01	92.29 ± 0.03	92.62 ± 0.00	92.63 ± 0.01	92.63 ± 0.01
Hospital A	Hemoglobin	730d	84.34 ± 0.44	84.74 ± 0.08	85.02 ± 0.02	85.12 ± 0.00	85.12 ± 0.01
Hospital A	Leukocyte	30d	87.51 ± 0.01	83.44 ± 0.01	88.49 ± 0.01	88.70 ± 0.01	88.67 ± 0.01
Hospital A	Leukocyte	730d	82.48 ± 0.04	80.70 ± 0.01	83.36 ± 0.03	83.46 ± 0.04	83.47 ± 0.02
Hospital A	Platelet	30d	94.22 ± 0.02	91.10 ± 0.02	94.87 ± 0.03	94.92 ± 0.00	94.91 ± 0.00
Hospital A	Platelet	730d	85.80 ± 0.07	83.15 ± 0.03	87.12 ± 0.04	87.20 ± 0.02	87.25 ± 0.01
Hospital A	Readmission	30d	85.77 ± 0.01	86.72 ± 0.02	86.75 ± 0.01	86.77 ± 0.04	86.81 ± 0.01
Hospital A	Readmission	730d	78.30 ± 0.02	78.57 ± 0.00	78.58 ± 0.04	78.75 ± 0.03	78.79 ± 0.01
MIMIC-IV	Creatinine	30d	93.51 ± 0.28	94.22 ± 0.00	94.13 ± 0.10	94.14 ± 0.07	94.14 ± 0.07
MIMIC-IV	Creatinine	730d	90.92 ± 0.00	91.67 ± 0.06	91.62 ± 0.01	91.63 ± 0.02	91.68 ± 0.04
MIMIC-IV	Death	30d	84.04 ± 1.03	88.37 ± 0.22	88.65 ± 0.10	88.65 ± 0.32	88.65 ± 0.32
MIMIC-IV	Death	730d	79.59 ± 0.06	82.63 ± 0.02	82.75 ± 0.04	82.74 ± 0.09	82.54 ± 0.05
MIMIC-IV	Hematocrit	30d	88.76 ± 0.12	89.74 ± 0.06	89.90 ± 0.04	89.82 ± 0.09	89.86 ± 0.00
MIMIC-IV	Hematocrit	730d	84.42 ± 0.13	85.60 ± 0.02	85.73 ± 0.02	85.77 ± 0.02	85.63 ± 0.06
MIMIC-IV	Hemoglobin	30d	88.71 ± 0.10	89.38 ± 0.06	89.40 ± 0.10	89.36 ± 0.06	89.52 ± 0.03
MIMIC-IV	Hemoglobin	730d	84.74 ± 0.00	85.82 ± 0.00	85.76 ± 0.04	85.84 ± 0.04	85.83 ± 0.04
MIMIC-IV	Leukocyte	30d	85.77 ± 0.12	87.54 ± 0.00	87.70 ± 0.00	87.57 ± 0.07	87.58 ± 0.07
MIMIC-IV	Leukocyte	730d	83.53 ± 0.48	85.71 ± 0.04	85.76 ± 0.05	85.76 ± 0.05	85.72 ± 0.05
MIMIC-IV	Platelet	30d	94.62 ± 0.00	95.39 ± 0.19	95.81 ± 0.12	95.68 ± 0.12	95.68 ± 0.12
MIMIC-IV	Platelet	730d	85.89 ± 0.09	87.26 ± 0.11	87.46 ± 0.16	87.29 ± 0.00	86.91 ± 0.00
MIMIC-IV	Readmission	30d	69.69 ± 0.15	71.92 ± 0.00	72.20 ± 0.07	72.21 ± 0.05	72.26 ± 0.00
MIMIC-IV	Readmission	730d	70.27 ± 0.08	74.11 ± 0.07	74.02 ± 0.08	74.20 ± 0.09	74.20 ± 0.09

Table 17: **GFF vocabulary-size win rates against XGB-Past.** Each cell reports the number and percentage of task comparisons in which XGB-GFF outperforms XGB-Past for that dataset, horizon, and retained generated-feature vocabulary size. The combined row aggregates over all dataset–task–horizon comparisons.

Dataset	Horizon	$V = 25\%$	$V = 50\%$	$V = 75\%$	$V = 100\%$
Hospital A	30d	4/7 (57.1%)	6/7 (85.7%)	7/7 (100.0%)	7/7 (100.0%)
Hospital A	730d	4/7 (57.1%)	6/7 (85.7%)	7/7 (100.0%)	7/7 (100.0%)
MIMIC-IV	30d	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)
MIMIC-IV	730d	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)	7/7 (100.0%)
Combined	All	22/28 (78.6%)	26/28 (92.9%)	28/28 (100.0%)	28/28 (100.0%)

Table 18: **Target-adjacent feature-removal ablation.** Held-out AUROC (%) is mean ± standard deviation over five refit seeds. Δ^* is XGB-GFF* minus XGB-Past; Δ_{full} is full XGB-GFF minus XGB-GFF*.

Dataset	Endpoint	Horizon	XGB-Past	XGB-GFF*	Δ^*	XGB-GFF	Δ_{full}
Hospital A	Creatinine	30d	95.02 ± 0.19	91.09 ± 0.03	-3.92	96.07 ± 0.02	+4.97
Hospital A	Death	30d	82.19 ± 0.05	83.35 ± 0.01	+1.16	83.50 ± 0.04	+0.15
Hospital A	Hematocrit	30d	91.28 ± 0.32	92.05 ± 0.02	+0.77	92.43 ± 0.00	+0.38
Hospital A	Hemoglobin	30d	91.82 ± 0.01	92.37 ± 0.01	+0.55	92.63 ± 0.01	+0.26
Hospital A	Leukocyte	30d	87.51 ± 0.01	81.25 ± 0.02	-6.26	88.67 ± 0.01	+7.42
Hospital A	Platelet	30d	94.22 ± 0.02	91.39 ± 0.03	-2.82	94.91 ± 0.00	+3.52
Hospital A	Readmission	30d	85.77 ± 0.01	86.75 ± 0.00	+0.98	86.81 ± 0.01	+0.06
Hospital A	Creatinine	730d	90.73 ± 0.41	87.76 ± 0.01	-2.97	91.60 ± 0.02	+3.84
Hospital A	Death	730d	77.64 ± 0.08	78.25 ± 0.02	+0.61	78.37 ± 0.02	+0.12
Hospital A	Hematocrit	730d	83.87 ± 0.07	84.12 ± 0.01	+0.25	84.36 ± 0.01	+0.24
Hospital A	Hemoglobin	730d	84.34 ± 0.44	84.84 ± 0.04	+0.50	85.12 ± 0.01	+0.27
Hospital A	Leukocyte	730d	82.48 ± 0.04	75.18 ± 0.09	-7.30	83.47 ± 0.02	+8.29
Hospital A	Platelet	730d	85.80 ± 0.07	83.49 ± 0.01	-2.31	87.25 ± 0.01	+3.77
Hospital A	Readmission	730d	78.30 ± 0.02	78.85 ± 0.01	+0.56	78.79 ± 0.01	-0.07
MIMIC-IV	Creatinine	30d	93.51 ± 0.28	92.04 ± 0.14	-1.48	94.14 ± 0.07	+2.11
MIMIC-IV	Death	30d	84.04 ± 1.03	88.27 ± 0.17	+4.23	88.65 ± 0.32	+0.38
MIMIC-IV	Hematocrit	30d	88.76 ± 0.12	89.52 ± 0.06	+0.76	89.86 ± 0.00	+0.34
MIMIC-IV	Hemoglobin	30d	88.71 ± 0.10	89.30 ± 0.06	+0.59	89.52 ± 0.03	+0.22
MIMIC-IV	Leukocyte	30d	85.77 ± 0.12	82.74 ± 0.06	-3.03	87.58 ± 0.07	+4.84
MIMIC-IV	Platelet	30d	94.62 ± 0.00	93.31 ± 0.21	-1.31	95.68 ± 0.12	+2.37
MIMIC-IV	Readmission	30d	69.69 ± 0.15	72.06 ± 0.00	+2.37	72.26 ± 0.00	+0.20
MIMIC-IV	Creatinine	730d	90.92 ± 0.00	89.30 ± 0.06	-1.62	91.68 ± 0.04	+2.37
MIMIC-IV	Death	730d	79.59 ± 0.06	82.51 ± 0.01	+2.92	82.54 ± 0.05	+0.03
MIMIC-IV	Hematocrit	730d	84.42 ± 0.13	85.41 ± 0.05	+1.00	85.63 ± 0.06	+0.21
MIMIC-IV	Hemoglobin	730d	84.74 ± 0.00	85.61 ± 0.07	+0.87	85.83 ± 0.04	+0.22
MIMIC-IV	Leukocyte	730d	83.53 ± 0.48	80.01 ± 0.08	-3.52	85.72 ± 0.05	+5.71
MIMIC-IV	Platelet	730d	85.89 ± 0.09	84.96 ± 0.08	-0.93	86.91 ± 0.00	+1.95
MIMIC-IV	Readmission	730d	70.27 ± 0.08	74.17 ± 0.00	+3.90	74.20 ± 0.09	+0.03

Appendix I. Subgroup AUROC Analysis

We evaluated subgroup AUROC for XGB-GFF and XGB-Past using age and sex subgroup membership extracted at the discharge prediction time. Table 19 reports seed-averaged AUROC for every dataset, endpoint, horizon, and subgroup. Table 20 reports the percentage of dataset–endpoint–horizon experiments in which the seed-averaged XGB-GFF AUROC is higher than the corresponding seed-averaged XGB-Past AUROC.

Table 19: **Subgroup AUROC by task.** AUROC values are averaged over five XGBoost random seeds for each dataset–endpoint–horizon setting. Subgroup columns show XGB-Past (P) and XGB-GFF (G).

			Age >65		Age ≤65		Female		Male	
Dataset	Endpoint	Horizon	P	G	P	G	P	G	P	G
Hospital A	creatinine	30d	0.954	0.962	0.946	0.960	0.961	0.969	0.940	0.952
Hospital A	death	30d	0.795	0.818	0.874	0.875	0.831	0.840	0.813	0.830
Hospital A	hematocrit	30d	0.911	0.923	0.913	0.924	0.912	0.924	0.913	0.924
Hospital A	hemoglobin	30d	0.922	0.930	0.911	0.919	0.916	0.925	0.920	0.928
Hospital A	leukocyte	30d	0.871	0.882	0.879	0.892	0.885	0.893	0.866	0.881
Hospital A	platelet	30d	0.936	0.945	0.948	0.953	0.953	0.958	0.932	0.940
Hospital A	readmission	30d	0.842	0.854	0.880	0.888	0.848	0.861	0.866	0.874
Hospital A	creatinine	730d	0.905	0.915	0.914	0.923	0.916	0.922	0.897	0.908
Hospital A	death	730d	0.761	0.770	0.808	0.823	0.774	0.778	0.778	0.787
Hospital A	hematocrit	730d	0.823	0.828	0.858	0.864	0.840	0.845	0.837	0.842
Hospital A	hemoglobin	730d	0.836	0.842	0.852	0.862	0.842	0.851	0.845	0.851
Hospital A	leukocyte	730d	0.821	0.828	0.831	0.846	0.835	0.842	0.815	0.828
Hospital A	platelet	730d	0.846	0.859	0.872	0.890	0.859	0.872	0.853	0.869
Hospital A	readmission	730d	0.766	0.770	0.810	0.817	0.777	0.783	0.788	0.792
MIMIC-IV	creatinine	30d	0.919	0.929	0.947	0.951	0.938	0.947	0.927	0.933
MIMIC-IV	death	30d	0.788	0.852	0.884	0.911	0.877	0.905	0.807	0.869
MIMIC-IV	hematocrit	30d	0.857	0.876	0.911	0.916	0.884	0.899	0.891	0.899
MIMIC-IV	hemoglobin	30d	0.855	0.868	0.913	0.916	0.879	0.892	0.895	0.898
MIMIC-IV	leukocyte	30d	0.857	0.871	0.857	0.878	0.842	0.867	0.868	0.882
MIMIC-IV	platelet	30d	0.937	0.946	0.953	0.966	0.945	0.963	0.945	0.951
MIMIC-IV	readmission	30d	0.663	0.686	0.717	0.745	0.689	0.723	0.700	0.718
MIMIC-IV	creatinine	730d	0.894	0.896	0.924	0.935	0.900	0.906	0.910	0.918
MIMIC-IV	death	730d	0.744	0.776	0.817	0.840	0.820	0.848	0.772	0.802
MIMIC-IV	hematocrit	730d	0.799	0.815	0.875	0.886	0.844	0.854	0.845	0.858
MIMIC-IV	hemoglobin	730d	0.815	0.827	0.870	0.879	0.843	0.857	0.852	0.859
MIMIC-IV	leukocyte	730d	0.825	0.854	0.842	0.859	0.813	0.842	0.853	0.869
MIMIC-IV	platelet	730d	0.848	0.871	0.868	0.869	0.856	0.867	0.858	0.867
MIMIC-IV	readmission	730d	0.665	0.708	0.731	0.767	0.691	0.731	0.710	0.749

Table 20: **Subgroup AUROC win rates.** Win rate is the percentage of dataset–endpoint–horizon experiments for which the seed-averaged XGB-GFF subgroup AUROC is higher than the corresponding seed-averaged XGB-Past subgroup AUROC.

Dataset	Subgroup	Experiments	GFF wins	Win rate	Mean AUROC$_P$	Mean AUROC$_G$	Mean ΔAUROC
Hospital A	Age >65	14	14/14	100.0%	0.856	0.866	+0.010
Hospital A	Age ≤65	14	14/14	100.0%	0.878	0.888	+0.010
Hospital A	Female	14	14/14	100.0%	0.868	0.876	+0.008
Hospital A	Male	14	14/14	100.0%	0.862	0.872	+0.010
MIMIC-IV	Age >65	14	14/14	100.0%	0.819	0.841	+0.022
MIMIC-IV	Age ≤65	14	14/14	100.0%	0.865	0.880	+0.015
MIMIC-IV	Female	14	14/14	100.0%	0.844	0.864	+0.020
MIMIC-IV	Male	14	14/14	100.0%	0.845	0.862	+0.017

Table 21: **Rollout generation, storage, and cache-construction costs.** Wall-clock rollout time is extrapolated from measured one-rollout held-out runs to 50 rollouts per indexed discharge. Storage footprints report cached raw generated trajectories, all-code generated TTEs, and GFF probability features. TTE construction timings are for the held-out split; GFF probability construction timings are for train, tuning, and held-out splits.

Dataset	Discharges	Trajectories	Rollout latency	Rollout wall-clock	Raw rollouts	TTE cache	GFF cache	TTE construction	GFF construction
Hospital A	1,531,893	76,594,650	698 ms	297 h all splits; 7.59 h held-out	71 GB	17 GB	225 MB	22.8 s	4.85 min
MIMIC-IV	397,737	19,886,850	759 ms	83.8 h all splits; 2.24 h held-out	12 GB	3.2 GB	81 MB	12.5 s	3.92 min

Table 22: **Training and cached inference costs.** Training times are for the death-at-30-days benchmark unless noted. ETHOS pretraining times extrapolate one measured epoch-equivalent run to 200 epochs. XGB refit-only times use one seed with fixed selected hyperparameters and include warm cached feature loading, model fitting, held-out feature loading, prediction, and output writing; parenthetical values extrapolate the same per-setting cost over the 14 task–horizon settings. Cached inference latencies include held-out feature loading, model/scoring computation, and output writing. Zero-Shot loads 50-rollout TTEs and outputs risks; rollout generation is excluded from timing in this table.

Dataset	Method	Training time	Cached inference latency per discharge
Hospital A	ETHOS pretraining	5.61 h	N/A
Hospital A	Supervised transformer	24.5 min	0.134 ms
Hospital A	Zero-Shot	N/A	0.0036 ms
Hospital A	XGB-Past refit only	54.6 s (12.8 min over 14 settings)	0.0169 ms
Hospital A	XGB-GFF refit only	1.60 s (22.5 s over 14 settings)	0.0075 ms
MIMIC-IV	ETHOS pretraining	2.55 h	N/A
MIMIC-IV	Supervised transformer	5.45 min	0.247 ms
MIMIC-IV	Zero-Shot	N/A	0.0064 ms
MIMIC-IV	XGB-Past refit only	32.4 s (7.57 min over 14 settings)	0.0417 ms
MIMIC-IV	XGB-GFF refit only	1.87 s (26.2 s over 14 settings)	0.0264 ms

Appendix J. Compute and Storage Costs

We measured the main computational costs of the generative-feature pipeline on one NVIDIA GH200 GPU for neural rollout and transformer benchmarks, and CPU XGBoost benchmarks using cached feature matrices. Rollout generation is the dominant cost: for 50 generated futures per discharge, rollout latency was 698 ms per discharge in Hospital A and 759 ms per discharge in MIMIC-IV. Once generated trajectories are cached, converting trajectories to time-to-event (TTE) caches, scoring cached trajectories, and XGBoost inference are substantially cheaper. In the tables below, XGBoost refit and inference timings report a single model with fixed selected hyperparameters, matching the single-seed supervised transformer timing; the main XGBoost evaluation protocol still averages over five refit seeds as described above.

Appendix K. XGB-GFF Versus Past-and-Future Features

As an additional ablation, we compared XGB-GFF against an XGBoost model trained on the concatenation of past-window tabular features and generated-future features (XGB-Past+GFF). Both methods use the same task-specific XGBoost tuning and five-seed refit protocol. XGB-GFF outperformed XGB-Past+GFF in 3/14 Hospital A comparisons (21.4%) and 1/14 MIMIC-IV comparisons (7.1%). Across all 28 dataset–endpoint–horizon comparisons, the XGB-GFF win rate was 4/28 (14.3%).

Table 23: **XGB-GFF versus XGB-Past+GFF.** Held-out AUROC (%) for generated-future features alone and for concatenated past-window plus generated-future features. Positive Δ indicates higher AUROC for XGB-GFF.

Dataset	Task	Horizon	XGB-GFF	XGB-Past+GFF	Δ
Hospital A	Creatinine	30d	96.07 ± 0.02	96.03 ± 0.00	+0.04
Hospital A	Creatinine	730d	91.60 ± 0.02	91.84 ± 0.04	-0.24
Hospital A	Death	30d	83.50 ± 0.04	84.14 ± 0.11	-0.64
Hospital A	Death	730d	78.37 ± 0.02	79.21 ± 0.04	-0.84
Hospital A	Hematocrit	30d	92.43 ± 0.00	92.43 ± 0.01	0.00
Hospital A	Hematocrit	730d	84.36 ± 0.01	84.62 ± 0.08	-0.26
Hospital A	Hemoglobin	30d	92.63 ± 0.01	92.61 ± 0.02	+0.03
Hospital A	Hemoglobin	730d	85.12 ± 0.01	85.36 ± 0.04	-0.25
Hospital A	Leukocyte	30d	88.67 ± 0.01	88.98 ± 0.04	-0.31
Hospital A	Leukocyte	730d	83.47 ± 0.02	83.99 ± 0.05	-0.52
Hospital A	Platelet	30d	94.91 ± 0.00	95.01 ± 0.03	-0.10
Hospital A	Platelet	730d	87.25 ± 0.01	87.23 ± 0.05	+0.02
Hospital A	Readmission	30d	86.81 ± 0.01	86.98 ± 0.02	-0.17
Hospital A	Readmission	730d	78.79 ± 0.01	79.81 ± 0.02	-1.02
MIMIC-IV	Creatinine	30d	94.14 ± 0.07	94.30 ± 0.17	-0.16
MIMIC-IV	Creatinine	730d	91.68 ± 0.04	91.90 ± 0.03	-0.23
MIMIC-IV	Death	30d	88.65 ± 0.32	89.04 ± 0.19	-0.39
MIMIC-IV	Death	730d	82.54 ± 0.05	83.26 ± 0.16	-0.72
MIMIC-IV	Hematocrit	30d	89.86 ± 0.00	89.87 ± 0.00	-0.01
MIMIC-IV	Hematocrit	730d	85.63 ± 0.06	86.17 ± 0.00	-0.54
MIMIC-IV	Hemoglobin	30d	89.52 ± 0.03	89.29 ± 0.00	+0.23
MIMIC-IV	Hemoglobin	730d	85.83 ± 0.04	86.08 ± 0.00	-0.24
MIMIC-IV	Leukocyte	30d	87.58 ± 0.07	87.85 ± 0.08	-0.28
MIMIC-IV	Leukocyte	730d	85.72 ± 0.05	86.31 ± 0.05	-0.59
MIMIC-IV	Platelet	30d	95.68 ± 0.12	95.88 ± 0.00	-0.20
MIMIC-IV	Platelet	730d	86.91 ± 0.00	87.58 ± 0.27	-0.67
MIMIC-IV	Readmission	30d	72.26 ± 0.00	72.60 ± 0.06	-0.34
MIMIC-IV	Readmission	730d	74.20 ± 0.09	74.65 ± 0.05	-0.45

Table 24: **XGB-Past+GFF aggregate wins.** Win rate is the percentage of endpoint–horizon comparisons in which XGB-Past+GFF outperforms XGB-GFF. Mean ΔAUROC is computed as XGB-Past+GFF minus XGB-GFF, in AUROC percentage points. Win and loss deltas average over only the comparisons won or lost by XGB-Past+GFF.

Dataset	XGB-Past+GFF wins	Win rate	Mean ΔAUROC	Mean win Δ	Mean loss Δ
Hospital A	11/14	78.6%	+0.30	+0.39	-0.03
MIMIC-IV	13/14	92.9%	+0.33	+0.37	-0.23
Combined	24/28	85.7%	+0.32	+0.38	-0.08

Appendix L. XGB-GFF Versus Zero-Shot Calibration

We additionally compared held-out calibration and Brier score for direct zero-shot ETHOS endpoint probabilities and XGB-GFF. Zero-shot probabilities are computed as the fraction of generated task-level trajectories in which the endpoint occurs by horizon; XGB-GFF probabilities are averaged over five XGBoost refit seeds for calibration plots. XGB-GFF achieved lower Brier score in 13/14 Hospital A comparisons (92.9%) and 12/14 MIMIC-IV comparisons (85.7%), for a combined win rate of 25/28 comparisons (89.3%). Mean zero-shot minus XGB-GFF Brier improvement was +0.0041 in Hospital A and +0.0044 in MIMIC-IV.

Table 25: **Brier score for XGB-GFF versus zero-shot.** Held-out Brier score is lower when better. XGB-GFF values are mean ± standard deviation over five XGBoost refit seeds; zero-shot values use the direct ETHOS rollout probability for the target endpoint. Positive Δ indicates lower Brier score for XGB-GFF.

Dataset	Endpoint	Horizon	Zero-Shot	XGB-GFF	Δ
Hospital A	Creatinine	30d	0.0425	0.0419 ± 0.0004	+0.0006
Hospital A	Creatinine	730d	0.0823	0.0791 ± 0.0002	+0.0032
Hospital A	Death	30d	0.0500	0.0487 ± 0.0001	+0.0013
Hospital A	Death	730d	0.1938	0.1607 ± 0.0001	+0.0330
Hospital A	Hematocrit	30d	0.0829	0.0816 ± 0.0000	+0.0013
Hospital A	Hematocrit	730d	0.1481	0.1577 ± 0.0086	-0.0096
Hospital A	Hemoglobin	30d	0.0800	0.0789 ± 0.0000	+0.0011
Hospital A	Hemoglobin	730d	0.1458	0.1415 ± 0.0001	+0.0043
Hospital A	Leukocyte	30d	0.1076	0.1052 ± 0.0000	+0.0023
Hospital A	Leukocyte	730d	0.1714	0.1629 ± 0.0001	+0.0084
Hospital A	Platelet	30d	0.0475	0.0458 ± 0.0000	+0.0017
Hospital A	Platelet	730d	0.1041	0.0991 ± 0.0000	+0.0049
Hospital A	Readmission	30d	0.1437	0.1419 ± 0.0000	+0.0019
Hospital A	Readmission	730d	0.1348	0.1321 ± 0.0000	+0.0027
MIMIC-IV	Creatinine	30d	0.0267	0.0271 ± 0.0002	-0.0004
MIMIC-IV	Creatinine	730d	0.0697	0.0661 ± 0.0002	+0.0037
MIMIC-IV	Death	30d	0.0226	0.0222 ± 0.0000	+0.0004
MIMIC-IV	Death	730d	0.1449	0.1242 ± 0.0003	+0.0207
MIMIC-IV	Hematocrit	30d	0.0441	0.0432 ± 0.0000	+0.0009
MIMIC-IV	Hematocrit	730d	0.1134	0.1083 ± 0.0001	+0.0051
MIMIC-IV	Hemoglobin	30d	0.0455	0.0443 ± 0.0001	+0.0012
MIMIC-IV	Hemoglobin	730d	0.1169	0.1097 ± 0.0001	+0.0072
MIMIC-IV	Leukocyte	30d	0.0857	0.0853 ± 0.0004	+0.0005
MIMIC-IV	Leukocyte	730d	0.1502	0.1472 ± 0.0003	+0.0030
MIMIC-IV	Platelet	30d	0.0197	0.0195 ± 0.0002	+0.0001
MIMIC-IV	Platelet	730d	0.0572	0.0573 ± 0.0000	-0.0001
MIMIC-IV	Readmission	30d	0.1801	0.1746 ± 0.0000	+0.0055
MIMIC-IV	Readmission	730d	0.1954	0.1812 ± 0.0002	+0.0142

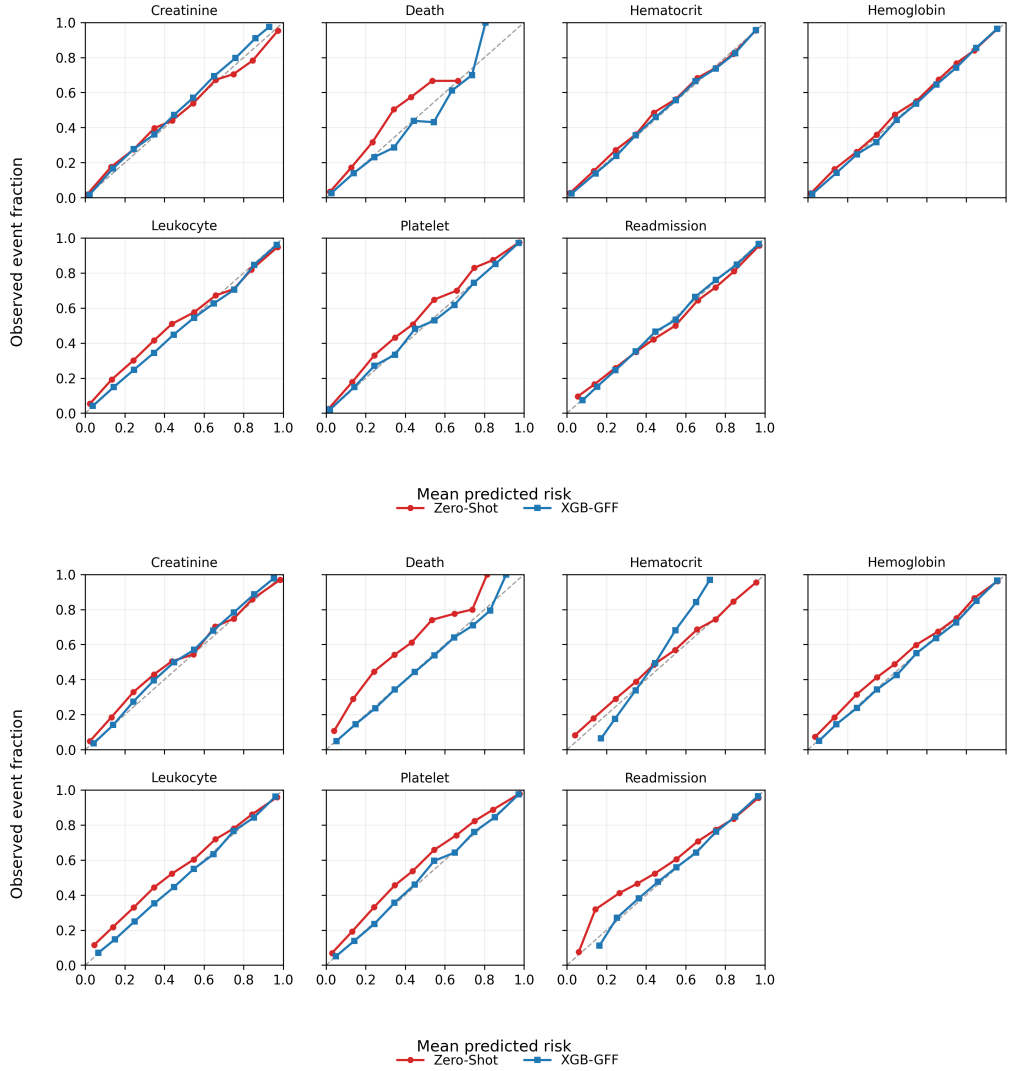


Figure 9: **Calibration curves for Hospital A.** Zero-shot and XGB-GFF calibration curves are shown together by endpoint for 30-day and 730-day horizons. The diagonal line indicates perfect calibration.

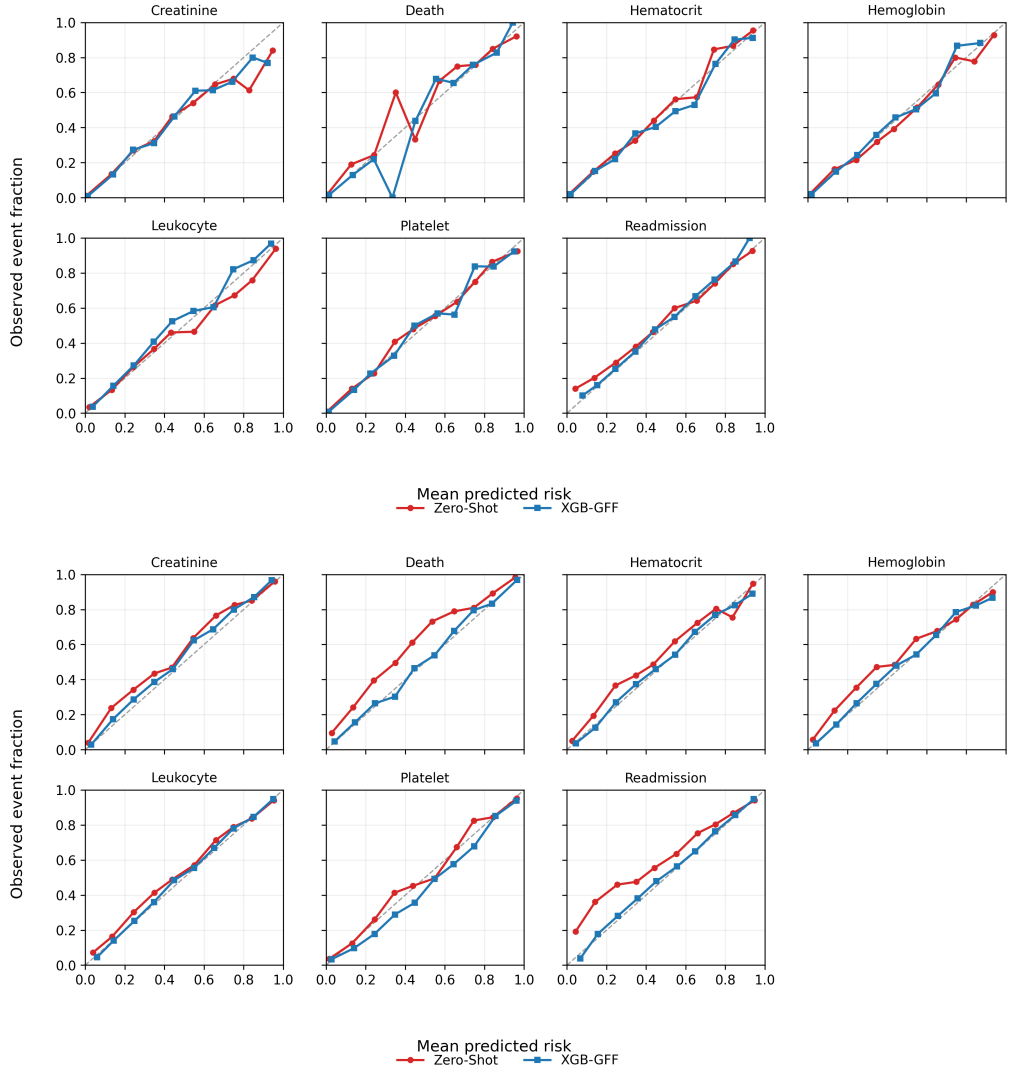


Figure 10: **Calibration curves for MIMIC-IV.** Zero-shot and XGB-GFF calibration curves are shown together by endpoint for 30-day and 730-day horizons. The diagonal line indicates perfect calibration.

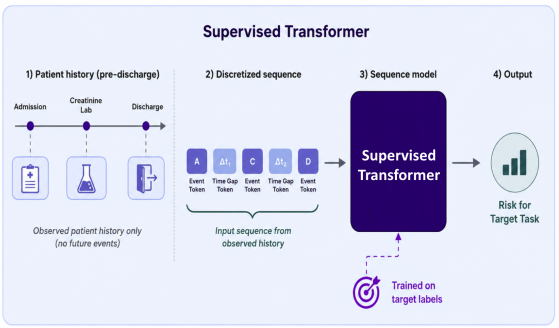
Across the 28 dataset–endpoint–horizon settings, the mean absolute difference between generated-trajectory and held-out endpoint prevalence was 3.4 percentage points overall: 1.5 points at 30 days and 5.2 points at 730 days. The two largest absolute differences were the 730-day death endpoints in Hospital A and MIMIC-IV.

Table 26: **Endpoint prevalence in held-out data and generated trajectories.** Real prevalence is the held-out endpoint fraction by horizon. Trajectory prevalence is the mean zero-shot rollout event probability for the same endpoint and horizon. Positive Δ indicates higher generated-trajectory prevalence.

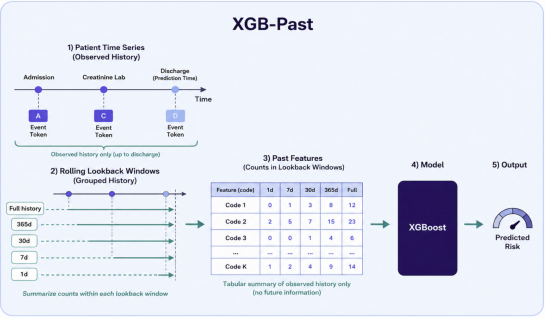
Dataset	Endpoint	Horizon	Real prev.	Traj. prev.	Δ	N
Hospital A	Creatinine	30d	13.0%	12.1%	-0.8 pp	39,135
Hospital A	Creatinine	730d	21.5%	18.5%	-2.9 pp	39,135
Hospital A	Death	30d	5.8%	3.9%	-2.0 pp	39,135
Hospital A	Death	730d	28.0%	14.6%	-13.4 pp	39,135
Hospital A	Hematocrit	30d	20.2%	18.9%	-1.3 pp	39,135
Hospital A	Hematocrit	730d	33.7%	30.1%	-3.6 pp	39,135
Hospital A	Hemoglobin	30d	19.4%	18.1%	-1.3 pp	39,135
Hospital A	Hemoglobin	730d	33.3%	29.1%	-4.2 pp	39,135
Hospital A	Leukocyte	30d	24.3%	21.0%	-3.2 pp	39,135
Hospital A	Leukocyte	730d	45.9%	39.8%	-6.1 pp	39,135
Hospital A	Platelet	30d	13.0%	11.1%	-1.9 pp	39,135
Hospital A	Platelet	730d	22.4%	17.5%	-5.0 pp	39,135
Hospital A	Readmission	30d	47.8%	47.7%	-0.1 pp	39,135
Hospital A	Readmission	730d	80.1%	78.4%	-1.7 pp	39,135
MIMIC-IV	Creatinine	30d	4.4%	4.0%	-0.4 pp	10,640
MIMIC-IV	Creatinine	730d	14.3%	10.7%	-3.6 pp	10,640
MIMIC-IV	Death	30d	3.3%	2.0%	-1.4 pp	10,640
MIMIC-IV	Death	730d	21.3%	12.1%	-9.3 pp	10,640
MIMIC-IV	Hematocrit	30d	6.4%	5.5%	-0.9 pp	10,640
MIMIC-IV	Hematocrit	730d	19.2%	14.7%	-4.5 pp	10,640
MIMIC-IV	Hemoglobin	30d	6.3%	5.4%	-0.9 pp	10,640
MIMIC-IV	Hemoglobin	730d	19.7%	14.5%	-5.2 pp	10,640
MIMIC-IV	Leukocyte	30d	15.2%	14.5%	-0.7 pp	10,640
MIMIC-IV	Leukocyte	730d	39.0%	35.8%	-3.3 pp	10,640
MIMIC-IV	Platelet	30d	4.2%	3.8%	-0.4 pp	10,640
MIMIC-IV	Platelet	730d	10.4%	8.7%	-1.7 pp	10,640
MIMIC-IV	Readmission	30d	28.9%	23.1%	-5.8 pp	10,640
MIMIC-IV	Readmission	730d	68.7%	60.1%	-8.6 pp	10,640

Appendix M. Graphical Overview of Experimental Approaches

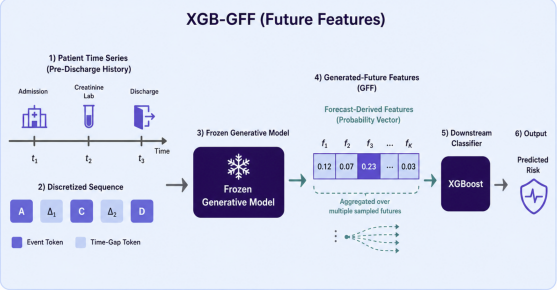
Figure 11 provides graphical overviews of the four supervised and representation-based approaches evaluated in the main experiments. All approaches begin from the same discretized pre-discharge patient history but differ in how that history is represented and supplied to the downstream predictor. The supervised transformer is trained end-to-end for each target task; XGB-Past summarizes the observed history using code counts over rolling lookback windows; XGB-GFF applies XGBoost to forecast-derived event-probability features produced by a frozen generative model; and linear probing fits logistic regression to the final-token embedding from the same frozen generative model. Direct zero-shot prediction and the construction of the generated-future feature vector are illustrated separately in Figure 1.



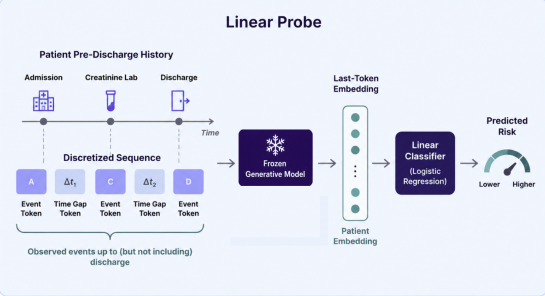
(a) **Supervised Transformer**



(b) **XGB-Past**



(c) **XGB-GFF**



(d) **Linear Probe**

Figure 11: **Graphical overview of the supervised and representation-based experimental approaches.** All methods begin from the same discretized pre-discharge patient history. **(a)** The supervised transformer is trained end-to-end for each target task. **(b)** XGB-Past aggregates observed code counts over multiple rolling lookback windows and supplies the resulting tabular features to XGBoost. **(c)** XGB-GFF supplies horizon-specific generated-future feature probabilities from a frozen generative model to XGBoost. **(d)** Linear probing extracts the final-token embedding from the frozen generative model and fits a logistic-regression classifier. Figure 1 separately illustrates direct zero-shot forecasting and the construction of generated-future features from sampled trajectories.